ЕВРОПЕЙСКИЙ УНИВЕРСИТЕТ В САНКТ-ПЕТЕРБУРГЕ
EUROPEAN UNIVERSITY AT ST PETERSBURG

С.С. Валландер

Лекции по статистике и эконометрике

Санкт-Петербург
2005

Валландер С.С. **Лекции по статистике и эконометрике.** — СПб.: Изд-во Европ. ун-та в С.-Петербурге, 2005. — 248 с.

ISBN 5-94380-036-7

Рецензенты:
Профессор кафедры теории вероятностей и математической статистики СПбГУ
д. ф.-м. н. В.Б. НЕВЗОРОВ

Декан ф-та экономики Европейского университета в Санкт-Петербурге
д. ф.-м. н. С.Л. ПЕЧЕРСКИЙ

Издание осуществлено при финансовой поддержке Института "Открытое общество" (Фонд Сороса), Россия. Грант HBC201

Без объявления
© Европейский университет в Санкт-Петербурге, 2005
ISBN 5-94380-036-7
© С.С. Валландер, 2005

Оглавление

Предисловие

За последние годы автору довелось прочитать ряд курсов математической статистики и эконометрики для слушателей магистерской программы факультета экономики Европейского университета в Санкт-Петербурге (ЕУСПб). Значительная часть излагавшегося материала вошла в настоящие “Лекции...”. Не совсем традиционный стиль изложения связан с тем, что лекции читались для аудитории, уже имеющей высшее образование, причем значительную ее часть составляли выпускники не экономических вузов. Запас базовых математических знаний слушателей, пришедших после разных институтов, сильно различался. Поэтому изложение по возможности (и умению лектора) строилось таким образом, чтобы материал можно было воспринимать на разных уровнях. Я надеюсь, что основные идеи и понятия были доступны всем, и, в то же время, некоторые технически более сложные детали предназначались для более подготовленных слушателей. Эти особенности я старался сохранить и в печатном тексте.

Кроме того, по моему убеждению, чрезвычайно важным аспектом изучения статистики и эконометрики являются общие концепции. Подготовка аудитории позволяла уделить концептуальным вопросам несколько больше внимания, чем это делается в курсах утилитарной или, напротив, формалистической направленности.

Наконец, будет не лишним упомянуть о том, что общие курсы математической статистики и эконометрики составляют только часть большого сбалансированного комплекса курсов, включенных в магистерскую программу по экономике в ЕУСПб (параллельно слушателям предлагаются курсы социально-экономической статистики и прикладной эконометрики, а также предоставляется возможность освоить специализированные компьютерные пакеты, предназначенные для статистических и эконометрических расчетов). Эти обстоятельства в значительной мере объясняют почти полное отсутствие в лекциях конкретных иллюстрирующих примеров.

Работа над книгой — это длительный и непростой труд. Как говорил Уинстон Черчилль, “Написание книги напоминает любовный роман: сначала она для вас развлечение, затем становится вашей любовницей, потом превращается в вашу госпожу и наконец — в тирана” (цитируется по изданию: Уинстон Черчилль. Мускулы мира. Изд-во “ЭКСМО”, М., 2002, с.513). Надеюсь, что некоторым читателям “Лекции. . .” принесут пользу.

Разумеется, за все имеющиеся в тексте оплошности и неточности несу ответственность только я.

Благодарности. В первую очередь хочу поблагодарить коллектив факультета экономики ЕУСПб за прекрасные возможности для работы, постоянную стимулирующую поддержку и общую творческую атмосферу.

При написании “Лекций” автор пользовался поддержкой гранта ИОО НВС201. Мои благодарности и этому институту.

Благодарю своих давних коллег по Санкт-Петербургскому (Ленинградскому) государственному университету В.Б.Невзорова и Я.Ю.Никитина, ознакомившихся с отдельными частями рукописи и высказавшими много полезных и конструктивных замечаний, способствовавших улучшению текста.

Глава 1

Основания статистики

При изучении оснований статистики, как, видимо, и любой другой науки, приходится немалое внимание уделять правильному словоупотреблению, аккуратной терминологии и точным определениям. Мы будем, естественно, использовать русскую версию языка статистики, обращаясь в необходимых случаях к английской. В основном эти две версии согласуются, однако иногда их сравнение позволяет обеспечить более глубокое понимание сути проблемы. Кроме того, английский язык является, де факто, языком международного научного общения, и знание хотя бы некоторых английских специальных терминов и выражений становится просто необходимым.

В настоящей главе мы затронем некоторые фундаментальные идеи, которые обычно не рассматриваются в начальных курсах статистики. Знакомство с элементами теории вероятностей и статистики является крайне желательным, можно сказать — необходимым.

1.1 Статистические данные и случайные величины

При построении теоретической модели статистических данных, как правило, постулируется, что имеющиеся конкретные числа (или наборы чисел) можно представлять себе как "реализовавшиеся значения"некоторых случайных величин. При этом возможны различные понимания случайности и механизма ее возникновения в рассматриваемом явлении, а потому и различные трактовки понятия случайной величины. Остановимся на этом более подробно. Начнем с одного часто встречающегося типа задач.

Простой случайный выбор ([8]). Предположим, что в поле зрения исследователя находится конечная совокупность

объектов (чаще всего, большая и, может быть, труднообозримая). Каждый из этих объектов (индивидуумов, фирм,...) может быть охарактеризован одним или несколькими числами. Выбирая случайным образом один объект из совокупности, исследователь измеряет его характеристики (предполагается, что это возможно) и, тем самым, получает "эмпирические данные". Выражение "выбирая случайным образом"подразумевает активное участие исследователя в выборе, т.е. организацию им некоторого случайного механизма реализации этого эксперимента, а термин "простой выбор"означает, что все объекты рассматриваемой совокупности считаются равноправными, т.е. выбираются с одной и той же вероятностью.

Цель статистического исследования — сделать по эмпирическим данным тот или иной вывод об изучаемом явлении, процессе и т.д. В рассматриваемом примере по характеристикам выбранного объекта исследователь, видимо, хочет судить о всей совокупности изучаемых объектов. Разумеется, это трудно сделать по **одному** объекту, и практически всегда подобный выбор "повторяется". Организовать повторения можно по-разному. Один из наиболее распространенных способов организации так и называется — "повторный выбор". Он характеризуется тем, что выбранный объект каждый раз "возвращается"в изучаемую совокупность, а следующий выбор совершается независимо от всех предыдущих. О других способах организации повторений будет сказано чуть позже.

Примером повторного выбора является изучаемая в курсе теории вероятностей последовательность симметричных испытаний Бернулли. Каждое испытание при этом следует понимать как выбор одного из двух исходов (скажем, одной из двух сторон монеты). Испытания Бернулли по определению независимы, так что выбор действительно является повторным. Для статистики последовательность **симметричных** испытаний Бернулли неинтересна, т.к. изучаемая совокупность из двух объектов ("герб"и "решка") очень проста и, собственно, изучать-то нечего[1]. Повторный выбор обычно используется в тех ситуациях, когда изучаемая совокупность действительно большая, а по относительно небольшой выборке удается достаточно содержательным образом судить о всей, как говорят иногда, "генеральной"совокупности.

[1]Другое дело, что предположение симметричности может оказаться сомнительным — тогда его следует проверять, а это уже типичная статистическая задача, к тому же, не вполне тривиальная.

Мы незаметно подошли к обсуждению традиционной статистической терминологии, порожденной обсуждаемым примером. Последовательность выбираемых объектов называется (случайной) **выборкой**, в обсуждаемом случае повторного выбора — **повторной выборкой**, а вся совокупность объектов, из которой производится выбор — генеральной совокупностью. Фактически же термин "выборка"относится к характеристикам выбранных объектов, т.е. к эмпирическим данным. Семантически мы трактуем, следуя [8], "выбор"как процесс, а "выборку"как результат этого процесса. В английской терминологии выбор — это sampling, выборка — sample (random sample), а генеральная совокупность — population. В учебной литературе по общей статистике можно найти разъяснения и практические рекомендации по организации выбора в различных реальных задачах (см., например, [5]). Мы упомянем лишь так называемый "бесповторный"выбор, при котором ранее выбранные объекты не возвращаются в генеральную совокупность. Если объем выборки пренебрежимо мал по сравнению с объемом всей генеральной совокупности, различиями между повторным и бесповторным выбором можно пренебречь. Заметим также, что в более общих моделях статистических данных зависимые наблюдения (бесповторный выбор — простейший случай зависимости) широко распространены (см., например, [20]).

Подводя итог обсуждению модели простого случайного выбора, еще раз подчеркнем, что случайность в этой модели возникает извне, по воле исследователя, а понятие случайной величины нам, по существу, не потребовалось. Во многих социально-экономических задачах подобные активные эксперименты невозможны, а часто и само представление о генеральной совокупности становится крайне расплывчатым. Поэтому мы сейчас рассмотрим более общую и, как следствие, более абстрактную модель случайности, приспособленную для описания значительно более широкого круга явлений. В отличие от предыдущей, весьма прагматичной, эта модель имеет в первую очередь концептуальный характер. Подобная тенденция типична для современных изложений теории вероятностей и математической статистики (см., например, [12], [1]). Мы, впрочем, не собираемся углубляться в сложные математические конструкции и постараемся обойтись необходимым минимумом.

Удобно иметь в голове конкретный пример, достаточно сложный, чтобы мотивировать общность модели, и достаточно наглядный, чтобы

его можно было обсуждать и на полубытовом уровне. Итак, рассмотрим эволюцию обменного курса рубля к доллару, устанавливаемого в каком-либо конкретном финансовом учреждении, или эволюцию стоимости какой-либо ценной бумаги на соответствующем финансовом рынке. Для нас существенным будет лишь то обстоятельство, что на формирование, скажем, завтрашнего курса или завтрашней стоимости оказывают влияние столь многочисленные факторы, что ни перечислить, ни, тем более, учесть их оказывается невозможным. Тем самым, завтрашнее значение выбранной характеристики оказывается (сегодня) по меньшей мере **не** определимым. Принимаемая нами концепция в двух словах состоит в том, что это завтрашнее значение можно трактовать как **случайное**, а завтра нам станет известным **реализовавшееся значение** этой случайной величины.

Для более подробного и точного описания этой случайности мы постулируем (и это есть уточнение выбранной концепции), что изучаемое нами явление описывается некоторым множеством (часто говорят — пространством) элементарных событий (исходов) Ω, представляющих возможные варианты состояния изучаемого мира (например, финансового рынка). Эти элементарные исходы чаще всего непосредственно не наблюдаются (и именно поэтому наш постулат является частью **теоретической концепции**), но некоторую информацию о состоянии изучаемого мира можно извлечь, отслеживая, происходят ли те или иные наблюдаемые события, или измеряя ту или иную наблюдаемую величину. Стоит подчеркнуть, что элементарный исход нужно понимать как нечто действительно не подлежащее уточнению, т.е. знание этого исхода (если бы оно было возможно) однозначно определяло бы все значения всех характеристик (прошлых, настоящих и будущих) изучаемого явления.

На математическом уровне событие отождествляется с множеством благоприятствующих ему элементарных исходов, т.е. с подмножеством пространства Ω. Запас событий определяется желанием и возможностями исследователя приписывать им осмысленные вероятности. Практически всегда предполагается, что система событий является **алгеброй** множеств, т.е. объединение, пересечение и разность событий снова являются событиями. Более того, предполагается, что эта алгебра **счетно-замкнута**, т.е. пределы монотонных последовательностей событий также являются событиями. Счетно-замкнутая алгебра множеств часто называется **сигма-алгеброй** ($\equiv \sigma$

-алгеброй). О вероятностях событий в статистических задачах будет сказано дальше.

В большинстве своем обсуждаемые события связаны с определенными характеристиками нашего явления. Мы только что видели, что эти характеристики являются **функциями** элементарного исхода, т.е. состояния изучаемого мира. Функции, определенные на пространстве элементарных исходов Ω, называются **случайными величинами**. Тем самым, понятие случайной величины является органической частью принятой концепции.

Сделаем еще несколько уточняющих замечаний о случайных величинах и событиях. Нам потребуются только числовые случайные величины, т.е. функции, принимающие значения в множестве вещественных чисел $\mathbf{R}$, и их многомерные варианты, иногда называемые случайными векторами. Точное математическое определение включает требование определенной согласованности между запасом событий и запасом случайных величин. Именно, случайными величинами называются **измеримые** функции на пространстве Ω. По определению функция $X : \Omega \longrightarrow \mathbf{R}$ измерима, если для любого замкнутого промежутка $[a, b] \subset \mathbf{R}$ прообраз этого промежутка относительно отображения X —

$$X^{-1}([a, b]) = \{\omega \in \Omega : X(\omega) \in [a, b]\}$$

является событием (более наглядно, но менее точно, это событие можно обозначить $\{X \in [a, b]\}$). С точки зрения пользователя это требование выполняется почти автоматически, хотя возможны патологические или казуистические контрпримеры. В подробных курсах теории вероятностей доказывается, что в этом определении замкнутые промежутки можно заменить открытыми или даже произвольными борелевскими множествами (последние нам не потребуются, так что мы не даем точного определения). Можно доказать, что основные арифметические операции над случайными величинами снова дают случайные величины (с обычной оговоркой о невозможности деления на 0), а также что (поточечный) предел последовательности случайных величин снова является случайной величиной. С точки зрения пользователя эти свойства являются, конечно, сами собой разумеющимися.

Важное преимущество трактовки случайных величин как функций с общей областью определения Ω заключается в том, что совместное

их рассмотрение (например, обсуждение их совместных распределений вероятностей — см. следующий параграф и [12]) не создает никаких проблем. Так, мы можем две числовые случайные величины X_1 и X_2 "склеить"в случайный вектор $X = (X_1, X_2)$. При этом события $\{X_i \in [a_i, b_i]\}, i = 1, 2$ автоматически породят событие, относящееся к случайному вектору X:

$$\begin{aligned} \{X \in [a_1, b_1] \times [a_2, b_2]\} &= \{X_1 \in [a_1, b_1], X_2 \in [a_2, b_2]\} \\ &= \{X_1 \in [a_1, b_1]\} \cap \{X_2 \in [a_2, b_2]\}. \end{aligned}$$

Сейчас самое время обратить внимание читателя на запятую в центре среднего выражения. По общепринятому соглашению она понимается как знак пересечения (или, что, по существу, то же самое, как логическая связка "И") — ср. с последним выражением. Подобное использование запятой будет часто встречаться в следующих разделах.

В англоязычной литературе термину "случайная величина"соответствует "random variable". Здесь мы сталкиваемся с некоторым расхождением в терминологии, причем русский вариант выглядит более предпочтительным. Вообще, термином "переменная"злоупотреблять не стоит, поскольку он вызывает весьма расплывчатые догадки и, возможно, вопросы о причинах "переменности"этой величины — "кто и как ее меняет"(а термин "величина"возник у нас, между прочим, как бы сам собой!). Мы еще вернемся к этому обсуждению в гл.6,7.

1.2 Случайные величины и вероятности — кое-что о постановке статистических задач

Хорошо известно одно из основных отличий курса теории вероятностей от курса математической статистики (мы, разумеется, утрируем): в курсе теории вероятностей учат, как по вероятностям некоторых "базисных"событий искать вероятности прочих событий, а в курсе математической статистики интересуются тем, как эти "базисные"вероятности извлечь из статистических данных.

В логическом плане вероятностные понятия явно предшествуют статистическим. "Вероятностник"(probabilist) предполагает, что на σ-алгебре событий задана вероятностная мера (каждому событию приписано неотрицательное число, называемое его вероятностью,

с выполнением известных свойств, главным из которых является аддитивность, даже счетная аддитивность). Статистик (statistician), соглашаясь с ним в целом, подчеркивает, что имеющаяся у него априорная информация о случайном явлении не позволяет эту вероятностную меру однозначно определить, и потому работает со всеми априори допустимыми вероятностными мерами, а иногда, скрепя сердце, добавляет какие-либо кажущиеся осмысленными требования, урезающие это слишком обширное множество априори допустимых мер. При первой возможности статистик старается тестировать добавленные требования и с легкостью отказывается от них, если обнаруживает, что эмпирические данные его к тому вынуждают (правда после этого ему приходится, иногда в тяжелых муках, изобретать альтернативную постановку задачи).

Главное здесь в том, что исследователь пытается, опираясь на статистические данные, решить, какая из априори допустимых возможностей реализована "в жизни"("в природе", "в обществе", "на финансовом рынке"...). Во многих случаях полное исследование явления не входит в задачу статистика, и он интересуется лишь допустимыми вероятностными мерами на более узкой алгебре событий — алгебре, порожденной конечным набором случайных величин (доступных ему наблюдений). Другими словами, он интересуется совместным распределением вероятностей для этих случайных величин и не касается других величин (и вероятностей), относящихся к тому же случайному явлению.

Мы приближаемся к очень важному обсуждению: как формулируются типичные статистические задачи, и каковы отличия прикладной статистики от математической (теоретической). Оговоримся сразу, что подобное обсуждение, помещенное в самом начале, должно рассматриваться как сугубо предварительное, не претендующее на полную ясность. Может быть, читателю будет полезно иногда возвращаться к этому параграфу по мере изучения последующих глав.

Удобно выделить два существенных этапа исследования: "от статистических данных к статистической модели"и "от статистической модели к статистическому выводу".

Построение модели статистических данных, а также (по крайней мере, иногда) и модели всего изучаемого явления — прикладная часть исследования. Во многих случаях приходится углубляться в содержательный предметный анализ явления и выходить за рамки

собственно статистики. В применении к социально-экономической проблематике такой анализ составляет ядро отдельной научной дисциплины, называемой **эконометрикой**.

Модель статистических данных, говоря упрощенно, задает исследователю алгебру событий и совокупность априори допустимых вероятностных мер на ней. Статистическая практика показывает, что в процессе работы эта модель может (часто неоднократно) модифицироваться, сравниваться с альтернативными моделями, тестироваться разнообразными способами, пока не накопится достаточная уверенность в ее "адекватности". Слово "адекватность"мы заключаем в кавычки, поскольку в серьезных задачах всегда остается тень сомнения.

Для окончательного выбора модели нет четких правил — это скорее искусство статистика, чем наука. Такое положение дел вполне согласуется с тем обстоятельством, что статистические выводы (см. ниже) практически не бывают абсолютно надежными, а умение сомневаться (**разумно** сомневаться) — первостепенная черта статистика (и эконометриста), как исследователя.

В определенные моменты у исследователя возникает ощущение, что текущая модель заслуживает того, чтобы в ее рамках заняться получением статистических выводов (с возможным возвратом после этого к обсуждению модели). Правила перехода (в рамках фиксированной модели) от исходных данных к статистическим выводам — иногда они называются статистическими решающими правилами (decision rules или statistical inference procedures; в частных задачах используются и более узкие термины, см. ниже) — в центре теоретической части исследования. Их нужно построить, обосновать, изучить, сравнить с альтернативными правилами и т.д. и, в конце концов, применить к конкретным наборам чисел (последнее, впрочем, уже не теория).

Затем статистику целесообразно приостановиться, оглядеться вокруг и осознать полученные выводы. Только после этого имеет смысл планировать конкретные дальнейшие действия. Иногда по пословице: "Пировали — веселились, подсчитали — прослезились".

Поговорим теперь немного о крайних точках статистического исследования — о данных и о выводах.

Как уже упоминалось в предыдущем параграфе, статистические данные как числа — это "реализовавшиеся значения"случайных

величин. Сами эти случайные величины представляют тем самым теоретический конструкт статистических данных. Детализируем обозначения и терминологию, стараясь не слишком отклоняться от традиционных и не забывая об аккуратности и здравом смысле.

Обычно, хотя и не всегда, статистические данные естественным образом разделены на части, отвечающие отдельным наблюдениям. Такие части мы будем выделять в наших обозначениях индексом, например, наблюдения $X_1, X_2, \ldots, X_T$. Каждое наблюдение трактуется как случайная величина (в простейшем случае — одномерная) или ее реализовавшееся значение. Обычно из контекста видно, какое из двух толкований имеется в виду. В редких случаях, когда оба толкования используются в одной формуле, реализовавшееся значение мы будем отмечать дополнительным индексом "эмп."(эмпирическое) или"эксп."(экспериментальное). Так, выражение

$$\mathsf{P}(X_1 = X_{1,\text{эмп.}})$$

следует понимать как вероятность того, что случайная величина X_1 примет значение $X_{1,\text{эмп.}}$ Наряду с подобными выражениями будут употребляться и более короткие, вида $\mathsf{P}(X_1 = x)$. Здесь буквой x обозначено одно из **возможных** значений случайной величины X_1, которому **не** приписывается роль **реализовавшегося**.

Совокупность наблюдений обычно линейно упорядочена в виде последовательности. При этом номер наблюдения чаще всего имеет одно из двух толкований — либо момент времени, либо номер объекта (скажем, фирмы) из совокупности одновременно рассматриваемых объектов. В первом случае последовательность наблюдений называется time series (временной ряд), а во втором — cross-section (общепринятого русского эквивалента нет, один из вариантов перевода — пространственные данные). Иногда это различие удобно подчеркнуть обозначением индекса: $t = 1, \ldots T$ или $i = 1, \ldots N$. В отдельных задачах встречаются "двумерные"массивы данных X_{it} — так называемые панельные данные (panel data).

Следуя установившейся традиции (о ее происхождении см. предыдущий параграф), мы иногда будем называть последовательность наблюдений выборкой, а если соответствующие случайные величины независимы и одинаково распределены (independent identically distributed, сокращенно iid или IID) — повторной выборкой. При этом никакой "генеральной совокупности"в общем случае иметь в виду

не следует. Характеристики случайных величин, составляющих выборку (распределения вероятностей, математические ожидания, дисперсии, ковариации и т.д.), мы будем называть теоретическими (в англоязычных текстах можно встретить прилагательное populational) характеристиками, в противовес эмпирическим, о которых пойдет речь в следующем параграфе.

В принципе, весь набор статистических данных можно рассматривать как одно (многомерное) наблюдение, но это редко бывает удобно — подразделение на естественные части дает дополнительную структуру набора данных, которая часто отражается в структуре априори допустимых вероятностных мер (повторная выборка — типичный пример: каждая априори допустимая мера — произведение (одинаковых) распределений отдельных наблюдений).

По характеру множества априори допустимых мер можно выделить параметрические и непараметрические модели. Четкой грани между ними иногда нет, но в целом обычно считается, что в параметрической модели совокупность априори допустимых мер определяется конечным набором числовых параметров, различающих эти меры. Например, совокупность одномерных нормальных распределений $\mathbf{N}(a, \sigma^2)$ задается двумя параметрами — математическим ожиданием a и дисперсией σ^2. Фиксация этих двух параметров однозначно определяет закон распределения. Напротив, совокупность всевозможных одномерных распределений с конечными математическим ожиданием и дисперсией параметрической считать не следует, т.к. фиксация этих характеристик еще не задает закон распределения — возможны совершенно разные распределения с одинаковыми средними значениями и дисперсиями.

Завершая пока разговор о статистических данных, отметим еще, что сама принимаемая нами концепция, согласно которой их можно трактовать как случайные величины (или их значения), требует тщательного анализа. С совершенно разных, но одинаково важных, точек зрения об этом можно прочитать в [11], [5].

Обсудим теперь возможные типы статистических выводов. Традиционно выделяют два таких типа, каждому из которых отвечает свой класс задач.

В задачах **оценивания** статистический вывод представляет собой конечный набор чисел — оцененных характеристик модели, и в этом смысле имеет арифметический характер (мы слегка упрощаем картину — к числам, разумеется должны быть сделаны надлежащие комментарии

и разъяснения). Оценивание (estimation) представляет собой процесс переработки исходных статистических данных в этот набор чисел — оценок (estimates) теоретических характеристик. Исследователь интерпретирует эти оценки как приближенные значения неизвестных ему теоретических характеристик. В английском языке имеется также не имеющий русского эквивалента термин "estimator"для правила вычисления оценки (т.е. фактически, для соответствующей формулы). Более подробно о задачах оценивания мы говорим дальше, в главе 2.

В задачах **проверки статистических гипотез** (hypotheses testing) вывод имеет логический характер — "ДА"или "НЕТ", т.е. гипотеза подтверждается или отвергается. Мы увидим дальше, что иногда одну и ту же (по существу) задачу можно сформулировать и как задачу проверки гипотезы, и как задачу оценивания, так что изложенную классификацию следует рассматривать скорее как нечто вспомогательное. Тем не менее, такая структуризация оказывается часто очень удобной и методически полезной. По естественным причинам (подробно задачи проверки гипотез мы рассматриваем в главе 4) отрицательный вывод — отвержение статистической гипотезы — на практике делается значительно чаще и в более решительной форме, чем положительный. Правило получения вывода в задачах проверки гипотез называется критерием (criterion) или тестом (test, testing procedure) проверки.

Во всех случаях статистический вывод представляет собой умозаключение исследователя, базирующееся на доступной ему информации, содержащейся в статистических данных. Такая информация заведомо является неполной, а основанный на ней вывод нельзя считать достоверным. Это — важнейшая особенность статистики: выводы по своей природе неточны. В задачах оценивания получаемые числа лишь приближенно соответствуют теоретическим характеристикам явления, а при проверке гипотезы потенциально можно отвергнуть ее, в то время как "на самом деле"она справедлива, или же принять, в то время как "на самом деле"она ложна. Выражение, заключенное в кавычки, подчеркивает то обстоятельство, что даже в рамках выбранной модели данных исследователь имеет дело с целым семейством априори допустимых вероятностных мер, лишь одна из которых отвечает реальной ситуации.

Неточность статистического вывода и "ущерб", возникающий от последствий неправильного вывода, можно включить в модель в

виде так называемой функции потерь, переводящей эту неточность и этот ущерб в числовую форму. Разумеется, каждый выбор функции потерь несет оттенок субъективности и открывает возможности для дискуссий. Тем не менее, функция потерь часто позволяет сравнивать между собой различные решающие правила и выбирать из них оптимальное (оптимальные). Мы будем неоднократно далее возвращаться к обсуждению проблемы оптимальности.

1.3 Эмпирическая мера, принцип соответствия и асимптотические мотивы в статистике

Формулируя свой неточный вывод, статистик, тем не менее, имеет надежду не ошибиться. Попробуем проанализировать, какие соображения позволяют ему надеяться на это, и в какой степени. Такой анализ во многом основан на рассуждениях, применимых лишь в частных случаях. Мы будем предполагать, что статистические данные образуют повторную выборку, т.е. конечную последовательность $X_1, \dots X_N$ независимых одинаково распределенных величин. Часть наших аргументов остается осмысленной и при более слабых ограничениях, но мы не будем на этом останавливаться.

Повторная выборка характеризуется распределением вероятностей $\mathcal{P}$ одного из наблюдений (разумеется, любого), совместные же вероятности восстанавливаются с использованием независимости. Каждая фиксация распределения $\mathcal{P}$ определяет тем самым одну из априори допустимых мер (обратное очевидно). При помощи набора $X_{1,\text{эмп.}}, \dots X_{N,\text{эмп.}}$ определяется эмпирическая мера $\mathcal{P}_N^*$ — дискретный закон распределения, приписывающий каждому из значений $X_{i,\text{эмп.}}$ вероятность, равную $1/N$ (вероятности этого вида, соответствующие совпадающим значениям, суммируются ("склеиваются"); так, если $X_{1,\text{эмп.}} = X_{2,\text{эмп.}}$, то этому значению приписывается вероятность $2/N$). Эмпирическая мера представляет собой некую карикатуру на закон распределения $\mathcal{P}$ и порождает принцип соответствия между теоретическими объектами (характеристиками распределения $\mathcal{P}$) и их эмпирическими аналогами. Соответствие начинается с констатации аналогичности двух разнородных объектов: самого теоретического закона $\mathcal{P}$ и эмпирической меры $\mathcal{P}_N^*$, а затем продолжается на вторичные (по отношению к $\mathcal{P}$) характеристики: если $f(\mathcal{P})$ — какая-либо

теоретическая характеристика (функционал от распределения $\mathcal{P}$), то ей ставится в соответствие аналогичная характеристика $f(\mathcal{P}_N^*)$ эмпирического распределения.

Приведем несколько типичных хорошо известных примеров (для одномерных наблюдений).

1. Эмпирический аналог математического ожидания $\mathbf{E} = \mathbf{E}_{\mathcal{P}}$:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

   Эта величина обычно называется выборочным (или эмпирическим) средним значением.

2. Эмпирический аналог дисперсии $\mathbf{V} = \mathbf{V}_{\mathcal{P}}(= \mathsf{var})$:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

   — выборочная или эмпирическая дисперсия.

3. Эмпирическая функция распределения:

$$F_N^*(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{]-\infty,x[}(X_i).$$

В приведенных формулах мы опустили, и это не случайно, дополнительный индекс "эмп."Дело в том, что все наши "эмпирические"объекты, начиная с эмпирической меры, можно понимать двояко, точно так же, как наблюдения, которые мы понимаем и как случайные величины, и как их реализовавшиеся значения. Тем самым, $\bar{X}$ можно трактовать и как случайную величину, и как число. Аналогично, $F_N^*(x)$ — и как обычную функцию числового аргумента x, и как случайную функцию того же аргумента.

Принцип соответствия говорит нам о том, что, скорее всего, эмпирические характеристики можно рассматривать как приближенные значения теоретических. Известны два подхода к более точной формулировке этой идеи.

Первый из них называется асимптотическим и связан с изучением введенного соответствия при растущем числе наблюдений, т.е. при

$N \to \infty$. Об этом подходе мы поговорим более подробно чуть ниже в этом параграфе. Второй подход можно охарактеризовать как оптимизационный — рассматриваются различные функции от выборки (часто они называются статистиками), тем или иным способом вводится мера отклонения их от интересующего исследователя теоретического объекта (этой мерой отклонения может быть, скажем, функция потерь) и, наконец, решается задача минимизации этого отклонения при фиксированном объеме выборки. Часто оказывается, что эмпирические характеристики являются решениями такой экстремальной задачи. Подобные оптимизационные соображения используются как в задачах оценивания, так и в задачах проверки гипотез, и мы более подробно будем обсуждать их дальше.

Здравый смысл подсказывает нам, что выборка большего объема должна содержать больше информации, так что основанный на ней статистический вывод окажется более точным, и потому асимптотические соображения оказываются полезными прежде всего в тех задачах, где число наблюдений принципиально может быть сделано весьма большим, а сами эти наблюдения не требуют крупных затрат (на практике последнее обстоятельство часто оказывается весьма существенным).

Напротив, оптимизация показывает, какой степени приближения (конечно, не любой) можно добиться, располагая выборкой фиксированного объема, т.е., на другом языке, как извлечь из эмпирических данных максимально возможную информацию об интересующей нас характеристике теоретического распределения вероятностей.

Асимптотическая теория включает, прежде всего, утверждения о поведении эмпирических характеристик в пределе, когда число наблюдений стремится к бесконечности. Типичными являются при этом результаты о сходимости этих эмпирических характеристик к пределу в подходящем смысле. Этого, однако, недостаточно, поскольку пользователи имеют дело с конечными выборками и делают выводы по ним. Поэтому очень важными (и, чаще всего, очень трудными для доказательства, но эта сторона медали пользователей редко интересует) являются границы погрешностей, т.е. отклонений допредельных значений от предельных. Часто подобные границы удается получить лишь для оптимальных или близких к ним решающих правил, и тогда асимптотическая задача смыкается с оптимизационной. В

следующем параграфе мы обсуждаем различные понятия сходимости, использующиеся в статистике, и формулируем простейшие результаты об этой сходимости. Более сложные утверждения асимптотического характера обсуждаются в последующих главах.

Асимптотическая теория для неодинаково распределенных наблюдений усложняется тем обстоятельством, что приходится предполагать тот или иной характер этой неодинаковости, причем подобные предположения следует согласовывать со спецификой конкретной задачи, а не с удобством математических доказательств. Мы будем затрагивать эти вопросы только по мере необходимости.

1.4 Предельные переходы в статистике

Простейший статистический объект, с которым приходится совершать предельный переход, — последовательность случайных величин, т.е. измеримых функций с **общей** областью определения Ω. Для таких последовательностей чаще всего рассматривается **сходимость по вероятности**.

Говорят, что последовательность $\{X_n\}$ числовых случайных величин сходится по вероятности к случайной величине X, если для любого положительного числа ε вероятность события $\{|X_n - X| \geq \varepsilon\}$ стремится к нулю при $n \to \infty$:

$$\lim_{n\to\infty} \mathsf{P}(|X_n - X| \geq \varepsilon) = 0.$$

Иногда это определение формулируют в терминах противоположных событий:

$$\lim_{n\to\infty} \mathsf{P}(|X_n - X| < \varepsilon) = 1.$$

В литературе по прикладной статистике и эконометрике часто используется обозначение $\mathrm{p}-\lim X_n$ для предела по вероятности. В математической литературе чаще пишут $X_n \xrightarrow{\mathsf{P}} X$ (буква P здесь является символом вероятностной меры, дающей способ вычисления вероятностей, и может при необходимости заменяться другим символом, обозначающим аналогичный объект, например, Q, μ или $\tilde{\mathbf{P}}$). Эта запись содержит, тем самым, больше информации о способе предельного перехода.

Нестрого говоря, данное выше определение означает, что с ростом n величины X_n и X постепенно сближаются ближе чем на наперед

заданное ε "с подавляющей вероятностью", т.е. на все большей и большей (хотя и зависящей от n) части своей общей области определения Ω.

Следует иметь в виду, что приведенное определение ничего не говорит о сходимости значений этих случайных величин в какой-нибудь конкретной точке $\omega \in \Omega$ их области определения. Более того, в некоторых точках (элементарных исходах) ω числовая последовательность $X_n(\omega)$ вполне может вообще не сходиться к $X(\omega)$ (или никуда не сходиться). Как это ни парадоксально, такое положение дел вполне устраивает статистиков. Напротив, сходимость последовательности случайных величин поточечно, т.е. для каждого $\omega \in \Omega$, в статистике практически не используется. Проиллюстрируем ситуацию конкретным и очень важным примером.

Пусть $\{X_n\}$ -последовательность независимых одинаково распределенных случайных величин, принимающих только значения 0 и 1, причем вероятности

$$\mathsf{P}(X_n = 1) = p$$

и

$$\mathsf{P}(X_n = 0) = q(= 1 - p)$$

не зависят от номера n. Если объявить событие $\{X_n = 1\}$ успехом в n-м испытании, мы получим последовательность испытаний Бернулли с вероятностью успеха p. Обозначим традиционным образом через S_N число успехов в первых N испытаниях (очевидно, $S_N = X_1 + \cdots + X_N$) и рассмотрим относительную частоту успеха — последовательность

$$\left\{ \frac{S_N}{N} \right\}.$$

Согласно интуитивному смыслу вероятностей, эта относительная частота должна сближаться с ростом N с вероятностью соответствующего события (успеха), т.е. с p. Тем не менее, можно указать бесчисленное множество реализаций бесконечной последовательности испытаний, для которых это не так. Если $0 < p < 1$, такими будут, например, реализация, состоящая из сплошных успехов, и реализация, состоящая из сплошных неудач (и много других, в том числе содержащие лишь конечное число успехов или неудач; читатель может предложить свои примеры подобных последовательностей). Каждой реализации отвечает по крайней мере одна точка $\omega \in \Omega$, и во всех упомянутых выше случаях

$$\frac{S_N(\omega)}{N} \not\to p \qquad \text{при} \qquad N \to \infty.$$

С другой стороны, закон больших чисел Бернулли утверждает, что S_N/N стремится к p по вероятности. Эта теорема (доказываемая в курсе теории вероятностей) исторически (около 1700 г.) была первым асимптотическим утверждением такого рода, получившим точную формулировку и полное обоснование.

Мы уже упоминали в предыдущих параграфах, что элементарные исходы, как правило, ненаблюдаемы. В обсуждаемом примере это очень наглядно видно — для определения такого исхода нужно знать бесконечную реализацию испытаний (может быть, и еще что-то), а наблюдаемы лишь конечные последовательности. Тем самым, определение сходимости по вероятности (и не только в рассматриваемом частном примере) не может войти в противоречие с эмпирическими данными — наблюдаются обычно лишь события положительной вероятности, в то время как элементарные исходы имеют нулевую вероятность.

Обсудим теперь наш пример с позиций статистики. Испытания Бернулли с неизвестной вероятностью успеха p могут в определенных ситуациях выступать в роли модели статистических данных, при этом сами эти данные образуют последовательность $X_{1,\text{эмп.}},\dots,\ X_{N,\text{эмп.}}$, состоящую из нулей и единиц. Относительная частота S_N/N по принципу соответствия может рассматриваться как оценка неизвестного параметра p (о свойствах этой оценки см. дальше). Очень важно осознавать, что одна и та же последовательность S_N/N имеет (согласно закону больших чисел) пределом разные значения p, в зависимости от способа вычисления вероятностей. Если зафиксировать ту из априори допустимых мер, которая соответствует испытаниям Бернулли с некоторым конкретным значением p, то по **этой** вероятности пределом последовательности S_N/N будет **именно это** p. Собственно здесь и кроется возможность оценить неизвестную вероятность p — типичные эмпирические данные как бы автоматически ведут себя нужным образом. Так же обстоит дело и в других задачах оценивания (более точно об этом пойдет речь дальше) —разные априори допустимые вероятностные меры имеют и асимптотически различимые множества типичных реализаций. Мы еще будем возвращаться к этому примеру по мере введения других видов предельного перехода.

Сформулируем наиболее распространенные условия, гарантирующие сходимость по вероятности.

Условие Чебышёва. Если $\{X_n\}$ — последовательность случайных величин с конечными математическими ожиданиями и дисперсиями, причем

$$\mathbf{E}X_n \to 0, \mathbf{V}X_n \to 0,$$

то $X_n \to 0$ по вероятности.

Этот результат почти сразу вытекает из неравенства Чебышёва, доказываемого в курсе теории вероятностей:

$$\mathsf{P}(|X - \mathbf{E}X| \geq \varepsilon) \leq \frac{\mathbf{V}X}{\varepsilon^2}.$$

Действительно, подставляя X_n вместо X и учитывая сходимость дисперсий $\mathbf{V}X_n$ к нулю, получаем, что $X_n - \mathbf{E}X_n \to 0$ по вероятности. Теперь, используя сходимость математических ожиданий $\mathbf{E}X_n$ к нулю, получаем искомое: $X_n \to 0$ по вероятности. Тонкости этого рассуждения, связанные с одновременным использованием сходимости по вероятности и сходимости числовых последовательностей, мы опускаем.

Закон больших чисел Хинчина ([12]). Если $\{X_n\}$ — последовательность независимых одинаково распределенных величин с конечными математическими ожиданиями, то последовательность

$$\bar{X} = \frac{X_1 + \cdots X_N}{N}$$

сходится по вероятности к общему значению этих математических ожиданий[2].

Закон больших чисел для зависимых наблюдений будет обсуждаться в приложении С.

Вторым видом предельного перехода, используемым в статистике, является предел с вероятностью 1, он же предел почти всюду или почти наверное. Соответствующее определение основывается на том обстоятельстве, что для любой последовательности случайных величин $\{X_n\}$ определена вероятность

$$\mathsf{P}\{\omega : \exists \lim X_n(\omega)\}$$

($\exists$ - квантор существования). Если эта вероятность равна 1, то говорят, что последовательность X_n сходится с вероятностью 1.

Этот вид сходимости можно равносильным образом описать так (ср. с определением сходимости по вероятности):

[2]В такой общности закон больших чисел был доказан уже в XX веке — примерно в 1925 г.

для любого положительного числа ε

$$\lim_{N\to\infty} \mathsf{P}(\exists n \geq N : |X_n - X| \geq \varepsilon) = 0.$$

Из последнего соотношения сразу же вытекает, что из сходимости почти всюду следует сходимость по вероятности.

Приведем без доказательства наиболее важное условие сходимости с вероятностью 1[3].

Усиленный закон больших чисел Колмогорова ([12]). Если X_n — последовательность независимых одинаково распределенных величин с конечными математическими ожиданиями, то последовательность

$$\bar{X} = \frac{X_1 + \cdots X_N}{N}$$

сходится с вероятностью 1 к общему значению этих математических ожиданий.

Очевидно, этот результат усиливает теорему Хинчина. Частным случаем теоремы Колмогорова является усиленный закон больших чисел Бореля для испытаний Бернулли:

$$\mathsf{P}\left(\frac{S_N}{N} \to p\right) = 1. \tag{1.1}$$

Символ P здесь является сокращением — речь идет о способе подсчета вероятностей, связанном с испытаниями Бернулли с той вероятностью успеха p, которая присутствует внутри круглых скобок, в выражении $\left(\frac{S_N}{N} \to p\right)$. Для большей строгости можно было бы включить символ p вероятности успеха в обозначение вероятностной меры и писать P_p. Подобное усложнение обозначений не следует использовать без острой необходимости (как сказал бы пользователь — и так понятно).

Написанное выше соотношение (1.1) еще выразительнее, чем в контексте сходимости по вероятности, показывает, что различные априори допустимые меры (в данном случае они характеризуются различными p) сосредоточены на реализациях с принципиально различными свойствами — с разными частотами успехов, и как раз эта особенность и позволяет делать статистические выводы.

Роль в статистике сходимости почти всюду во многом связана с тем, что из утверждения о более сильной сходимости легче извлекать

[3]И этот результат доказан в XX веке — около 1930 г.

теоретические следствия. Прямое прикладное значение этой сходимости значительно меньше, чем сходимости по вероятности.

Еще один вид предельного перехода лишь косвенно связан со случайными величинами. Это — слабая сходимость вероятностных распределений. Мы ограничимся обсуждением одномерного случая, в котором можно обойтись соответствующими функциями распределения.

Говорят, что последовательность $\{F_n\}$ функций распределения слабо сходится к функции распределения F, если для каждой точки $x \in \mathbf{R}$, в которой F непрерывна,

$$F_n(x) \to F(x).$$

"Слабость"здесь следует понимать по отношению к поточечной сходимости — не в каждой точке, а лишь в точках непрерывности предельной функции.

В этом определении вообще не фигурируют случайные величины, порождающие рассматриваемые законы распределения. Для случайных величин никакой сходимости не предполагается (формально, они могут даже иметь совершенно разные области определения), более того, в типичных для приложений случаях сходимости случайных величин и не будет. Тем не менее, условно говорят, что эти величины сходятся по распределению. Иногда слабая сходимость обозначается так: $F_n \xrightarrow{w} F$ (weak — слабый).

В статистике слабая сходимость появляется во многих так называемых предельных теоремах. Часто при этом предельный закон распределения непрерывен, а тогда слабая сходимость превращается в поточечную. Более того, можно доказать, что в этом случае (когда F непрерывна) поточечная сходимость оказывается равномерной.

Примерами предельных теорем являются центральная предельная теорема (это — собирательный термин для целого ряд сходных теорем, см. одну из них ниже), теорема Пуассона (см. следующий параграф), теорема Пирсона (см. главу 4).Сформулируем наиболее важный для статистики вариант центральной предельной теоремы — теорему Леви́, а также ее частный случай для испытаний Бернулли — интегральную теорему Муавра-Лапласа.

Теорема Леви́([12]). Пусть $\{X_n\}$ последовательность независимых одинаково распределенных случайных величин с конечными математическими ожиданиями $a = \mathbf{E}X_n$ и конечными ненулевыми дисперсиями $\sigma^2 = \mathbf{V}X_n \neq 0$. Обозначим $S_N = X_1 + \cdots + X_N$. Тогда

последовательность центрированных и нормированных сумм

$$\frac{S_N - Na}{\sigma\sqrt{N}}$$

сходится по распределению к нормальному закону, т.е. функции распределения

$$F_N(x) = \mathsf{P}\left(\frac{S_N - Na}{\sigma\sqrt{N}} < x\right)$$

слабо (а также поточечно и равномерно) сходятся к стандартной нормальной функции распределения

$$F_N(x) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-t^2/2\right) dt.$$

В частном случае испытаний Бернулли речь идет о величинах

$$\frac{S_N - Np}{\sqrt{Npq}}$$

(остальная часть формулировки сохраняется).

Несколько слов о соотношении между законом больших чисел Хинчина (или Колмогорова) и центральной предельной теоремой Леви́ (обе формулировки относятся к одинаково распределенным наблюдениям!). По закону больших чисел

$$\frac{S_N}{N} - a \to 0$$

по вероятности. В то же время согласно центральной предельной теореме

$$\sqrt{N}\left(\frac{S_N}{N} - a\right) = \sigma\frac{S_N - Na}{\sigma\sqrt{N}}$$

по распределению сходится к нормальному закону. Сравнивая эти соотношения и пренебрегая различиями между разными понятиями сходимости, можно образно сказать, что

$$\frac{S_N}{N} - a$$

сходится к нулю со скоростью, обратно пропорциональной $\sqrt{N}$. То же самое можно символически записать в виде

$$\frac{S_N}{N} - a \approx \frac{\sigma\mathbf{N}(0,1)}{\sqrt{N}} = \mathbf{N}\left(0, \frac{\sigma^2}{N}\right) \tag{1.2}$$

(здесь $\mathbf{N}(0,1)$ понимается как символ нормально распределенной величины со стандартными параметрами). В этом смысле центральная предельная теорема уточняет закон больших чисел и дает определенное представление о том, с какой точностью

$$\bar{X} = \frac{S_N}{N}$$

можно истолковать как приближенное значение для математического ожидания a (например, для вероятности успеха p в случае испытаний Бернулли).

Нормальную аппроксимацию (1.2) можно использовать для решения различных статистических задач. Выбирая типичную для многих эконометрических задач надежность 95% и пользуясь "правилом 5%", отвечающим ей, получаем

$$\mathsf{P}\left(\left|\frac{S_N}{N} - a\right| \le 1.96\frac{\sigma}{\sqrt{N}}\right) \approx 0.95$$

(приблизительность здесь происходит почти исключительно из погрешности нормальной аппроксимации (1.2); погрешностями вычислений по сравнению с ней обычно можно пренебречь).

Если дисперсия σ^2 наблюдений известна, последнее соотношение показывает (на приблизительно 95%-ом уровне надежности) точность приближенного значения (оценки) $\bar{X}$для неизвестного математического ожидания a. К сожалению, в типичных случаях σ^2 следует считать неизвестным (так называемый "мешающий"параметр). Вопрос о мешающих параметрах далее будет обсуждаться более подробно, а сейчас мы ограничимся кратким изложением частного случая испытаний Бернулли, когда $\sigma^2 = p(1-p), a = p$ и мешающего параметра фактически нет. Получается нелинейное неравенство

$$|\bar{X} - p| \le z\sqrt{\frac{p(1-p)}{N}}$$

(мы заменили выбранное ранее конкретное табличное значение 1.96 общим символом z), которое можно решить относительно p (задача сводится к квадратному неравенству) и получить равносильное двойное неравенство вида

$$p_- \le p \le p_+, \tag{1.3}$$

где $p_\pm$ выражаются через z и эмпирические данные (т.е. через N и $\bar{X}$). В параграфе 3.1 более подробно излагается практическая сторона

соответствующих вычислений. Итоговым результатом (1.3) можно воспользоваться либо для нахождения точности (на соответствующем уровне надежности) приближенного значения $\bar{X}$ для p, либо (если так сформулирована задача) для проверки гипотезы. Если гипотеза имеет вид $p = p_0$, где p_0 — гипотетическое значение вероятности, то неравенство (1.3) позволяет отвергнуть (если $p_0 \notin [p_-, p_+]$) или принять ее (в противном случае) на указанном уровне надежности.

Описанные выше манипуляции с нормальным распределением являются типичным примером рассуждения, которое можно назвать **использованием шаблона** (точнее, шаблонного распределения). В качестве такового выступает нормальный закон. Далее мы увидим, что в статистике имеется еще несколько шаблонных распределений — хи-квадрат, Стьюдента, Фишера, Колмогорова и некоторые другие. Важность шаблона определяется важностью и широтой того круга задач, которые могут быть решены с его помощью. В этом смысле нормальное распределение несомненно стоит на первом месте. В любом учебнике по математической статистике или эконометрике приводятся таблицы шаблонных распределений, а компьютерные пакеты приводят нужные табличные значения в отчетах о проделанных вычислениях.

Последний предельный переход, который мы рассмотрим в этом параграфе, связан с эмпирической мерой $\mathcal{P}_N^*$ и соответствующей функцией распределения $F_N^*(x)$. Новых определений здесь не потребуется, однако сам предельный переход оказывается чуть более сложным: следует учесть, что эмпирическая функция распределения кроме основного аргумента x зависит еще от элементарного исхода ω, т.е. фактически является функцией двух аргументов $F_N^*(x, \omega)$.

Предположим сначала, что $x \in \mathbf{R}$ зафиксировано. Тогда $\{F_N^*(x)\}$ — последовательность обычных случайных величин. Более того, это — последовательность средних арифметических. Поэтому усиленный закон больших чисел сразу же позволяет сделать вывод, что

$$F_N^*(x) \to F(x)$$

с вероятностью 1.Точно так же можно доказать, что для любого фиксированного промежутка B (или даже любого фиксированного борелевского множества)

$$\mathsf{P}(\mathcal{P}_N^*(B) \to \mathcal{P}(B)) = 1.$$

Некоторое усовершенствование этого рассуждения, которое мы не приводим, позволяет доказать более сильный результат:

Теорема Гливенко-Кантелли([1]). Для любой повторной выборки с вероятностью 1

$$\sup_x |F_N^*(x) - F(x)| \to 0, N \to \infty.$$

Таким образом, мы видим, что эмпирическая мера $\mathcal{P}_N^*$ и ее функция распределения сходятся к соответствующим теоретическим объектам. Неудивительно, что сближение (в том или ином смысле) эмпирических объектов с теоретическими можно обнаружить и для многих производных характеристик — функционалов от эмпирической меры. Это отчасти объясняет важность принципа соответствия.

1.5 Основные параметрические семейства распределений

При построении статистических и эконометрических моделей постоянно возникают разнообразные конкретные распределения вероятностей. В большинстве случаев они включаются в обширные семейства, зависящие от одного или нескольких параметров. Мы сейчас перечислим несколько наиболее важных семейств распределений, которые будут далее использоваться в качестве примеров, и приведем их основные характеристики. Для некоторых семейств мы укажем распространенные обозначения (одно из них, для нормального распределения, уже фигурировало в предыдущих параграфах). Знак принадлежности $\in$ будет применяться для фиксации того обстоятельства, что случайная величина имеет то или иное распределение (например, запись $X \in \mathbf{N}$ будет означать, что случайная величина X имеет нормальное распределение).

I. *Двухпараметрическое семейство нормальных распределений* $\mathbf{N}(a, \sigma^2)$.

Стандартное нормальное распределение задается плотностью

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), x \in \mathbf{R},$$

и функцией распределения

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt.$$

Плотность общего нормального распределения с параметрами $a \in \mathbf{R}$, $\sigma > 0$ выражается через стандартную нормальную плотность φ с помощью преобразований сдвига и масштаба:

$$p(x) = \frac{1}{\sigma}\varphi\left(\frac{x-a}{\sigma}\right).$$

Аналогично обстоит дело и с функцией распределения $F(x)$:

$$F(x) = \Phi\left(\frac{x-a}{\sigma}\right)$$

Поэтому, если $X \in \mathbf{N}(a, \sigma^2)$, то $\frac{X-a}{\sigma} \in \mathbf{N}(0,1)$. Параметр сдвига a задает математическое ожидание, а параметр масштаба σ — стандартное отклонение (квадратный корень из дисперсии) нормального распределения: если $X \in \mathbf{N}(a, \sigma^2)$, то

$$\mathbf{E}X = a, \mathbf{V} = \sigma^2.$$

Моменты нормального распределения более высоких порядков выражаются через основные параметры. Центральные моменты (они, очевидно, не зависят от сдвига) имеют вид:

$$\mathbf{E}(X-a)^{2k+1} = 0, k = 0, 1, 2, \ldots,$$

$$\mathbf{E}(X-a)^{2k} = \frac{(2k)!}{k!}\left(\frac{\sigma^2}{2}\right)^k = (2k-1)!!\sigma^{2k}, k = 1, 2, \ldots.$$

Начальные моменты можно выразить через центральные при помощи формулы бинома Ньютона:

$$X^k = [(X-a)+a]^k = \sum_{i=0}^{k} C_k^i (X-a)^i a^{k-i}.$$

Вычисляя математическое ожидание правой части, получаем требуемое выражение для начальных моментов. Приведем еще три "табличных"вероятности, относящиеся к нормальному распределению. Эти вероятности постоянно используются в иллюстративных примерах. В формулах предполагается, что $X \in \mathbf{N}(a, \sigma^2)$.

Это

$$\mathsf{P}(|X-a| > 1.96\sigma) \approx 0.05$$

(правило "пяти процентов");

$$\mathsf{P}(X-a > 1.65\sigma) \approx 0.05$$

(одностороннее правило "пяти процентов");

$$\mathsf{P}(|X - a| > 3\sigma) \approx 0.9973$$

(правило "трех сигма").

Иногда удобно вырожденное распределение (т.е. распределение константы a) считать нормальным распределением с $\sigma = 0$:

$$a \in \mathbf{N}(a, 0)$$

.

II. *Двухпараметрическое семейство гамма-распределений* $\Gamma(\alpha, p)$.

Плотность гамма-распределения сосредоточена на положительной полуоси $]0, \infty[$ и задается формулой

$$p(x) = \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, x > 0.$$

Параметр $\alpha > 0$ имеет (обратный) масштабный смысл: если $X \in \Gamma(\alpha, p)$, то $\alpha X \in \Gamma(1, p)$. Параметр $p > 0$ иногда называется параметром формы. О свойствах гамма-функции Эйлера $\Gamma(p)$ см. приложение А. В том же приложении объясняются формулы для моментов гамма-распределения:

$$\mathbf{E}X = \frac{p}{\alpha}, \mathbf{V}X = \frac{p}{\alpha^2},$$

$$\mathbf{E}(X^k) = \alpha^{-k} p(p+1) \cdots (p+k-1).$$

Частным случаем гамма-распределения при $p = 1$ является показательное распределение с плотностью

$$p(x) = \alpha e^{-\alpha x}, x > 0.$$

Другой частный случай —

$$\Gamma(1/2, n/2)$$

— называется распределением хи-квадрат с n степенями свободы и обозначается χ^2_n. Это распределение обычно возникает в качестве шаблонного.

Иногда бывает полезно включить в определение гамма-семейства в качестве третьего параметра сдвиг.

III. *Семейство бета-распределений* $B(p_1, p_2)$.

Плотность бета-распределения сосредоточена на промежутке $\langle 0, 1\rangle$ и задается формулой

$$p(x) = \frac{\Gamma(p_1, p_2)}{\Gamma(p_1)\Gamma(p_2)} x^{p_1-1}(1-x)^{p_2-1}, 0 < x < 1.$$

Оба параметра p_1 и p_2 предполагаются положительными. Значения плотности в концевых точках 0 и 1 не имеют значения (плотность всегда определяется с точностью до почти всюду), поэтому мы обозначили промежуток треугольными скобками, не уточняя, включены ли в него концы.

Формулы для моментов бета-распределения

$$\mathbf{E}X = \frac{p_1}{p_1 + p_2}, \mathbf{V}X = \frac{p_1 p_2}{(p_1 + p_2)^2(p_1 + p_2 + 1)}$$

также обсуждаются в приложении А.

Частным случаем бета-распределения при $p_1 = p_2 = 1$ является равномерное распределение с плотностью

$$p(x) = 1, 0 < x < 1.$$

Это семейство можно расширить, делая сдвиг и масштабное преобразование. В частности, так получается двухпараметрическое семейство равномерных распределений на $\langle a, b\rangle$:

$$p(x) = \frac{1}{b-a}, a < x < b.$$

Для него $\mathbf{E}X = \frac{a+b}{2}, \mathbf{V}X = \frac{(b-a)^2}{12}$.

IV. *Семейство распределений Бернулли* $B_n(p)$.

$(n = 1, 2, \ldots; 0 \le p \le 1)$.

Распределение $B_n(p)$, известное также и под названием биномиального, дискретно и сосредоточено в точках $0, 1, \ldots n$. Соответствующие вероятности задаются формулой Бернулли:

$$P_n(k, p) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, \ldots n.$$

При $p = 0$ и $p = 1$ получаем вырожденные распределения в точках 0 и n соответственно.

Как известно, число успехов S_n в n испытаниях Бернулли с вероятностью успеха p имеет распределение $B_n(p)$, при этом

$$\mathbf{E}S_n = np, \mathbf{V}S_n = np(1-p).$$

V. *Семейство распределений Пуассона* $\Pi(\lambda)$ $(\lambda \geq 0)$.

Распределение $\Pi(\lambda)$ дискретно и сосредоточено на множестве $\mathbf{Z}_+ = \{0, 1, \dots\}$ целых неотрицательных чисел. Соответствующие вероятности задаются формулой

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, \dots$$

При $\lambda = 0$ получаем вырожденное распределение в точке 0.

Теорема Пуассона утверждает, что распределение Бернулли $B_n(p)$ превращается в распределение Пуассона $\Pi(\lambda)$, если $n \to \infty$ и $p \to 0$, причем $np \to \lambda$. Это отчасти объясняет, почему пуассоновская случайная величина имеет математическое ожидание и дисперсию, равные λ.

VI. *Семейство логнормальных распределений.*

Говорят, что случайная величина X имеет логнормальное распределение, если $\ln X$ имеет нормальное распределение. Плотность логнормального распределения имеет вид

$$f(x) = \frac{1}{x} p(\ln x), x > 0,$$

где $p(\cdot)$ — плотность нормального распределения. Соответствующее математическое ожидание равно

$$\mathbf{E}X = \exp(a + \frac{\sigma^2}{2}).$$

Формулу для дисперсии мы не приводим.

Логарифмическое преобразование часто используется в связи с дисконтированием денежных потоков.

Многомерное нормальное распределение обсуждается в Приложении В, а многомерный аналог распределения Бернулли — полиномиальное распределение — в параграфе 4.7.

1.6 Свертки распределений и их роль в статистике

Пусть X_1 и X_2 — независимые случайные величины, имеющие распределения $\mathcal{P}_1$ и $\mathcal{P}_2$ соответственно. Тогда распределение их суммы $X_1 + X_2$ называется сверткой распределений $\mathcal{P}_1$ и $\mathcal{P}_2$ и обозначается $\mathcal{P}_1 * \mathcal{P}_2$.

Если распределения $\mathcal{P}_1$ и $\mathcal{P}_2$ непрерывны и заданы своими плотностями p_1 и p_2, то их свертка — также непрерывное распределение,

имеющее плотность

$$p(z) = (p_1 * p_2)(z) = \int_{\mathbf{R}} p_1(z-y)p_2(y)dy = \int_{\mathbf{R}} p_1(x)p_2(z-x)dx.$$

Аналогичная формула справедлива и для дискретных величин. Выпишем ее в наиболее существенном случае, когда X_1 и X_2 — целочисленные величины:

$$\mathsf{P}(X_1 + X_2 = n) = \sum_k \mathsf{P}(X_1 = k)\mathsf{P}(X_2 = n-k).$$

Роль сверток в статистике определяется двумя взаимосвязанными обстоятельствами. Во-первых, суммирование независимых величин, образующих выборку, или как-то связанных с ней, — операция, постоянно присутствующая в большинстве рассуждений. Во-вторых, некоторые основные параметрические семейства распределений "выдерживают"свертку, воспроизводятся при сложении независимых величин (точные формулировки приведены ниже). Такая воспроизводимость сильно облегчает изучение многих классических статистических моделей и уменьшает количество возникающих при этом шаблонов.

Начнем с нормального распределения, которое воспроизводится по обоим параметрам:

$$\mathbf{N}(a_1, \sigma_1^2) * \mathbf{N}(a_2, \sigma_2^2) = \mathbf{N}(a_1 + a_2, \sigma_1^2 + \sigma_2^2).$$

Средние значения и дисперсии, как всегда при сложении независимых величин, складываются, а потому главным новым обстоятельством здесь является воспроизведение свойства нормальности.

Перейдем теперь к гамма-семейству. Для него имеется только частичная воспроизводимость — воспроизводимость по параметру формы p:

$$\Gamma(\alpha, p_1) * \Gamma(\alpha, p_2) = \Gamma(\alpha, p_1 + p_2).$$

В частности,

$$\chi^2_{n_1} * \chi^2_{n_2} = \chi^2_{n_1+n_2}.$$

Отсюда, как мы сейчас увидим, вытекает, что χ^2_n — распределение суммы квадратов n независимых величин, распределенных по стандартному нормальному закону: если $X_1, \ldots, X_n \in \mathbf{N}(0,1)$ — независимые случайные величины, то $X_1^2 + \cdots + X_n^2 \in \chi^2_n$. Ввиду свойства

воспроизводимости, это следствие достаточно доказать при $n = 1$, что делается прямым счетом: при $z > 0$

$$\mathsf{P}(X_1^2 < z) = \mathsf{P}(-\sqrt{z} < X_1 < \sqrt{z}) = \Phi(\sqrt{z}) - \Phi(-\sqrt{z}),$$

так что плотность величины X_1^2 записывается как

$$\frac{d}{dz}(\Phi(\sqrt{z}) - \Phi(-\sqrt{z})) = \frac{1}{\sqrt{z}}\phi(\sqrt{z}) = \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2}.$$

Последнее выражение является плотностью распределения

$$\Gamma(1/2, 1/2) = \chi_1^2,$$

правда, записанной без использования гамма-функции.

Центральная предельная теорема в форме Леви́ (см. предыдущий параграф), примененная к указанной выше последовательности $X_1^2, X_2^2, \ldots$, утверждает, что центрированное и нормированное распределение χ_n^2 слабо сходится при $n \to \infty$ к стандартному нормальному закону. Это свойство можно символически записать в виде аппроксимации

$$\chi_n^2 \approx \mathbf{N}(n, 2n).$$

Аналогично, при больших p

$$\Gamma(\alpha, p) \approx \mathbf{N}(p/\alpha, p/\alpha^2).$$

Заметим, впрочем, что в литературе приводятся и другие, как утверждается, более точные, способы аппроксимации χ_n^2 нормальным законом, например,

$$\mathsf{P}(\chi_n^2 < x) \approx \Phi(\sqrt{2x} - \sqrt{2n-1}).$$

(см. [1], [19])

Для распределений Бернулли и Пуассона также имеется частичная воспроизводимость:

$$B_{n_1}(p) * B_{n_2}(p) = B_{n_1+n_2}(p),$$

$$\Pi(\lambda_1) * \Pi(\lambda_2) = \Pi(\lambda_1 + \lambda_2)$$

(первую из этих формул легко истолковать в терминах испытаний Бернулли).

Глава 2

Теория оценивания

Мы начинаем главу с краткого описания основных понятий и разбора простейших примеров. Во второй части главы излагаются более сложные вопросы, включая теорию достаточных статистик и асимптотическую эффективность.

2.1 Точечные оценки. Состоятельность и эффективность

Как уже упоминалось в параграфе 1.2, оцениванию подлежат параметры теоретического распределения вероятностей. В параметрической модели описание теоретического распределения включает некоторый (конечный) набор "базисных"параметров, задание которых однозначно определяет это распределение. Оценивать при этом можно как сами эти базисные параметры, так и функции от них (это зависит от цели исследования). В непараметрической модели параметром (лучше сказать, оцениваемым функционалом) можно считать любую числовую характеристику теоретического распределения, интересующую статистика.

Во всех случаях точечной оценкой (estimator) некоторого параметра или функционала θ может быть объявлена статистика, т.е. функция от выборки, предлагаемая в качестве правила вычисления приближенного значения этого параметра. Разумеется, не любая статистика пригодна для этого. Простейшее требование к оценке — состоятельность — формулируется на асимптотическом языке.

Оценка $\hat{\theta}$ параметра θ называется **состоятельной**, если она стремится к нему по вероятности при $N \to \infty$.

Это определение требует некоторых разъяснений. Прежде всего отметим, что $\hat{\theta}$ следует понимать как функцию с довольно сложной областью определения — выборку призвольного объема N статистика $\hat{\theta}$ "перерабатывает"в приближенное значение параметра. Поэтому ее область определения (как функции от выборки) состоит из "одномерной"части, на которой задана функция одного аргумента $\hat{\theta}_1(X_1)$, "двумерной"части, на которой задана функция двух аргументов $\hat{\theta}_2(X_1, X_2)$ и т.д. (кавычки поставлены по той причине, что сами наблюдения могут быть и многомерными).

Кроме того, оценку, как функцию от случайных величин, можно понимать и как случайную величину (суперпозицию функции, о которой шла речь в предыдущем абзаце, и выборки).

С учетом сделанного разъяснения определение состоятельности означает, что последовательность случайных величин

$$\hat{\theta}_1 = \hat{\theta}_1(X_1), \hat{\theta}_2 = \hat{\theta}_2(X_1, X_2), \cdots$$

по вероятности сходится к θ. Остается уточнить, по какой вероятности, или, точнее, по каким вероятностям. Предварительный разговор об этом уже шел в параграфе 1.4. Имеется в виду следующее. Каждому значению функционала θ в рассматриваемой модели отвечает некоторая совокупность априори допустимых (в качестве теоретического распределения) вероятностных мер, имеющих именно это значение параметра. Состоятельность означает, что $\hat{\theta}_N \to \theta$ **по каждой** из этих вероятностей.

Приведем полезный пример (статистика $\bar{X}$), которому можно придать как параметрическую, так и непараметрическую форму.

Первый параметрический вариант. Для выборки, имеющей распределение Пуассона $\Pi(\lambda)$, статистика $\bar{X}$ (выборочное среднее значение) является состоятельной оценкой параметра λ. Это утверждение вытекает из закона больших чисел Хинчина (см. параграф 1.4). При этом каждому значению λ отвечает единственная априори допустимая мера, порожденная указанным распределением Пуассона $\Pi(\lambda)$ (произведение распределений Пуассона, отвечающих отдельным наблюдениям[1]), и $\bar{X} \to \lambda$ по ней.

Второй параметрический вариант. Для выборки, имеющей нормальное распределение $\mathbf{N}(a, \sigma^2)$ с неизвестными параметрами,

[1]Поскольку объем выборки N растет до бесконечности, удобно рассматривать априори допустимые меры на сигма-алгебре, порожденной бесконечной последовательностью наблюдений. Нам не потребуются детали их определения.

статистика $\bar{X}$ (снова по теореме Хинчина) является состоятельной оценкой параметра a. При этом каждому значению a отвечает однопараметрическое семейство априори допустимых мер, порожденных нормальными распределениями с этим a и различными дисперсиями σ^2. По каждой из этих вероятностей $\bar{X}$ стремится к a.

Непараметрический вариант. Для выборки с конечным математическим ожиданием $\mathbf{E}$ статистика $\bar{X}$ (все по той же теореме Хинчина) является состоятельной оценкой математического ожидания. При этом каждому значению математического ожидания отвечает обширное (непараметрическое) семейство априори допустимых мер — всевозможные распределения вероятностей P (произведения распределений $\mathcal{P}$ отдельных наблюдений[2]), дающие это математическое ожидание ($\mathbf{E}_{\mathcal{P}} = \mathbf{E}$). По отношению к каждой из них

$$\bar{X} \xrightarrow{\mathsf{P}} \mathbf{E}.$$

Таким образом, свойство состоятельности означает, что рассматриваемая оценка "приспособлена"именно к тому параметру (функционалу), который мы желаем с ее помощью оценивать.

Иногда оказывается полезным несколько более узкое определение. Оценка $\hat{\theta}$ называется сильно состоятельной, если она сходится к оцениваемому параметру с вероятностью 1. Все приведенные выше комментарии к определению состоятельности переносятся с минимальными изменениями и на случай сильной состоятельности.

Закон больших чисел является одним из основных способов проверки состоятельности той или иной оценки. Проиллюстрируем его использование на более сложном примере эмпирической (=выборочной) дисперсии S^2 повторной выборки. Точнее, докажем, что S^2 — состоятельная оценка дисперсии теоретического распределения (к априори допустимым при этом относятся только произведения одинаковых распределений с конечной дисперсией). Для этого заметим, что

$$S^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2$$

(эта формула является специализацией на случай эмпирического распределения $\mathcal{P}_N^*$ общей формулы $\mathbf{V}X = \mathbf{E}(X^2) - (\mathbf{E}X)^2$). Первое

[2]См. предыдущую сноску.

слагаемое

$$\frac{1}{N}\sum_{i=1}^{N} X_i^2$$

сходится по вероятности к общему значению вторых моментов $\mathbf{E}(X_i^2)$ (теорема Хинчина для последовательности квадратов $X_1^2,\ X_2^2, \dots$). В то же время $\bar{X}$ сходится по вероятности к общему значению математических ожиданий $\mathbf{E}X_i$. Пользуясь стандартными формулами для предела произведения и разности (в случае предела по вероятности они справедливы, хотя и нуждаются в специальном доказательстве), заключаем, что S^2 сходится по вероятности к общему значению выражений $\mathbf{E}(X_i^2) - (\mathbf{E}X_i)^2 = \mathbf{V}X_i$. Это и есть состоятельность эмпирической дисперсии.

Ниже сформулирован еще один полезный результат, позволяющий устанавливать состоятельность, не используя прямо исходное определение. Для его формулировки нам потребуется одно важное понятие, которое в дальнейшем будет многократно использоваться.

Оценка $\hat{\theta}$ параметра или функционала θ называется **несмещенной**, если $\mathbf{E}\hat{\theta} = \theta$. Более общим образом, она называется **асимптотически** несмещенной, если $\mathbf{E}\hat{\theta} \to \theta$ при $N \to \infty$.

Аналогично тому, что подразумевалось в определении состоятельности, имеется в виду, что равенство (соотв. сходимость) справедливо при любом выборе априори допустимой меры с данным значением параметра или функционала (именно по априори допустимой мере вычисляется математическое ожидание)[3].

Величина $b(\theta) = \mathbf{E}\hat{\theta} - \theta$ называется смещением оценки $\hat{\theta}$. Если функционал θ не определяет теоретическое распределение единственным образом, то смещение может зависеть не только от θ, но и от выбора априори допустимого распределения.

Состоятельные оценки, как правило, являются асимптотически несмещенными:

Достаточные условия состоятельности. Предположим, что оценка $\hat{\theta}$ является асимптотически несмещенной и что $\mathbf{V}\hat{\theta} \to 0$ при $N \to \infty$. Тогда $\hat{\theta}$ — состоятельная оценка параметра θ.

Для получения этих условий воспользуемся тем же приемом, что и в доказательстве простейших вариантов закона больших чисел —

[3] В дальнейшем подобные комментарии, как правило, будут опускаться.

неравенством Чебышёва:

$$\mathsf{P}(|\hat{\theta} - \mathbf{E}\hat{\theta}| \geq \varepsilon) \leq \frac{\mathbf{V}\hat{\theta}}{\varepsilon^2},$$

правая часть которого, по предположению, стремится к нулю при $N \to \infty$. Для несмещенной оценки левую часть неравенства можно заменить на $\mathsf{P}(|\hat{\theta} - \theta| \geq \varepsilon)$, откуда сразу следует состоятельность (надо воспользоваться определением сходимости по вероятности). Если же оценка только асимптотически несмещенная, требуется незначительное усложнение рассуждения. При достаточно больших N по определению предела числовой последовательности

$$|\mathbf{E}\hat{\theta} - \theta| \leq \varepsilon/2.$$

Поэтому неравенство $|\hat{\theta} - \theta| \geq \varepsilon$ влечет $|\hat{\theta} - \mathbf{E}\hat{\theta}| \geq \varepsilon/2$, так что

$$\mathsf{P}(|\hat{\theta} - \theta| \geq \varepsilon) \leq \mathsf{P}(|\hat{\theta} - \mathbf{E}\hat{\theta}| \geq \varepsilon/2) \leq \frac{\mathbf{V}\hat{\theta}}{(\varepsilon/2)^2},$$

а последнее выражение стремится к нулю.

Для задач с фиксированным объемом выборки (обычно такая формулировка возникает в случаях, когда большой объем выборки по тем или иным причинам не может быть получен), свойство состоятельности почти полностью теряет свое значение, и на первый план выступает тот ущерб, который возникает от расхождения оценки и оцениваемого параметра. Чаще всего этот ущерб измеряют средним значением функции потерь. При этом со времен Гаусса (начало XIX века) принято считать, что наиболее естественной является квадратичная функция потерь. Эта трактовка приводит к определению сравнительной эффективности. Говорят, что оценка $\hat{\theta}$ **эффективнее** оценки $\tilde{\theta}$, если

$$\mathbf{E}(\hat{\theta} - \theta)^2 \leq \mathbf{E}(\tilde{\theta} - \theta)^2.$$

Ввиду важности этого и последующих определений, подчеркнем еще раз, что символ $\mathbf{E}$ относится к априори допустимым мерам, и что неравенство должно выполняться для каждого значения θ и для каждой такой меры. Отсюда сразу же следует, что две оценки могут оказаться несравнимыми. Например, оценка $\tilde{\theta} \equiv \theta_0$, где θ_0 — конкретное возможное значение параметра, будет несравнима с оценкой $\hat{\theta} \equiv \theta_0'$ — другое конкретное значение параметра. Разумеется, приведенный пример малосодержателен, однако саму возможность несравнимости

он иллюстрирует крайне выразительно. Нетрудно выделить и причину этого явления. Обе оценки несостоятельны и смещены (за исключением случая, когда одна из них совпадает с истинным значением параметра, но надеяться на это — уже не статистический подход, а гадание). Если сузить каким-нибудь содержательным образом класс рассматриваемых оценок, то в пределах этого класса оценки могут оказаться сравнимыми — от сравнительной эффективности иногда удается перейти к "абсолютной": оценка $\hat{\theta}$ называется эффективной в данном классе оценок, если она эффективнее любой другой оценки этого класса.

Примерами таких содержательных классов оценок (см. далее параграф 4) в параметрических моделях являются K_0 — класс несмещенных оценок и K_b — класс оценок с фиксированным смещением $b = b_N(\theta)$. Еще один подобный класс — класс эквивариантных оценок — будет определен позже, в параграфе 9.

В непараметрических моделях класс априори допустимых теоретических распределений скорее всего окажется слишком широким, и не будет существовать эффективной оценки в K_0. Например, для нормального распределения эффективной несмещенной оценкой математического ожидания является $\bar{X}$, а для математического ожидания равномерного распределения существуют и более эффективные оценки (мы будем обсуждать эти примеры в параграфе 3). Поэтому для непараметрической модели, допускающей оба эти распределения, эффективной несмещенной оценки не существует.

Некоторым расширением свойства эффективности является асимптотическая эффективность. Оценка $\hat{\theta}$ называется асимптотически эффективной в данном классе K, если для любой другой оценки $\tilde{\theta}$ этого класса

$$\overline{\lim} \frac{\mathbf{E}(\hat{\theta} - \theta)^2}{\mathbf{E}(\tilde{\theta} - \theta)^2} \leq 1.$$

Символ верхнего предела использован по той причине, что для некоторых оценок настоящий предел может не существовать. Как правило, при такой асимптотической трактовке эффективности класс K состоит только из состоятельных и асимптотически несмещенных оценок (может быть, с какими-нибудь дополнительными ограничениями). Более подробно об асимптотической эффективности мы будем говорить в параграфе 8.

Оценка, являющаяся состоятельной, несмещенной и эффективной (в классе K_0), в большинстве случаев рассматривается как наилучший

рецепт оценивания. К сожалению, далеко не всегда ее удается найти. Более того, вполне может оказаться (см. [1]), что для данного параметра вообще не существует несмещенных оценок. Собственно, подобные казусы и объясняют, в значительной степени, введение расширенных — асимптотических — вариантов несмещенности и эффективности.

2.2 Общие принципы построения оценок

В первую очередь следует назвать уже упоминавшийся в первой главе принцип соответствия и основанные на нем процедуры подстановки. Напомним (см. параграф 1.3), что этот принцип подчеркивает аналогию между функционалами $f(\mathcal{P})$ от теоретического распределения $\mathcal{P}$ и их выборочными вариантами — функционалами $f(\mathcal{P}_N^*)$ от эмпирического распределения, которые, собственно, и предлагаются в качестве оценок. При необходимости этот принцип может слегка модифицироваться — подстраиваться под специфику задачи. Например, при оценивании плотности теоретического распределения может потребоваться предварительное "сглаживание"эмпирического распределения. Другая возможная модификация обсуждается чуть ниже.

Рассмотрим наиболее известную и популярную реализацию сформулированной выше идеи — метод моментов и его обобщения. Для простоты рассмотрим сначала параметрическую модель с единственным одномерным параметром $\theta \in \Theta \subset \mathbb{R}$, т.е. будем считать, что имеется однопараметрическое семейство $\mathcal{P}_\theta$ априори допустимых мер. На всякий случай полезно подчеркнуть, что при этом подразумевается обратимость параметризации — разным значениям $\theta \in \Theta$ отвечают разные распределения $\mathcal{P}_\theta$. Символом $\mathbf{E}_\theta$ при необходимости будем обозначать соответствующее математическое ожидание. Как известно, моментом порядка k случайной величины X называется математическое ожидание $\mathbf{E}(X^k)$. В современной статистической и эконометрической литературе (см. [1] и [19]) принята более широкая трактовка моментов — любое выражение вида $\mathbf{E}g(X)$, где g — какая-нибудь подходящая функция, называется моментом случайной величины X. Выберем g так, чтобы "моментная функция"

$$m(\theta) = \mathbf{E}_\theta g(X_1)$$

была определена при всех $\theta \in \Theta$ и обратима, так что

$$\theta = m^{-1}(\mathbf{E}_\theta g(X_1)).$$

Метод моментов предлагает оценивать моментную функцию (как и положено для математического ожидания по принципу соответствия) эмпирическим средним

$$\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g(X_i),$$

а сам параметр θ — соответствующим прообразом

$$\hat{\theta} = m^{-1}(\bar{g}) \tag{2.1}$$

Согласно закону больших чисел, $\bar{g}$ — состоятельная оценка моментной функции. Поэтому, в предположении, что m^{-1} непрерывна, $\hat{\theta}$ — состоятельная оценка θ (даже сильно состоятельная).

Если $\bar{g}$ не попадает в область определения $m(\Theta)$ обратной функции m^{-1}, формулу (2.1) следует модифицировать. Например, можно заменить в ней $\bar{g}$ на ближайшую к ней точку множества $m(\Theta)$.

Очевидно, что метод моментов дает обширное множество оценок параметра θ — при разных g (примеры мы рассмотрим в следующем параграфе).

В более общем случае r-мерного векторного параметра θ конструкция оценки (2.1) практически полностью сохраняется. Единственное изменение — в том, что функция g и моментная функция $m(\theta)$ также должны браться векторнозначными размерности r.

Имеются обобщения метода моментов, пригодные и в непараметрических моделях. Пусть $\{\mathcal{P}\}$ — совокупность априори допустимых теоретических распределений, $\theta = f(\mathcal{P})$ — некоторый функционал на этом множестве (параметр, подлежащий оцениванию). Предположим, что функция $g(x, \theta)$ такова, что уравнение

$$\mathbf{E}_{\mathcal{P}} g(X_1, \theta) = 0$$

имеет единственное решение для каждой априори допустимой меры и что это решение воспроизводит функционал f, т.е. имеет вид $\theta = f(\mathcal{P})$. Тогда оценкой обобщенного метода моментов (GMM) или M-оценкой параметра θ называется решение уравнения

$$\sum_{i=1}^{N} g(X_i, \theta) = 0. \tag{2.2}$$

Обычный метод моментов, описанный выше, укладывается в GMM-схему при

$$g(x, \theta) = g(x) - m(\theta).$$

Некоторой модификацией понятия M-оценки является понятие $\hat{M}$-оценки. Если $\psi(x,\theta)$ — функция двух аргументов, то $\hat{M}$-оценкой параметра θ называется точка (глобального) максимума выражения

$$\sum_{i=1}^{N} \psi(X_i,\theta).$$

Если ψ дифференцируема по θ, то полагая

$$g(x,\theta) = \frac{\partial\psi}{\partial\theta}(x,\theta),$$

мы получаем уравнение (2.2) как необходимое условие максимума. Можно доказать, что при весьма незначительных ограничениях (см. [1]) M-оценки и $\hat{M}$-оценки сильно состоятельны.

Второй общий принцип построения оценок — принцип максимального правдоподобия. Он применим в параметрических моделях с обратимой параметризацией.

Предположим сначала, что априори допустимые распределения дискретны и сосредоточены на едином не более чем счетном множестве E. Пусть $p_\theta(e), e \in E$ — соответствующие вероятности. Рассмотрим вероятность

$$L(\theta) = \prod_{i=1}^{N} p_\theta(X_i) \tag{2.3}$$

как (случайную) функцию параметра θ (она называется функцией правдоподобия — likelihood function). Точка максимума функции правдоподобия объявляется оценкой максимального правдоподобия $\hat{\theta}_{ML}$ параметра θ.

Этот рецепт основан на том обстоятельстве, что реализация случайной функции $L(\theta)$ задает вероятность "реализовавшейся выборки":

$$\mathsf{P}_\theta(X_1 = X_{1,\text{эмп.}}, \ldots, X_N = X_{N,\text{эмп.}}),$$

а реализовалась она, видимо, потому, что эта вероятность достаточно велика, немного утрируя — максимально велика.

Для непрерывного параметрического семейства распределений, заданного плотностью $p_\theta(x)$, функция правдоподобия определяется (через эту плотность) той же формулой (2.3), а рецепт построения оценки $\hat{\theta}_{ML}$ сохраняется.

Удобно сразу же заметить, что идея максимизации правдоподобия пригодна и для более общих схем наблюдений (скажем, для зависимых или неодинаково распределенных наблюдений). Нужно лишь заменить произведение вероятностей или плотностей совместной вероятностью или плотностью. У нас будут возможности воспользоваться этим замечанием в последующих (эконометрических) главах.

Очевидно, что точка максимума функции правдоподобия лежит в множестве $\{\theta : L(\theta) > 0\}$. Поэтому можно перейти к логарифмам и искать максимумы логарифмической функции правдоподобия

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{N} \ln p_\theta(X_i)$$

(они будут в тех же точках).

Таким образом, $\hat{\theta}_{ML}$ является $\hat{M}$-оценкой, а если p_θ дифференцируема по θ, то и M-оценкой. Необходимое условие максимума в гладком случае имеет вид

$$\frac{dl(\theta)}{d\theta} = 0$$

и называется уравнением правдоподобия.

В учебнике Боровкова [1] объясняется, как оценки максимального правдоподобия получаются методом подстановки. Кроме того, в этой книге можно найти унифицированное изложение дискретного и непрерывного случаев на языке доминирующих мер и доминируемых семейств распределений вероятностей.

Свойства оценок максимального правдоподобия будут подробно рассмотрены в следующих параграфах.

2.3 Примеры оценивания

Пример 0. Оценивание простейших моментов — математического ожидания $\mathbf{E}_\mathcal{P}$ и дисперсии $\mathbf{V}_\mathcal{P}$.

Напрашивающимися оценками являются $\bar{X}$ — эмпирическое среднее значение и S^2 — эмпирическая дисперсия. Состоятельность этих оценок была уже выведена из закона больших чисел ранее. Поскольку

$$\mathbf{E}\bar{X} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{E}X_i = \mathbf{E}_\mathcal{P},$$

эта оценка является несмещенной. С другой стороны,

$$\mathbf{E}S^2 = \mathbf{E}(\overline{X^2} - \bar{X}^2) =$$
$$= \mathbf{E}(X_1^2) - \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbf{E}(X_i^2) + 2\sum_{1\le i<j\le N}\mathbf{E}(X_iX_j)\right) =$$
$$= \mathbf{E}(X_1^2) - \frac{1}{N}\mathbf{E}(X_1^2) - \frac{N^2-N}{N^2}(\mathbf{E}X_1)^2 = \frac{N-1}{N}\mathbf{V}X_1,$$

следовательно, эмпирическая дисперсия S^2 смещена. В то же время она является асимптотически несмещенной (т.к. $\frac{N-1}{N} \to 1$ при $N \to \infty$). Для большинства моделей более предпочтительным является исправленный (несмещенный) вариант эмпирической дисперсии

$$S^2_{\text{испр.}} = \frac{N}{N-1}S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

Оценка $S^2_{\text{испр.}}$, очевидно, является состоятельной и несмещенной (для $\mathbf{V}_{\mathcal{P}}$).

Все приведенные соображения применимы как в параметрических, так и в непараметрических моделях. Что же касается эффективности, то, как было замечено в параграфе 1, ее имеет смысл исследовать только в параметрических моделях. В этом плане интерес представляет распределение Пуассона, параметр λ которого является одновременно и математическим ожиданием и дисперсией. Для него у нас есть уже две оценки — $\bar{X}$ и $S^2_{\text{испр.}}$ — обе состоятельные и несмещенные. Можно проверить, что первая из них эффективнее второй (для этого нужны скучные вычисления), но вряд ли целесообразно сейчас это делать — нужны общие методы исследования эффективности, разговор о которых еще впереди.

Пример 0 (продолжение). Оценивание коэффициента корреляции $\rho_{\mathcal{P}}$.

Предположим, что повторная выборка $X_1, \ldots, X_N$ состоит из двумерных случайных величин. Компоненты X_i', X_i'' отдельного наблюдения не предполагаются независимыми между собой. Для подобных двумерных выборок к простейшим моментам относятся, помимо математических ожиданий и дисперсий, ковариация

$$\mathsf{cov}_{\mathcal{P}} = \mathsf{cov}(X_i', X_i'')$$

и коэффициент корреляции

$$\rho_{\mathcal{P}} = \rho(X_i', X_i'').$$

Состоятельной оценкой ковариации $\mathsf{cov}_{\mathcal{P}}$ является эмпирическая ковариация

$$\overline{\mathsf{cov}} = \frac{1}{N}\sum_{i=1}^{N}(X_i' - \overline{X'})(X_i'' - \overline{X''})$$
$$= \frac{1}{N}\sum_{i=1}^{N} X_i'X_i'' - \overline{X'}\cdot\overline{X''} = \overline{X'X''} - \overline{X'}\cdot\overline{X''}.$$

Так же как и эмпирическая дисперсия, эта оценка лишь асимптотически несмещена. Исправить ее можно точно так же как и эмпирическую дисперсию (проверьте!):

$$\overline{\mathsf{cov}}_{\text{испр.}} = \frac{N}{N-1}\overline{\mathsf{cov}},\ \mathbf{E}\overline{\mathsf{cov}}_{\text{испр.}} = \mathsf{cov}_{\mathcal{P}}.$$

Оценкой коэффициента корреляции

$$\rho_{\mathcal{P}} = \frac{\mathsf{cov}(X_i', X_i'')}{\sqrt{\mathbf{V}X_i'\mathbf{V}X_i''}}$$

по методу моментов является эмпирический коэффициент корреляции

$$r = \frac{\overline{\mathsf{cov}}}{S'S''} = \frac{\overline{\mathsf{cov}}_{\text{испр.}}}{S'_{\text{испр.}}S''_{\text{испр.}}}.$$

Здесь S'^2 и S''^2 — эмпирические дисперсии компонент выборки. Полезно отметить, что при вычислении r можно пользоваться как исправленными, так и неисправленными вариантами эмпирических дисперсий и ковариации (или просто соответствующими суммами) — это обстоятельство отражает безразмерность коэффициента корреляции.

Очевидно, что эмпирический коэффициент корреляции r состоятельно оценивает теоретический коэффициент $\rho_{\mathcal{P}}$. Ожидать несмещенности этого нелинейного выражения, конечно, не приходится.

В следующих примерах мы будем обсуждать оценки максимального правдоподобия, возвращаясь к методу моментов лишь в случаях, не укладывающихся в схему примера 0.

Пример 1. Вероятность успеха p ($p = \mathbf{E}X_1$).

Функция правдоподобия имеет вид

$$L(p) = p^{S_N}(1-p)^{N-S_N}, 0 \le p \le 1,$$

где $S_N = X_1 + \cdots + X_N$ — общее (суммарное) число успехов в N испытаниях. Если $S_N = N$, функция $L(p)$ оказывается степенной: $L(p) = p^N$, так что $\hat{p}_{ML} = 1$. Аналогично, если $S_N = 0$, $\hat{p}_{ML} = 0$. В остальных случаях $L(p)$ обращается в 0 (т.е. в минимум) на концах отрезка $[0,1]$, а точку максимума следует искать дифференцированием. Во внутренних точках отрезка $[0,1]$ можно перейти к логарифмической функции правдоподобия $l(p)$ и написать

$$\frac{dl(p)}{dp} = \frac{S_N}{p} - \frac{N - S_N}{1-p}.$$

Приравнивая производную нулю, получаем

$$\hat{p}_{ML} = \frac{S_N}{N} = \bar{X}.$$

Остается лишь отметить, что выделенные в начале рассуждения особые случаи также укладываются в эту формулу. Таким образом, мы не получили ничего нового по сравнению с примером 0. Впрочем, было бы удивительно, если бы обнаружилось что-нибудь иное.

Пример 2. Распределение Пуассона $\Pi(\lambda)$.

Логарифмическая функция правдоподобия имеет вид

$$l(\lambda) = \sum_{i=1}^{N} \ln p_\lambda(X_i) = \sum_{i=1}^{N} \ln\left[\frac{\lambda^{X_i}}{X_i!}e^{-\lambda}\right] =$$

$$= \sum_{i=1}^{N}[X_i \ln\lambda - \ln(X_i!) - \lambda] = \sum_{i=1}^{N} X_i \ln\lambda - N\lambda - \sum_{i=1}^{N}\ln(X_i!).$$

Дифференцируя по λ и приравнивая производную нулю, находим

$$\hat{\lambda}_{ML} = \bar{X}.$$

Без особого труда проверяется, что найдена именно точка максимума. Особо следует рассмотреть случай $\bar{X} = 0$.

Пример 3. Нормальное распределение $\mathbf{N}(a, \sigma^2)$.

$$l(a, \sigma^2) = \ln[(2\pi)^{-N/2}] - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - a)^2.$$

Дифференцируя по a и приравнивая производную нулю, получаем

$$\hat{a}_{ML} = \bar{X}$$

(вообще-то надо решать систему двух уравнений, но уравнение $\partial l/\partial a = 0$ решается без использования второго уравнения). Теперь, дифференцируя по σ и подставляя $\hat{a}_{ML}$, получаем

$$\hat{\sigma}^2_{ML} = S^2.$$

Можно было бы оценивать не σ, а σ^2 и дифференцировать по σ^2. Результат бы не изменился.

Стандартным способом — через матрицу вторых производных — можно проверить, что найденные оценки действительно определяют точку максимума. Как нам уже известно, оценка $\hat{\sigma}^2_{ML}$ смещена.

Аналогично можно проверить, что для двумерной нормально распределенной выборки с параметрами a', a'', σ'^2, σ''^2, ρ оценками максимального правдоподобия являются $\overline{X'}$, $\overline{X''}$, S'^2, S''^2, r.

Пример 4. Гамма-распределение.

Простейший вариант метода моментов дает (при $N \geq 2$)

$$\hat{\alpha} = \frac{\bar{X}}{S^2}, \hat{p} = \frac{\bar{X}^2}{S^2}.$$

Решить систему уравнений правдоподобия в элементарных функциях не удается, так что оба метода расходятся в своих рекомендациях.

Пример 5. Равномерное распределение на $\langle a, b\rangle$.

Поскольку

$$\mathbf{E} = \frac{a+b}{2}, \mathbf{V} = \frac{(b-a)^2}{12},$$

по методу моментов получаем

$$\hat{a} = \bar{X} - \sqrt{3}S, \hat{b} = \bar{X} + \sqrt{3}S.$$

В то же время

$$L(a,b) = (b-a)^{-N}, \quad \text{если} \quad a < X_1, \ldots, X_N < b.$$

Для увеличения значения функции правдоподобия следует сближать аргументы a и b, пока это возможно. Получаем

$$\hat{a}_{ML} = X_{\min}(= \min(X_1, \ldots, X_N)),$$

$$\hat{b}_{ML} = X_{\max}(= \max(X_1, \ldots, X_N)).$$

Эти оценки доставляют если не максимум, то, по крайней мере, супремум функции правдоподобия [4]. С точки зрения правдоподобия пунктуально отличать максимум от супремума представляется нецелесообразным (как и менять максимум на супремум в исторически сложившемся названии метода).

И здесь оба наши метода оценивания дают отличающиеся результаты, причем оценки максимального правдоподобия более соответствуют смыслу параметров. Заметим, впрочем, что они явно смещены "внутрь", т.е. $\hat{a}_{ML} \geq a, \hat{b}_{ML} \leq b$. Равномерное распределение удобно в качестве учебного примера. Во-первых, практически все вычисления можно провести явно, в элементарных функциях. Во-вторых, оно не регулярно и иллюстрирует некоторые эффекты, отсутствующие в регулярном случае (см. параграфы 4 и 5). По этой причине мы приведем несколько формул, характеризующих оценки максимального правдоподобия, и даже наметим их вывод. Для определенности будем работать с $X_{\max}$ (эмпирический минимум рассматривается аналогично, а формулы угадываются из соображений симметрии).

Сначала заметим, что

$$\begin{aligned}\mathsf{P}(X_{\max} < x) &= \mathsf{P}(X_1 < x, \ldots, X_N < x) \\ &= [\mathsf{P}(X_1 < x)]^N = \left(\frac{x-a}{b-a}\right)^N, a < x < b.\end{aligned}$$

Отсюда плотность величины $X_{\max}$ равна

$$\frac{N(x-a)^{N-1}}{(b-a)^N}, a < x < b.$$

Через нее находятся ($u = \frac{x-a}{b-a}$)

$$\begin{aligned}\mathbf{E}X_{\max} = \int_a^b x\frac{N(x-a)^{N-1}}{(b-a)^N}dx &= \int_0^1 [a+u(b-a)]Nu^{N-1}du \\ &= a + \frac{N}{N+1}(b-a), \\ \mathbf{E}(X_{\max}^2) &= \int_0^1 [a+u(b-a)]^2 Nu^{N-1}du \\ &= a^2 + 2\frac{N}{N+1}a(b-a) + \frac{N}{N+2}(b-a)^2\end{aligned}$$

[4] Это зависит от определения плотности в точках a и b.

и

$$\mathbf{V}X_{\max} = \mathbf{E}(X_{\max}^2) - (X_{\max})^2$$
$$= (b-a)^2\left[\frac{N}{N+2} - \left(\frac{N}{N+1}\right)^2\right] = \frac{N}{(N+1)^2(N+2)}(b-a)^2.$$

Поскольку

$$\mathbf{E}X_{\max} \to a + (b-a) = b$$

и

$$\mathbf{V}X_{\max} \to 0$$

при $N \to \infty$, оценка $\hat{b}_{ML}$ — состоятельная и асимптотически несмещенная. То же верно и для $\hat{a}_{ML}$.

Из формул

$$\mathbf{E}X_{\max} = a + \frac{N}{N+1}(b-a),$$

$$\mathbf{E}X_{\min} = b - \frac{N}{N+1}(b-a)$$

легко выводится (надо "решить"эти равенства относительно a и b), что линейные комбинации

$$\tilde{a} = \frac{N}{N-1}X_{\min} - \frac{1}{N-1}X_{\max},$$

$$\tilde{b} = \frac{N}{N-1}X_{\max} - \frac{1}{N-1}X_{\min}$$

являются несмещенными оценками a и b соответственно:

$$\mathbf{E}\tilde{a} = a, \mathbf{E}\tilde{b} = b.$$

Кроме того, из состоятельности оценок максимального правдоподобия $X_{\min}$ и $X_{\max}$ сразу же следует и состоятельность $\tilde{a}$ и $\tilde{b}$. В параграфе 7 будет установлено, что эти последние оценки еще и эффективны в классе K_0 несмещенных оценок.

Подведем некоторый итог рассмотрения примеров. Кроме ранее отмеченной проблемы поиска эффективных оценок обнаружилась еще одна трудность — невозможность во многих случаях аналитически решить уравнения правдоподобия. Полезный итеративный метод решения уравнений правдоподобия будет указан в параграфе 11.

2.4 Условия регулярности и неравенство Рао–Крамéра

Аккуратное математическое обоснование материала этого параграфа довольно громоздко и неинтересно для пользователей. Поэтому мы спрячем эти тонкости при помощи оборота "при некоторых условиях регулярности". В конце параграфа условия регулярности будут описаны неформально.

Итак, речь пойдет о несмещенных оценках одномерного параметра θ в параметрической модели, когда априори допустимое распределение вероятностей $\mathcal{P}_\theta$ однозначно характеризуется этим параметром. Для простоты будем предполагать, что область изменения θ — невырожденный промежуток. Предположим также, что логарифмическая функция правдоподобия $l(\theta)$ дифференцируема по θ и

$$I(\theta) = \mathbf{E}{l'}^2(\theta) < \infty$$

(здесь и далее в этом параграфе штрихом обозначено дифференцирование по θ). Функция $I(\theta)$ часто называется информацией Фишера.

Теорема. Пусть $\hat{\theta}$ — несмещенная оценка параметра θ. Тогда

$$\mathbf{V}(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

Неравенство такого вида справедливо "при некоторых условиях регулярности". Впервые оно доказано независимо друг от друга Фреше, Рао и Крамéром в 1943-45 г.г. и в литературе обычно называется неравенством Рао-Крамéра. Ниже приводится схема доказательства. В его основе лежат два равенства:

$$\mathbf{E}l'(\theta) = 0, \quad \mathbf{E}[\hat{\theta}l'(\theta)] = 1 \tag{2.4}$$

(они объясняются чуть ниже, именно в этих объяснениях потребуются "условия регулярности"). Само неравенство Рао-Крамéра из формул (2.4) получается так. Заметим сначала, что

$$\rho^2(\hat{\theta}, l'(\theta)) = \frac{\mathsf{cov}^2(\hat{\theta}, l'(\theta))}{\mathbf{V}\hat{\theta}\mathbf{V}l'(\theta)} \leq 1$$

(известное свойство коэффициента корреляции). Однако из (2.4) следует, что

$$\mathsf{cov}(\hat{\theta}, l'(\theta)) = \mathbf{E}[\hat{\theta} l'(\theta)] - \mathbf{E}\hat{\theta}\mathbf{E}l'(\theta) = 1,$$
$$\mathbf{V}l'(\theta) = \mathbf{E}{l'}^2(\theta) - (\mathbf{E}l'(\theta))^2 = I(\theta).$$

Поэтому

$$\mathbf{V}\hat{\theta} \geq \frac{\mathsf{cov}^2(\hat{\theta}, l'(\theta))}{\mathbf{V}l'(\theta)} = \frac{1}{I(\theta)}.$$

Перейдем теперь к доказательству равенств (2.4), считая для определенности, что совместное распределение выборки [5]

$$\vec{X} = (X_1, \ldots, X_N)^T$$

задается плотностью $p_\theta(\vec{x})$ (дискретный случай рассматривается аналогично). Имеем

$$\mathbf{E}l'(\theta) = \mathbf{E}\frac{{p'}_\theta(\vec{X})}{p_\theta(\vec{X})} = \int \frac{{p'}_\theta(\vec{x})}{p_\theta(\vec{x})} p_\theta(\vec{x}) d\vec{x}$$
$$= \int {p'}_\theta(\vec{x}) d\vec{x} = \left(\int p_\theta(\vec{x}) d\vec{x} \right)' = 1' = 0.$$

Точно так же

$$\mathbf{E}[\hat{\theta} l'(\theta)] = \mathbf{E}\left[\hat{\theta}(\vec{X}) \frac{{p'}_\theta(\vec{X})}{p_\theta(\vec{X})} \right] = \int \hat{\theta}(\vec{x}) \frac{{p'}_\theta(\vec{x})}{p_\theta(\vec{x})} p_\theta(\vec{x}) d\vec{x} =$$
$$= \int \hat{\theta}(\vec{x}) {p'}_\theta(\vec{x}) d\vec{x} = \left(\int \hat{\theta}(\vec{x}) p_\theta(\vec{x}) d\vec{x} \right)' = (\mathbf{E}\hat{\theta})' = \theta' = 1.$$

В обеих выкладках предполагается существование всех фигурирующих в них выражений, а также возможность "дифференцирования по параметру под знаком интеграла". Собственно в этом и состоят условия регулярности. Во второй выкладке дополнительно используется равенство $\mathbf{E}\hat{\theta} = \theta$ (несмещенность оценки). Условия дифференцирования по параметру можно найти в подробных курсах математического анализа. С точки зрения пользователя главное из них — отсутствие зависимости области интегрирования от параметра. Подробное обсуждение условий регулярности можно найти у Боровкова [1].

[5] Знак транспонирования присутствует по той причине, что вектор-столбец $\vec{X}$ записан в строчку.

Полезно отметить, что наш вывод неравенства Рао-Крамéра не использует ни независимости, ни одинаковой распределенности наблюдений $X_1, \ldots, X_N$. Для независимых наблюдений с плотностями $p_{\theta,j}(x_j)$ легко проверить, что

$$I(\theta) = \sum_{j=1}^{N} i_j(\theta),$$

где

$$i_j(\theta) = \mathbf{E}(\ln p_{\theta,j}(X_j))'^2.$$

Действительно,

$$l'^2(\theta) = [\sum_j (\ln p_{\theta,j}(X_j))']^2,$$

но удвоенные произведения, образующиеся при возведении в квадрат, имеют нулевые математические ожидания в силу независимости и первой формулы (2.4).

В частности, для повторной выборки $I(\theta) = Ni(\theta)$, где $i(\theta)$ — общее значение величин $i_j(\theta)$. Функцию $i(\theta)$ можно назвать удельной фишеровской информацией.

Следствие. Пусть $\hat{\theta} \in K_b$. Тогда

$$\mathbf{V}(\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{I(\theta)},$$

$$\mathbf{E}(\hat{\theta} - \theta)^2 \geq \frac{[1 + b'(\theta)]^2}{I(\theta)} + b^2(\theta).$$

Первое неравенство доказывается по той же схеме с использованием соотношения

$$\mathbf{E}[\hat{\theta} l'(\theta)] = (\mathbf{E}\hat{\theta})' = (\theta + b(\theta))' = 1 + b'(\theta).$$

Второе неравенство вытекает из формулы

$$\mathbf{E}(\hat{\theta} - \theta)^2 = \mathbf{E}[(\hat{\theta} - \mathbf{E}\hat{\theta}) + b(\theta)]^2 = \\ = \mathbf{V}(\hat{\theta}) + 2b(\theta)\mathbf{E}(\hat{\theta} - \mathbf{E}\hat{\theta}) + b^2(\theta) = \mathbf{V}(\hat{\theta}) + b^2(\theta).$$

Аналогичное (матричное) неравенство Рао–Крамéра имеет место для многомерного параметра:

$$C(\hat{\theta}) - I^{-1}(\theta) \geq 0.$$

Здесь $C(\hat{\theta})$ — матрица ковариаций случайного вектора $\hat{\theta}$, а

$$I(\theta) = \mathbf{E}[\operatorname{grad} l(\theta) \cdot \operatorname{grad} l(\theta)^T]$$

— матричный вариант информации Фишера. Запись $\cdots \geq 0$ означает, что слева стоит неотрицательно определенная матрица.

Связь неравенства Рао-Крамéра с эффективными оценками обсуждается в следующем параграфе.

2.5 Простейшие приемы нахождения эффективных оценок. Экспоненциальные семейства

Приемы, о которых идет речь, основаны на простом наблюдении. Если (в регулярном случае) оценка $\hat{\theta} \in K_0$ обращает неравенство Рао–Крамéра в равенство, то она эффективна. Приведем несколько примеров. В этих примерах удобно пользоваться следующими представлениями для $I(\theta)$ и $i_j(\theta)$:

$$I(\theta) = -\mathbf{E}l''(\theta), i_j(\theta) = -\mathbf{E}(\ln p_{\theta,j}(X_j))''.$$

Докажем первое из них (второе является следствием):

$$\begin{aligned} I(\theta) + \mathbf{E}l''(\theta) &= \mathbf{E}[l'^2(\theta) + l''(\theta)] \\ &= \mathbf{E}\left[\left(\frac{p'_\theta(\vec{X})}{p_\theta(\vec{X})}\right)^2 + \left(\frac{p'_\theta(\vec{X})}{p_\theta(\vec{X})}\right)'\right] \\ &= \mathbf{E}\left[\frac{p'^2_\theta(\vec{X})}{p^2_\theta(\vec{X})} + \frac{p''_\theta(\vec{X})p_\theta(\vec{X}) - p'^2_\theta(\vec{X})}{p^2_\theta(\vec{X})}\right] \\ &= \mathbf{E}\frac{p''_\theta(\vec{X})}{p_\theta(\vec{X})} = \int \frac{p''_\theta(\vec{x})}{p_\theta(\vec{x})} p_\theta(\vec{x}) d\vec{x} \\ &= \int p''_\theta(\vec{x}) d\vec{x} = \left(\int p_\theta(\vec{x}) d\vec{x}\right)'' = 0. \end{aligned}$$

Разумеется, в этой выкладке используются дополнительные предположения регулярности, связанные со второй производной.

Проверять условия регулярности для каждого отдельного примера мы не будем.

Пример 1. Оценка вероятности успеха.
Проверим, что $\hat{p} = \bar{X}$ эффективна. Для этого сосчитаем

$$I(p) = -\mathbf{E}[(S_N \ln p + (N - S_N)\ln(1-p))''] \\ = \mathbf{E}\left[\frac{S_N}{p^2} + \frac{N - S_N}{(1-p)^2}\right] = \frac{Np}{p^2} + \frac{N - Np}{(1-p)^2} \\ = \frac{N}{p} + \frac{N}{1-p} = \frac{N}{p(1-p)}.$$

Остается заметить, что

$$\mathbf{V}(\hat{p}) = \frac{\mathbf{V}(S_N)}{N^2} = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N} = \frac{1}{I(p)}.$$

Пример 2. Распределение Пуассона $\Pi(\lambda)$.
Докажем, что $\hat{\lambda}_{ML} = \bar{X}$ эффективна.

$$I(\lambda) = -\mathbf{E}l''(\lambda) = \mathbf{E}\frac{\sum_{i=1}^{N} X_i}{\lambda^2} = \frac{N\lambda}{\lambda^2} = \frac{N}{\lambda},$$

$$\mathbf{V}(\hat{\lambda}_{ML}) = \frac{\mathbf{V}(X_1 + \cdots + X_N)}{N^2} = \frac{N\lambda}{N^2} = \frac{\lambda}{N} = \frac{1}{I(\lambda)}.$$

К сожалению, далеко не всегда дело обстоит столь приятным образом. Общая картина выглядит так.

Теорема. Если несмещенная оценка $\hat{\theta}$ обращает неравенство Рао-Крамéра в равенство на всем промежутке изменения параметра θ, то она удовлетворяет уравнению правдоподобия

$$l'(\hat{\theta}) = 0.$$

Доказательство основано на анализе случаев, когда коэффициент корреляции $\rho(\hat{\theta}, l'(\theta))$ равен 1. Так будет, если $\hat{\theta}$ и $l'(\theta)$ линейно связаны:

$$\hat{\theta} = \alpha(\theta) l'(\theta) + \beta(\theta) \tag{2.5}$$

Коэффициенты α и β могут (и даже должны) зависеть от θ — в противном случае зависела бы от θ оценка $\hat{\theta}$, что противоречит определению. Вычисляя математическое ожидание обеих частей формулы (2.5), находим

$$\theta = \mathbf{E}\hat{\theta} = \alpha(\theta)\mathbf{E}l'(\theta) + \beta(\theta) = \beta(\theta).$$

Следовательно, тождественно по θ выполняется

$$\hat{\theta} = \alpha(\theta) l'(\theta) + \theta \tag{2.6}$$

Подставляя в (2.6) саму оценку $\hat{\theta}$, получаем $\alpha(\theta) l'(\theta) = 0$. Сокращая на коэффициент, получаем требуемый результат (мы опускаем исследование исключительных ситуаций, когда $\alpha(\theta) = 0$ — с вероятностью 1 они не реализуются; аккуратный анализ также требует некоторых условий регулярности).

Таким образом, кандидатами на роль эффективной оценки являются, в рамках нашего подхода, оценки максимального правдоподобия. К сожалению, они не обязаны быть несмещенными, и в этом случае неравенство Рао-Крамéра не обращается в равенство ни для какой (несмещенной) оценки, в том числе и для эффективной. В параграфе 7 мы обсуждаем другой, более действенный, подход к нахождению эффективных оценок.

Записывая соотношение (2.6) в виде

$$l'(\theta) = \alpha^{-1}(\theta)[\hat{\theta} - \theta]$$

и интегрируя по θ, получаем

$$l(\theta) = l(\theta_0) + \hat{\theta} \int_{\theta_0}^{\theta} \alpha^{-1}(t) dt - \int_{\theta_0}^{\theta} t \alpha^{-1}(t) dt.$$

Поэтому наше семейство плотностей должно при этом представляться в виде

$$p_\theta(\vec{x}) = h(\vec{x}) \exp\{\hat{\theta}(\vec{x}) A(\theta) + B(\theta)\},$$

где $A(\theta)$ и $B(\theta)$ — какие-то функции от параметра θ, а множитель $h(\vec{x})$, напротив, от параметра θ не зависит.

Семейства плотностей такого вида называются экспоненциальными семействами.

Таким образом, наш подход может дать эффективную оценку только для экспоненциальных семейств. Аналогично обстоит дело и в случае многомерного параметра. Мы ограничимся только аккуратным определением экспоненциальных семейств в этом случае.

Пусть $\theta \in \mathbb{R}^k$ — k-мерный параметр. Семейство плотностей (в дискретном случае — вероятностей) $p_\theta(\vec{x})$ называется экспоненциальным, если допускает представление вида

$$p_\theta(\vec{x}) = h(\vec{x}) \exp\{U(\vec{x})^T A(\theta) + B(\theta)\},$$

где $U(\vec{x})$ и $A(\theta)$ — вектор-функции (столбцы) со значениями в $\mathbb{R}^k$, $U(\vec{x})^T$ — транспонированный вектор, $h(\vec{x})$ и $B(\theta)$ — функции с числовыми значениями. Подчеркнем, что размерность значений вектор-функций U и A совпадает с размерностью параметра.

Почти все семейства распределений, перечисленные в параграфе 1.5, экспоненциальны.

Продолжим серию наших примеров.

Пример 3. Нормальное распределение $\mathbf{N}(a, \sigma^2)$.

Прежде всего заметим, что семейство нормальных плотностей экспоненциально:

$$p_{a,\sigma^2}(\vec{x}) = \exp\left\{ -\sum_{i=1}^{N} x_i^2 \cdot \frac{1}{2\sigma^2} + \sum_{i=1}^{N} x_i \cdot \frac{a}{\sigma^2} \right.$$
$$\left. -\frac{Na^2}{2\sigma^2} - -N\ln\sigma - \frac{N}{2}\ln(2\pi) \right\}.$$

Однако с оценками максимального правдоподобия не все в порядке — эмпирическая дисперсия S^2 смещена (а ее исправленный вариант уже не является оценкой максимального правдоподобия). Ввиду важности нормального распределения для статистики, выпишем информационную матрицу $I(a, \sigma^2)$, а также матрицу ковариаций вектора несмещенных оценок $(\bar{X}, S^2_{\text{испр.}})'$.

$$I = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{pmatrix}, I^{-1} = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{pmatrix}$$

$$C(\bar{X}, S^2_{\text{испр.}}) = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N-1} \end{pmatrix}.$$

Из сравнения двух последних матриц следует, что $\bar{X}$ имеет минимально возможную дисперсию, т.е. эффективна для a в двухпараметрическом случае, или, как иногда говорят, при наличии мешающего параметра σ. Сказать что-нибудь определенное об эффективности $S^2_{\text{испр.}}$ в рамках нашего подхода не представляется возможным (в дальнейшем мы увидим, что и эта оценка эффективна).

Пример 4. Гамма-распределение.

Ограничимся однопараметрическим семейством с параметром α при известном p (при $p = 1$ получается семейство показательных

распределений). Очевидно,

$$\hat{\alpha}_{ML} = \frac{p}{\bar{X}}.$$

Можно проверить, что эта оценка смещенная ($\mathbf{E}\hat{\alpha}_{ML} = \frac{Np}{Np-1}\alpha$), так что наш подход ответа не дает. Впрочем, для $\theta = \alpha^{-1}$ оценка максимального правдоподобия $\hat{\theta}_{ML} = p^{-1}\bar{X}$ является несмещенной. С помощью неравенства Рао-Крамéра без труда проверяется, что $\hat{\theta}_{ML}$ эффективна для θ в классе K_0 несмещенных оценок. Мы увидим позже (см. параграф 7), что несмещенная оценка

$$\frac{Np-1}{Np}\hat{\alpha}_{ML}$$

эффективна для α в K_0.

Пример 5. Равномерное распределение на $\langle a, b\rangle$.

Это семейство не удовлетворяет условиям регулярности, т.к. носитель плотности — промежуток $\langle a, b\rangle$ — зависит от параметров. Само неравенство Рао-Крамéра также не выполняется. Можно показать, что построенные в параграфе 3 эффективные несмещенные оценки $\tilde{a}$ и $\tilde{b}$ имеют дисперсии, убывающие обратно пропорционально N^2 (ср. с формулой для $\mathbf{V}X_{\min}$ в этом параграфе), в то время как неравенство Рао-Крамéра разрешало бы им убывать не быстрее, чем обратно пропорционально N. Такая "сверхэффективность"связана с тем, что параметры a и b — точки разрыва (нерегулярности) плотности. Извлечь из наблюдений информацию о таких характеристиках теоретического распределения, как правило, легче, чем о параметрах регулярного типа. Напомним, что эффективность оценок a и b будет доказана в параграфе 7.

2.6 Достаточные статистики

Основное определение этого параграфа опирается на общее понятие условного распределения. Краткое резюме теории условных распределений содержится в приложении D.

Итак, предположим, что задана параметрическая статистическая модель, т.е. семейство априори допустимых распределений вероятностей P_θ, где θ — конечномерный параметр, однозначно определяющий P_θ. Статистика $S = S(\vec{X})$ называется **достаточной** (для параметра θ),

если условное распределение выборки относительно S — $\mathsf{P}(\vec{X} \in B|S)$ — не зависит от параметра θ (точнее, существует вариант этого условного распределения, не зависящий от θ).

Неформально это определение означает, что вся информация о параметре, содержащаяся в выборке $\vec{X}$, фактически содержится уже в $S(\vec{X})$: свобода, остающаяся в выборке после фиксации значения статистики S, имеет "универсальный"характер, не имеющий отношения к θ. Можно сказать также, что достаточная статистика представляет выборочную информацию о параметре в сжатом виде, но без потерь (конечно, ее надо еще расшифровывать).

Полезно сразу же рассмотреть пример, дающий такое сжатое представление.

Пример 1. Модель испытаний Бернулли.

Априори допустимыми являются распределения P_p вида

$$\mathsf{P}_p(\vec{X} = \vec{x}) = \prod_{i=1}^{N} [p^{x_i}(1-p)^{1-x_i}] = p^{\sum x_i}(1-p)^{N-\sum x_i}$$

(мы представляем выборку $\vec{X}$ обычным образом — как последовательность независимых случайных величин X_i, принимающих значения 1 (успех) и 0 (неудача) с вероятностями p и $1-p$ соответственно). Докажем, что статистика $S = S_N = X_1 + \cdots + X_N$ (полное число успехов) является достаточной для p. Выберем некоторое k, $0 \le k \le N$, и согласующееся с ним $\vec{x}$, так что

$$S(\vec{x}) = x_1 + \cdots + x_N = k$$

(иначе условная вероятность будет нулевой). Тогда

$$\mathsf{P}_p(\vec{X} = \vec{x}|S = k) = \frac{\mathsf{P}_p(\vec{X} = \vec{x}, S = k)}{\mathsf{P}_p(S = k)}$$
$$= \frac{\mathsf{P}_p(\vec{X} = \vec{x})}{\mathsf{P}_p(S = k)} = \frac{p^k(1-p)^{N-k}}{C_N^k p^k (1-p)^{N-k}} = \frac{1}{C_N^k}.$$

Мы видим, что фиксация числа успехов k оставляет только свободу в порядке появления в выборке этих успехов и дополнительного числа неудач. Все такие порядки ("сочетания") оказываются условно равновероятными (а остальные комбинации успехов и неудач — условно невозможными). Таким образом, вся выборочная информация о

параметре p содержится уже в суммарном числе успехов S. Именно эта статистика и позволяет (см. параграфы 3 и 5) оценить p эффективным образом: $\hat{p} = S/N$.

Устанавливать достаточность, пользуясь определением, не всегда удобно, особенно в непрерывных моделях, поэтому чаще всего используют следующую теорему факторизации Неймана-Фишера:

Теорема факторизации. Статистика S достаточна в том и только в том случае, если функция правдоподобия $L(\theta)$ представляется (факторизуется) в виде

$$L(\theta) = h(\vec{X})\psi(S, \theta).$$

Мы докажем эту теорему только для семейств дискретных распределений. В непрерывном случае доказательство основано на тех же идеях, но технически значительно сложнее.

Пусть сначала функция правдоподобия факторизуется. Докажем, что S достаточна. Для этого рассмотрим некоторое s (значение функции S) и $\vec{x} \in S^{-1}(s)$. Тогда

$$\begin{aligned}\mathsf{P}_\theta(\vec{X} = \vec{x}|S = s) &= \frac{\mathsf{P}_\theta(\vec{X} = \vec{x}, S(X) = s)}{\mathsf{P}_\theta(S = s)} = \frac{\mathsf{P}_\theta(\vec{X} = \vec{x})}{\mathsf{P}_\theta(S = s)} \\ &= \frac{\mathsf{P}_\theta(\vec{X} = \vec{x})}{\sum_{\vec{y}\in S^{-1}(s)} \mathsf{P}_\theta(\vec{X} = \vec{y})} = \frac{h(\vec{x})\psi(S(\vec{x}), \theta)}{\sum_{\vec{y}\in S^{-1}(s)} h(\vec{y})\psi(S(\vec{y}), \theta)} \\ &= \frac{h(\vec{x})\psi(s, \theta)}{\sum_{\vec{y}\in S^{-1}(s)} h(\vec{y})\psi(s, \theta)} = \frac{h(\vec{x})}{\sum_{\vec{y}\in S^{-1}(s)} h(\vec{y})}.\end{aligned}$$

Для $\vec{x} \notin S^{-1}(s)$ рассматриваемая условная вероятность обращается в 0.

Обратно, предположим, что

$$\mathsf{P}_\theta(\vec{X} = \vec{x}|S = s)$$

не зависит от параметра θ. Обозначим ее $h(\vec{x})$. Указывать дополнительно ее зависимость от s не нужно, т.к. $s = S(\vec{x})$. Тогда (ср. с предыдущим рассуждением)

$$\frac{\mathsf{P}_\theta(\vec{X} = \vec{x})}{\mathsf{P}_\theta(S = s)} = h(\vec{x}).$$

Теперь обозначаем $\mathsf{P}_\theta(S = s)$ через $\psi(s, \theta)$ и получаем

$$\mathsf{P}_\theta(\vec{X} = \vec{x}) = h(\vec{x})\psi(s, \theta) = h(\vec{x})\psi(S(\vec{x}), \theta).$$

Теорема в дискретном варианте доказана.

Технические проблемы в доказательстве непрерывного варианта возникают по причине того, что множество $S^{-1}(s)$ может иметь сложную структуру (см. [1])

Факторизация, указанная в теореме Неймана-Фишера, неоднозначна — первый множитель можно домножить (а второй, соответственно, поделить) на произвольную строго положительную функцию от достаточной статистики S. Поэтому иногда удобнее рассматривать отношение правдоподобия

$$\frac{L(\theta)}{L(\theta')}.$$

Почти очевидно, что статистика S достаточна в том и только в том случае, если отношение правдоподобия является функцией от достаточной статистики:

$$\frac{L(\theta)}{L(\theta')} = Z(S;\theta,\theta').$$

В этом представлении уже нет упомянутой выше неоднозначности.

Предположим, что $p_\theta(\vec{x})$ — экспоненциальное семейство (см. параграф 5):

$$p_\theta(\vec{x}) = h(\vec{x})\exp\{U(\vec{x})^T A(\theta) + B(\theta)\}.$$

Очевидно, что эта формула уже является факторизацией, а $U(\vec{X})$ — достаточная статистика, размерность которой равна размерности параметра.

На этом пути сразу получаем:

Пример 2. $X_1 + \cdots + X_N$ и $\bar{X}$ — достаточные статистики для параметра λ распределения Пуассона. Эти две статистики эквивалентны в естественном смысле — взаимно однозначно определяют друг друга.

Пример 3. $(X_1 + \cdots + X_N, X_1^2 + \cdots + X_N^2)$ — достаточная статистика для двухпараметрического семейства нормальных распределений (см. параграф 5). Другой, эквивалентный, вариант достаточной статистики — $(\bar{X}, S^2)$. Действительно,

$$\bar{X} = \frac{1}{N}(X_1 + \cdots + X_N),$$

$$S^2 = \frac{1}{N}(X_1^2 + \cdots + X_N^2) - \frac{1}{N^2}(X_1 + \cdots + X_N)^2.$$

Формулы обратного преобразования читатель может вывести самостоятельно.

Пример 4. (Гамма-распределение.) Легко проверить, что $(X_1 + \cdots + X_N, X_1 \cdot X_2 \cdot \cdots \cdot X_N)$ — достаточная статистика. При известном p достаточной будет сумма $X_1 + \cdots + X_N$.

Пример 5. (Равномерное распределение.) Любая из статистик $(X_{\min}, X_{\max})$, $(\tilde{a}, \tilde{b})$ (см. параграф 2.3) является при $N \geq 2$ достаточной.

Рассмотрим модифицированную постановку задачи: пусть $a = \theta$, $b = 1 + \theta$. Соответствующее семейство плотностей — однопараметрическое. Но достаточной статистикой по-прежнему является пара $(X_{\min}, X_{\max})$ — наблюдается несоответствие размерностей. Оценивать несмещенным образом θ можно теперь двояко:

$$\theta^* = \tilde{a}, \theta^{**} = \tilde{b} - 1.$$

Почти очевидно, что эти оценки одинаково эффективны. А как найти самую эффективную в K_0 оценку? Мы вернемся к этому вопросу в параграфах 7 и 9.

В заключение параграфа заметим, что вариационный ряд $X_{(1)} = X_{\min}$, $X_{(2)}, \ldots, X_{(N)} = X_{\max}$ всегда является достаточной статистикой в случае повторных наблюдений — если его зафиксировать, остается лишь свобода в последовательности появления этих значений в выборке. По соображениям симметрии все такие последовательности равновероятны. В непрерывном случае можно считать, что все порядковые статистики различны (это событие почти достоверно — имеет вероятность 1). Тогда условное распределение приписывает вес $1/N!$ каждой перестановке вариационного ряда. В дискретном случае возможны совпадения, и условное распределение оказывается иным, но тоже описывается чисто комбинаторно.

В книге Боровкова [1] приводится пример — семейство сдвинутых распределений Коши с плотностью

$$p_\theta(x) = \frac{1}{\pi} \frac{1}{(x-\theta)^2 + 1}, x \in \mathbb{R},$$

для которого вариационный ряд является **минимальной** достаточной статистикой. По существу, этот пример показывает, что достаточные статистики могут быть практически бесполезными.

2.7 Достаточность и эффективность

Из неформального смысла достаточности становится правдоподобным, что искать эффективные оценки следует исключительно при помощи

достаточных статистик. Мы сейчас сформулируем соответствующий рецепт точно, считая для простоты, что θ — одномерный параметр. Буквой S будет обозначаться достаточная статистика. Свойства условных математических ожиданий обсуждаются в Приложении D.

Лемма. Пусть $T = T(\vec{X})$ — некоторая статистика. Тогда $\mathbf{E}_\theta(T|S)$ — также статистика.

Смысл этого утверждения в том, что указанное условное математическое ожидание не зависит от параметра θ. Лемма вытекает из того, что оно (т.е. ожидание) получается интегрированием по условному распределению (которое не зависит от θ):

$$\mathbf{E}_\theta(T|S) = \int T(\vec{x})\mathsf{P}_\theta(d\vec{x}|S).$$

В силу леммы можно опускать индекс θ у таких условных ожиданий.

Теорема Блекуэлла-Рао-Колмогорова. Пусть $\hat{\theta} \in K_b$ — оценка параметра θ. Тогда $\theta^* = \mathbf{E}(\hat{\theta}|S)$ — оценка того же класса K_b, более эффективная, чем $\hat{\theta}$ [6].

Доказательство. Заметим сначала, что

$$\mathbf{E}_\theta\theta^* = \mathbf{E}(\mathbf{E}(\hat{\theta}|S)) = \mathbf{E}_\theta\hat{\theta} = \theta + b(\theta).$$

Поэтому $\theta^* \in K_b$ — имеет то же смещение $b(\theta)$, что и $\hat{\theta}$. Далее,

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \theta^*)^2 + 2(\hat{\theta} - \theta^*)(\theta^* - \theta) + (\theta^* - \theta)^2.$$

Вычислим

$$\mathbf{E}_\theta[(\hat{\theta} - \theta^*)(\theta^* - \theta)] = \mathbf{E}_\theta[\mathbf{E}[(\hat{\theta} - \theta^*)(\theta^* - \theta)|S]]$$

(это равенство — формула полного математического ожидания — см. приложение D). Вынося "локально постоянный"множитель $\theta^* - \theta$, получаем для внутреннего (условного) ожидания

$$\begin{aligned}\mathbf{E}[(\hat{\theta} - \theta^*)(\theta^* - \theta)|S] &= (\theta^* - \theta)\mathbf{E}[\hat{\theta} - \theta^*|S] \\ &= (\theta^* - \theta)[\mathbf{E}(\hat{\theta}|S) - \mathbf{E}(\theta^*|S)] = (\theta^* - \theta)[\theta^* - \theta^*] = 0.\end{aligned}$$

Поэтому

$$\mathbf{E}_\theta[(\hat{\theta} - \theta^*)(\theta^* - \theta)] = 0$$

и

$$\mathbf{E}_\theta(\hat{\theta} - \theta)^2 = \mathbf{E}_\theta(\hat{\theta} - \theta^*)^2 + \mathbf{E}_\theta(\theta^* - \theta)^2 \geq \mathbf{E}_\theta(\theta^* - \theta)^2,$$

[6] Согласно приложению D, θ^* представляется в виде $f(S)$.

что и требовалось доказать.

Кстати, из проведенного рассуждения следует, что равенство эффективностей получается в единственном случае: $\theta^* = \hat{\theta}$ с вероятностью 1 (при этом уже первоначальная оценка $\hat{\theta}$ является функцией достаточной статистики).

Следствие. Эффективные в классах K_b оценки являются функциями достаточной статистики.

Разумеется, самый важный из всех классов K_b — класс несмещенных оценок.

Приведем два примера использования теоремы Блекуэлла-Рао-Колмогорова (справедливости ради следует отметить, что эффективные оценки в этих примерах нам уже известны).

Примеры 1 и 2. Оценка вероятности успеха и оценка параметра распределения Пуассона.

В обоих случаях берем (несостоятельную) несмещенную оценку X_1 и вычисляем для нее условное математическое ожидание при условии достаточной статистики $S = X_1 + \cdots + X_N$. Имеем по соображениям симметрии

$$\mathbf{E}(X_1|S) = \mathbf{E}(X_2|S) = \cdots = \mathbf{E}(X_N|S).$$

Сумма этих (одинаковых) величин есть

$$\mathbf{E}(S|S) = S.$$

Поэтому

$$\mathbf{E}(X_1|S) = \frac{S}{N} \quad (= \mathbf{E}(X_i|S), i = 2, \ldots, N).$$

Для модификации примера 5, обсуждавшейся в предыдущем параграфе, оценки $\tilde{a}$ и $\tilde{b} - 1$ параметра θ не могут быть улучшены этим приемом — теоремы Блекуэлла-Рао-Колмогорова здесь недостаточно для нахождения эффективной оценки.

Мы сейчас выделим дополнительное свойство достаточной статистики — полноту, позволяющее сразу указывать эффективные оценки.

Достаточная статистика S называется **полной**, если

$$\mathbf{E}f(S) \equiv 0 \Longrightarrow f(S) \equiv 0$$

(точнее, $\mathsf{P}_\theta(f(S) = 0) \equiv 1$). В этом определении символ $\equiv$ означает "тождественно по θ".

Теорема. Пусть S — полная достаточная статистика, $\hat{\theta} \in K_b$. Тогда оценка $\theta^* = \mathbf{E}(\hat{\theta}|S)$ эффективна в классе K_b.

Доказательство крайне просто. Пусть $\tilde{\theta} \in K_b$ эффективнее θ^*. Тогда $\tilde{\theta}^* = \mathbf{E}(\tilde{\theta}|S)$ еще эффективнее (в K_b). По одному из свойств условного математического ожидания, см. приложение D, $\theta^* - \tilde{\theta}^*$ — функция от S. Но $\mathbf{E}_\theta(\theta^* - \tilde{\theta}^*) = 0$, т.к. обе эти оценки имеют одинаковое смещение $b(\theta)$. По свойству полноты тогда $\tilde{\theta}^* = \theta^*$. Теорема доказана.

Проверка полноты достаточной статистики может оказаться трудной аналитической задачей. Проиллюстрируем на наших примерах, как она может проводиться.

Пример 1 мы оставим читателям в качестве упражнения.

Пример 2. Запишем подробно равенство

$$\mathbf{E}_\theta f(S) = 0.$$

Согласно параграфу 1.6 статистика $S = X_1 + \cdots + X_N$ имеет распределение Пуассона с параметром $N\lambda$. Поэтому получаем

$$\sum_{k=0}^{\infty} f(k) \frac{(N\lambda)^k}{k!} e^{-n\lambda} \equiv 0.$$

Сокращая экспоненту, получаем

$$\sum_{k=0}^{\infty} \frac{N^k f(k)}{k!} \lambda^k \equiv 0.$$

Из курса высшей математики известно, что если сходящийся степенной ряд тождественно равен нулю на некотором невырожденном промежутке, содержащем точку 0, то все его коэффициенты равны нулю. Поскольку $N^k/k! \neq 0$, получаем $f(k) = 0$ при всех $k = 0, 1, \ldots$.

Пример 4 (гамма-распределение). Снова мы ограничимся случаем известного p, когда достаточной статистикой является сумма $S = X_1 + \cdots + X_N$. По свойству воспроизводимости (см. параграф 1.6) случайная величина S имеет распределение $\Gamma(\alpha, Np)$. Поэтому равенство $\mathbf{E}_\alpha f(S) \equiv 0$ приобретает вид

$$\int_0^\infty f(x) \frac{\alpha^{Np}}{\Gamma(Np)} x^{Np-1} e^{-\alpha x} dx \equiv 0.$$

Выражение вида

$$G(\alpha) = \int_O^\infty g(x) e^{-\alpha x} dx$$

называется преобразованием Лапласа функции $g(x)$. В теории этого преобразования доказывается, что

$$G(\alpha) \equiv 0 \Longrightarrow g(x) = 0 \quad \text{почти всюду}$$

(мы не приводим точной формулировки соответствующих предположений о g). Таким образом, должно выполняться равенство

$$f(x)x^{Np-1} = 0,$$

откуда и следует $f(x) = 0$ (почти всюду).

Доказанная полнота S позволяет утверждать, что (см. параграф 5)

$$\tilde{\alpha} = \frac{Np-1}{S}$$

— эффективная несмещенная оценка параметра α. Можно доказать, что при неизвестных α и p достаточная статистика $(X_1 + \cdots + X_N, X_1 \cdot X_2 \cdot \cdots \cdot X_N)$ полна (см. [1]).

Пример 5 (равномерное распределение).

Нам потребуется плотность распределения $S = (X_{\min}, X_{\max})$ (в параграфе 3 была получена лишь индивидуальная плотность $X_{\max}$). Вычисления проводятся так:

$$\mathsf{P}(u < X_{\min}, X_{\max} < v) = \left(\frac{v-u}{b-a}\right)^N, a < u < v < b.$$

Совместная плотность $X_{\min}$ и $X_{\max}$ получается отсюда дифференцированием: следует взять вторую смешанную производную с противоположным знаком

$$p_S(u,v) = \frac{N(N-1)(v-u)^{N-2}}{(b-a)^N}, a < u < v < b.$$

Запишем теперь равенство $\mathbf{E}f(S) = 0$ в развернутом виде:

$$\int_a^b \left(\int_a^v f(u,v) \frac{N(N-1)(v-u)^{N-2}}{(b-a)^N} du \right) dv = 0.$$

Считая $N \geq 2$ и сокращая постоянный множитель, получаем

$$\int_a^b \left(\int_a^v f(u,v)(v-u)^{N-2} du \right) dv = 0.$$

Дифференцируя сначала по b, а затем по a, последовательно находим

$$\int_a^b f(u,b)(b-u)^{N-2}du = 0,$$

$$f(a,b)(b-a)^{N-2} = 0$$

(тождественно по a и b, $a<b$). Поэтому $f=0$ и достаточная статистика S полна. Отсюда следует (см. параграфы 6 и 3), что $(\tilde{a},\tilde{b})$ — эффективная несмещенная оценка двумерного параметра (a,b).

В модифицированной задаче статистика S, разумеется, не полна. Любая линейная комбинация оценок $\theta^*=\tilde{a}$ и $\theta^{**}=\tilde{b}-1$ вида

$$c\theta^* + (1-c)\theta^{**}$$

будет несмещенной и одновременно функцией от достаточной статистики.

Определим среди них оценку с минимальной дисперсией.

$$\mathbf{V}(c\theta^* + (1-c)\theta^{**}) = c^2\mathbf{V}\theta^* + 2c(1-c)\mathsf{cov}(\theta^*,\theta^{**}) + (1-c)^2\mathbf{V}\theta^{**}.$$

По соображениям симметрии

$$\mathbf{V}\theta^* = \mathbf{V}\theta^{**}.$$

Легко сообразить, что минимум квадратичного по c выражения, инвариантного при замене c на $1-c$, достигается при $c=1/2$.

Соответствующая оценка имеет вид

$$\frac{\theta^*+\theta^{**}}{2} = \frac{\tilde{a}+\tilde{b}-1}{2} = \frac{1}{2}\left(\frac{N}{N-1}X_{\min} - \frac{1}{N-1}X_{\max} + \frac{N}{N-1}X_{\max} - \frac{1}{N-1}X_{\min} - 1\right) = \frac{1}{2}\left(X_{\min}+X_{\max}-1\right).$$

Вычисления показывают, что

$$\mathbf{V}\left(\frac{\theta^*+\theta^{**}}{2}\right) = \frac{(b-a)^2}{2(N+1)(N+2)} = \frac{N+1}{2N}\mathbf{V}\theta^*,$$

так что полусумма почти вдвое эффективнее, чем каждая из оценок θ^*,θ^{**}.

Остался нерассмотренным самый важный пример 3 — нормальное распределение. Мы уже знаем, что $\bar{X}$ — эффективная несмещенная

оценка математического ожидания a (в том числе и при наличии мешающего параметра σ). Сформулируем аналогичный результат для дисперсии σ^2. Если предположить, что a известно, то эффективной в K_0 будет оценка

$$\overline{(X-a)^2} = \frac{1}{N}\sum_{i=1}^{N}(X_i - a)^2.$$

Этот результат мы оставляем читателю. В более реалистичной ситуации, когда a неизвестно (т.е. является мешающим параметром), эффективной в K_0 оценкой дисперсии σ^2, как уже упоминалось в параграфе 5, является

$$S^2_{\text{испр.}} = \frac{N}{N-1}S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

Мы сейчас докажем это, основываясь на идеях, близких к проверке полноты, хотя полнота при этом не будет ни доказываться, ни даже упоминаться. Итак, пусть $\widehat{\sigma^2}$ — некоторая несмещенная оценка дисперсии. Без ограничения общности можно считать ее функцией от достаточной статистики

$$(S_1, S_2) = (X_1 + \cdots + X_N, X_1^2 + \cdots + X_N^2)$$

и представить в виде

$$\widehat{\sigma^2} = S^2_{\text{испр.}} + f(S_1, S_2),$$

где $\mathbf{E}f(S_1, S_2) = 0$. Докажем, что $S^2_{\text{испр.}}$ и $f(S_1, S_2)$ не коррелируют. Этого достаточно, т.к. тогда

$$\mathbf{V}(\widehat{\sigma^2}) = \mathbf{V}(S^2_{\text{испр.}}) + \mathbf{V}f(S_1, S_2) \geq \mathbf{V}(S^2_{\text{испр.}})$$

(на самом деле, см. [1], достаточная статистика (S_1, S_2) полна). Имеем

$$\begin{aligned}\mathsf{cov}(S^2_{\text{испр.}}, f(S_1, S_2)) &= \mathbf{E}[S^2_{\text{испр.}} \cdot f(S_1, S_2)] \\ &= \mathbf{E}[(NS_2 - N^2S_1^2)f(S_1, S_2)]\end{aligned}$$

и мы проверим, что

$$\begin{aligned}\mathbf{E}[S_2 f(S_1, S_2)] &= \mathbf{E}[S_1^2 f(S_1, S_2)] \\ &= \mathbf{E}[S_1 f(S_1, S_2)] = 0 \quad (2.7)\end{aligned}$$

(отсюда сразу следует желаемая некоррелированность).

Запишем развернутым образом равенство $\mathbf{E}f(S_1, S_2) = 0$:

$$\int_{\mathbb{R}^N} p(\vec{x}) f(S_1(\vec{x}), S_2(\vec{x}))d\vec{x} = 0,$$

где

$$p(\vec{x}) = (2\pi)^{-N/2}\sigma^{-N}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - a)^2\}.$$

Сокращая постоянные множители, не обращающиеся в нуль и опуская аргумент $\vec{x}$ у функций S_1 и S_2, перепишем это равенство в виде

$$\int_{\mathbb{R}^N} f(S_1, S_2)\exp\{-\frac{1}{2\sigma^2}(S_2 - 2aS_1)\}d\vec{x} = 0. \tag{2.8}$$

Дифференцируя дважды по a, последовательно получаем

$$\int_{\mathbb{R}^N} S_1 f(S_1, S_2)\exp\{-\frac{1}{2\sigma^2}(S_2 - 2aS_1)\}d\vec{x} = 0,$$

$$\int_{\mathbb{R}^N} S_1^2 f(S_1, S_2)\exp\{-\frac{1}{2\sigma^2}(S_2 - 2aS_1)\}d\vec{x} = 0.$$

Восстанавливая сокращенные множители, записываем эти равенства в виде

$$\mathbf{E}[S_1 f(S_1, S_2)] = 0,$$

$$\mathbf{E}[S_1^2 f(S_1, S_2)] = 0.$$

Возвращаясь к (2.8) и дифференцируя теперь по σ, получаем аналогичным образом

$$\mathbf{E}[(S_2 - 2aS_1)f(S_1, S_2)] = 0,$$

откуда

$$\mathbf{E}[S_2 f(S_1, S_2)] = 0.$$

Все равенства (2.7) получены. Как уже было указано, из этого вытекает эффективность $S^2_{\text{испр.}}$.

Отметим в заключение параграфа еще один полезный факт.

Теорема. Оценки максимального правдоподобия являются функциями от достаточной статистики.

Доказательство. По теореме факторизации

$$L(\theta) = h(\vec{X})\psi(S, \theta),$$

где S — достаточная статистика. Поскольку первый множитель от параметра не зависит, точки максимума для функций $L(\theta)$ и $\psi(S, \theta)$ — одни и те же. Однако точка максимума $\psi(S, \theta)$, очевидно, зависит лишь от S.

2.8 Асимптотические свойства оценок максимального правдоподобия

В этом параграфе пойдет речь об основных асимптотических свойствах оценок максимального правдоподобия. Эти свойства сформулированы ниже в виде теорем 1 – 4. Мы не будем приводить ни доказательства этих сложных результатов, ни точные формулировки соответствующих условий регулярности, однако постараемся объяснить идейную сторону доказательств.

Нам потребуются некоторые предварительные определения.

Оценка $\hat{\theta}$ параметра θ называется асимптотически нормальной с коэффициентом разброса $\sigma^2 > 0$, если функция распределения величины

$$\sqrt{N}\frac{\hat{\theta} - \theta}{\sigma}$$

слабо сходится к функции распределения стандартного нормального закона:

$$\mathsf{P}(\sqrt{N}\frac{\hat{\theta} - \theta}{\sigma} < z) \longrightarrow \Phi(z), N \to \infty.$$

Коэффициент разброса σ^2 может при этом зависеть от θ.

Далее, будем говорить, следуя [1], что оценка $\hat{\theta}$ принадлежит классу $\tilde{K}_0$, если ее смещение обладает свойствами: 1) $\sqrt{N}b(\theta) \to 0$ при $N \to \infty$ и произвольном фиксированном θ; 2) производная $b'(\theta)$ существует, причем $b'(\theta) \to 0$ при $N \to \infty$ и произвольном фиксированном θ.

Теорема 1. При некоторых условиях регулярности оценка $\hat{\theta}_{ML}$ сильно состоятельна.

Теорема 2. При некоторых условиях регулярности оценка $\hat{\theta}_{ML}$ асимптотически нормальна с коэффициентом разброса $\frac{1}{i(\theta)}$.

Теорема 3. При некоторых условиях регулярности оценка $\hat{\theta}_{ML}$ лежит в классе $\tilde{K}_0$.

Теорема 4. При некоторых условиях регулярности оценка $\hat{\theta}_{ML}$ асимптотически эффективна в классе $\tilde{K}_0$.

Заметим сначала, что теорема 1 вытекает из теоремы 2. Далее, теорема 4 легко следует из теорем 2, 3 и неравенства Рао-Крамéра:

$$\mathbf{E}(\hat{\theta} - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{Ni(\theta)} + b^2(\theta).$$

Действительно, предположим, что $\hat{\theta} \in \tilde{K}_0$, и обозначим правую часть неравенства через $g_N(\theta)$. Из определения класса $\tilde{K}_0$ вытекает, что

$$Ng_N(\theta) \to \frac{1}{i(\theta)}, N \to \infty.$$

Из теоремы 2 следует, что

$$N\mathbf{V}\hat{\theta}_{ML} \to \frac{1}{i(\theta)}.$$

Поскольку

$$\mathbf{E}(\hat{\theta}_{ML} - \theta)^2 = \mathbf{V}\hat{\theta}_{ML} + b^2(\theta)$$

и по теореме 3 $\hat{\theta}_{ML} \in \tilde{K}_0$, получаем, что

$$N\mathbf{E}(\hat{\theta}_{ML} - \theta)^2 \to \frac{1}{i(\theta)}.$$

Наконец, для произвольной оценки $\hat{\theta}$ класса $\tilde{K}_0$ имеем

$$\frac{\mathbf{E}(\hat{\theta} - \theta)^2}{\mathbf{E}(\hat{\theta}_{ML} - \theta)^2} \geq \frac{g_N(\theta)}{\mathbf{E}(\hat{\theta}_{ML} - \theta)^2} = \frac{Ng_N(\theta)}{N\mathbf{E}(\hat{\theta}_{ML} - \theta)^2} \to \frac{1/i(\theta)}{1/i(\theta)} = 1.$$

Обсудим теперь теорему 2.

Определим функцию

$$Y(u) = l(\theta + \frac{u}{\sqrt{N}}) - l(\theta),$$

где l — логарифмическая функция правдоподобия, а θ — истинное значение параметра. Точку максимума функции $Y(u)$ обозначим u^*. Очевидно, что

$$\hat{\theta}_{ML} = \theta + \frac{u^*}{\sqrt{N}}.$$

Разложим $Y(u)$ по Тейлору:

$$Y(u) = l(\theta) + \frac{u}{\sqrt{N}}l'(\theta) + \frac{1}{2}u^2\left[\frac{l''(\theta)}{N} + o(1)\right] - l(\theta)$$
$$= u\frac{l'(\theta)}{\sqrt{N}} + \frac{1}{2}u^2\left[\frac{l''(\theta)}{N} + o(1)\right].$$

По определению

$$l(\theta) = \sum_{i=1}^{N} \ln p_\theta(X_i),$$

$$l'(\theta) = \sum_{i=1}^{N} (\ln p_\theta(X_i))',$$

$$l''(\theta) = \sum_{i=1}^{N} (\ln p_\theta(X_i))''.$$

Все эти суммы состоят из независимых одинаково распределенных величин. Из доказательства неравенства Рао-Крамéра мы знаем, что

$$\mathbf{E}(\ln p_\theta(X_i))' = 0,$$

$$\mathbf{V}[(\ln p_\theta(X_i))'] = -\mathbf{E}(\ln p_\theta(X_i))'' = i(\theta).$$

По центральной предельной теореме Левú (см. параграф 1.4) распределение величины

$$\xi_N = \frac{l'(\theta)}{\sqrt{N}}$$

слабо сходится к нормальному закону $\mathbf{N}(0, i(\theta))$. По теореме Хинчина (см. там же) величина

$$\frac{l''(\theta)}{N} \to -i(\theta)$$

по вероятности. Поэтому тейлоровское разложение можно переписать в виде

$$Y(u) = u\xi_N - \frac{u^2}{2} i(\theta)[1 + o(1)].$$

Тогда точка максимума этой функции запишется как

$$u^* = \frac{\xi_N}{i(\theta)}[1 + o(1)]. \tag{2.9}$$

Из последнего соотношения следует, что $\hat{\theta}_{ML}$ асимптотически нормальна с разбросом $\frac{1}{i(\theta)}$ (асимптотическая дисперсия $i(\theta)$ величины ξ_N умножается на квадрат постоянного множителя $\frac{1}{i(\theta)}$).

Перейдем, наконец, к теореме 3. Первое условие ($\sqrt{N}b(\theta) \to 0$) проверяется на основе соотношения

$$\sqrt{N}b(\theta) = \sqrt{N}\mathbf{E}(\hat{\theta}_{ML} - \theta) = \mathbf{E}u^*.$$

Достаточно сослаться на (7.11) и на сходимость распределения величины ξ_N к нормальному закону с нулевым средним значением:

$$\mathbf{E}u^* = \frac{\mathbf{E}\xi_N}{i(\theta)}[1 + o(1)] \to 0.$$

Второе условие ($b'(\theta) \to 0$) установить чуть сложнее. Имеем

$$1 + b'(\theta) = (\theta + b(\theta))' = (\mathbf{E}\hat{\theta}_{ML})' = \left(\int \hat{\theta}_{ML}(\vec{x})p_\theta(\vec{x})d\vec{x}\right)' =$$
$$= \int \hat{\theta}_{ML}(\vec{x})\frac{p'_\theta(\vec{x})}{p_\theta(\vec{x})}p_\theta(\vec{x})d\vec{x} = \int \hat{\theta}_{ML}(\vec{x})(\ln p_\theta(\vec{x}))'p_\theta(\vec{x})d\vec{x} =$$
$$= \mathbf{E}[\hat{\theta}_{ML}l'(\theta)] = \mathbf{E}[(\hat{\theta}_{ML} - \theta)l'(\theta)] =$$
$$= \mathbf{E}[(\hat{\theta}_{ML} - \theta)\sqrt{N}\xi_N] = \mathbf{E}[u^*\xi_N] = \mathbf{E}\left[\frac{\xi_N^2}{i(\theta)}(1 + o(1))\right] \to 1.$$

Отсюда вытекает искомое $b'(\theta) \to 0$.

Дадим неформальный комментарий к приведенным выше теоремам (см. также [1]). Рассматривать оценки, не принадлежащие классу $\tilde{K}_0$, по-видимому, просто нецелесообразно — неравенство Рао-Крамéра показывает, что их относительная эффективность ниже, по крайней мере, асимптотически. А тогда теорема 4, по существу, утверждает, что, в том же асимптотическом смысле, оценка максимального правдоподобия неулучшаема. При фиксированном N такое улучшение, конечно, может оказаться возможным (на величину $o(1/N)$).

Теоремами этого параграфа мы будем пользоваться и для других целей (см. параграф 3.1).

2.9 Эквивариантные оценки параметра сдвига

Как указывалось в параграфе 1, для нахождения эффективных оценок приходится разумным образом сужать класс всевозможных оценок. Несмещенные оценки (класс K_0) и оценки с фиксированным смещением (классы K_b) — примеры такого сужения. Сейчас мы рассмотрим еще один полезный класс оценок — эквивариантные оценки параметра сдвига (в книге Боровкова [1] можно найти аналогичное обсуждение эквивариантных оценок параметра масштаба, а также общую теорию эквивариантности).

Будем говорить, что θ — параметр сдвига, если параметрическое семейство плотностей $p(x;\theta)$ задается формулой

$$p(x;\theta) = p(x-\theta),$$

т.е. все плотности этого семейства получаются сдвигом аргумента из одной и той же плотности $p(x)$. Предположим также, что область изменения θ — вся числовая ось $\mathbb{R}$.

Оценка $\hat{\theta}$ параметра сдвига θ называется **эквивариантной**, если для любого $c \in \mathbb{R}$

$$\hat{\theta}(X_1 + c, \ldots, X_N + c) = \hat{\theta}(X_1, \ldots, X_N) + c.$$

Для краткости мы будем писать подобные равенства в виде

$$\hat{\theta}(\vec{X} + c^{\rightarrow}) = \hat{\theta}(\vec{X}) + c.$$

Здесь $c^{\rightarrow}$ —вектор, все компоненты которого равны c. Класс всех эквивариантных оценок параметра θ мы обозначим K_{eq}.

Статистику S будем называть инвариантной, если

$$S(\vec{X} + c^{\rightarrow}) = S(\vec{X}).$$

В очевидном смысле инвариантные статистики не содержат информации о параметре сдвига. Примеры таких статистик легко строятся с помощью статистики

$$S_0 = (X_2 - X_1, X_3 - X_1, \ldots, X_N - X_1).$$

Очевидно, любая статистика вида $f(S_0)$ инвариантна.

Нам потребуется простой вспомогательный результат об эквивариантных оценках.

Лемма. Если $\hat{\theta}$ — эквивариантная оценка, то

$$\mathbf{E}_\theta \hat{\theta} = \mathbf{E}_0 \hat{\theta} + \theta.$$

Действительно,

$$\mathbf{E}_\theta \hat{\theta} = \int_{\mathbb{R}^N} \hat{\theta}(\vec{x}) p(\vec{x} - \theta^{\rightarrow}) d\vec{x} = \int_{\mathbb{R}^N} \hat{\theta}(\vec{y} + \theta^{\rightarrow}) p(\vec{y}) d\vec{y}$$
$$= \int_{\mathbb{R}^N} [\hat{\theta}(\vec{y}) + \theta] p(\vec{y}) d\vec{y} = \mathbf{E}_0 \hat{\theta} + \theta.$$

Сформулируем теперь основной результат параграфа.

Теорема.

1. Оценка Питмена
$$\hat{\theta}^0 = \frac{\int_{-\infty}^{\infty} up(\vec{X} - u^{\rightarrow})du}{\int_{-\infty}^{\infty} p(\vec{X} - u^{\rightarrow})du}$$
является единственной эффективной в классе K_{eq} оценкой;

2. $\hat{\theta}^0$ — несмещенная оценка;

3. если $\hat{\theta} \in K_{eq}$, то
$$\hat{\theta} - \mathbf{E}_0(\hat{\theta}|S_0) = \hat{\theta}^0.$$

Доказательство.
Разобьем для удобства все доказательство на части.
1. Докажем, что оценки вида
$$\hat{\theta} - \mathbf{E}_0(\hat{\theta}|S_0), \hat{\theta} \in K_{eq}$$
— несмещенные.

Для этого заметим сначала (см. приложение D), что
$$\mathbf{E}_0(\hat{\theta}|S_0) = f(S_0)$$
— инвариантная статистика. Поэтому
$$\begin{aligned}\mathbf{E}_\theta f(S_0) &= \int_{\mathbb{R}^N} f(S_0(\vec{x}))p(\vec{x} - \theta^{\rightarrow})d\vec{x} \\ &= \int_{\mathbb{R}^N} f(S_0(\vec{y} + \theta^{\rightarrow}))p(\vec{y})d\vec{y} = \int_{\mathbb{R}^N} f(S_0(\vec{y}))p(\vec{y})d\vec{y} \\ &= \mathbf{E}_0 f(S_0) = \mathbf{E}_0(\mathbf{E}_0(\hat{\theta}|S_0)) = \mathbf{E}_0\hat{\theta}.\end{aligned}$$

Отсюда при помощи леммы
$$\mathbf{E}_\theta(\hat{\theta} - \mathbf{E}_0(\hat{\theta}|S_0)) = \mathbf{E}_\theta\hat{\theta} - \mathbf{E}_\theta f(S_0) = \mathbf{E}_0\hat{\theta} + \theta - \mathbf{E}_0\hat{\theta} = \theta.$$

2. Докажем, что оценки вида $\hat{\theta} - \mathbf{E}_0(\hat{\theta}|S_0)$ — эквивариантные. Запишем сначала такую оценку в виде
$$\hat{\theta}(\vec{X}) - f(S_0(\vec{X})).$$
Заменяя $\vec{X}$ на $\vec{X} + c^{\rightarrow}$, получаем
$$\hat{\theta}(\vec{X} + c^{\rightarrow}) - f(S_0(\vec{X} + c^{\rightarrow})) = \hat{\theta}(\vec{X}) + c - f(S_0(\vec{X})),$$
что и требовалось доказать.

3. Докажем, что для любой статистики S с конечным математическим ожиданием $\mathbf{E}_0(S)$ справедлива формула

$$\mathbf{E}_0(S|S_0) = \frac{\int_{-\infty}^{\infty} S(\vec{X} - u^{\rightarrow})p(\vec{X} - u^{\rightarrow})du}{\int_{-\infty}^{\infty} p(\vec{X} - u^{\rightarrow})du}.$$

Для этого обозначим правую часть написанного равенства через S^* и докажем два определяющих свойства условного математического ожидания (см. приложение D). Сначала проверим, что S^* есть функция от S_0. Для этого достаточно сделать замену переменной $v = X_1 - u$:

$$\begin{aligned} S^* = \int_{-\infty}^{\infty} & S(v, X_2 - X_1 + v, \ldots, X_N - X_1 + v) \\ & \cdot p(v, X_2 - X_1 + v, \ldots, X_N - X_1 + v)dv \\ & \cdot \left(\int_{-\infty}^{\infty} p(v, X_2 - X_1 + v, \ldots, X_N - X_1 + v)dv \right)^{-1}. \end{aligned}$$

Докажем теперь второе свойство — равенство математических ожиданий. Зафиксируем ограниченную функцию $Z = Z(S_0)$ и докажем, что

$$\mathbf{E}_0(ZS^*) = \mathbf{E}_0(ZS).$$

Подставляя определения, меняя порядок интегрирования и делая замену $\vec{y} = \vec{x} - u^{\rightarrow}$, получим

$$\begin{aligned} & \mathbf{E}_0(ZS^*) \\ & \quad = \int Z(S_0(\vec{x})) \left(\int S(\vec{x} - u^{\rightarrow})p(\vec{x} - u^{\rightarrow})du \right) \\ & \qquad \cdot \left(\int p(\vec{x} - v^{\rightarrow})dv \right)^{-1} p(\vec{x})d\vec{x} \\ & \quad = \int \left(\int Z(S_0(\vec{x}))S(\vec{x} - u^{\rightarrow})p(\vec{x} - u^{\rightarrow})p(\vec{x}) \right. \\ & \qquad \left. \cdot \left(\int p(\vec{x} - v^{\rightarrow})dv \right)^{-1} d\vec{x} \right) du \\ & \quad = \int \left(\int Z(S_0(\vec{y} + u^{\rightarrow}))S(\vec{y})p(\vec{y})p(\vec{y} + u^{\rightarrow}) \right. \\ & \qquad \left. \cdot \left(\int p(\vec{y} + u^{\rightarrow} - v^{\rightarrow})dv \right)^{-1} d\vec{y} \right) du. \end{aligned}$$

Воспользуемся теперь инвариантностью S_0, а затем снова поменяем порядок интегрирования. Получим

$$\mathbf{E}(ZS^*) = \int Z(S_0(\vec{y}))S(\vec{y})p(\vec{y}) \cdot \left(\int p(\vec{y}+u^{\rightarrow}) \left(\int p(\vec{y}+u^{\rightarrow}-v^{\rightarrow})dv \right)^{-1} du \right) d\vec{y}.$$

Остается заметить, что внутренний (по переменной u) интеграл равен 1:

$$\int p(\vec{y}+u^{\rightarrow}) \left(\int p(\vec{y}+u^{\rightarrow}-v^{\rightarrow})dv \right)^{-1} du = \int p(\vec{y}+u^{\rightarrow}) \left(\int p(\vec{y}+w^{\rightarrow})dw \right)^{-1} du = \frac{\int p(\vec{y}+u^{\rightarrow})du}{\int p(\vec{y}+w^{\rightarrow})dw} = 1.$$

Окончательно получаем

$$\mathbf{E}_0(ZS^*) = \int Z(S_0(\vec{y}))S(\vec{y})p(\vec{y})d\vec{y} = \mathbf{E}(ZS).$$

Равенство $\mathbf{E}_0(S|S_0) = S^*$ доказано.

4. Докажем утверждения 2 и 3 теоремы. Согласно предыдущему пункту доказательства

$$\hat{\theta} - \mathbf{E}_0(\hat{\theta}|S_0) = \hat{\theta}(\vec{X}) - \frac{\int \hat{\theta}(\vec{X}-u^{\rightarrow})p(\vec{X}-u^{\rightarrow})du}{\int p(\vec{X}-u^{\rightarrow})du} = \frac{\int [\hat{\theta}(\vec{X}) - \hat{\theta}(\vec{X}-u^{\rightarrow})]p(\vec{X}-u^{\rightarrow})du}{\int p(\vec{X}-u^{\rightarrow})du} = \frac{\int up(\vec{X}-u^{\rightarrow})du}{\int p(\vec{X}-u^{\rightarrow})du} = \hat{\theta}^0$$

(мы пользуемся эквивариантностью $\hat{\theta}$).

Таким образом, утверждение 3 теоремы доказано. Утверждение 2 теперь вытекает из п.1 доказательства.

5. Докажем, что для любой эквивариантной оценки $\hat{\theta}$ $\mathbf{E}_\theta(\hat{\theta}-\theta)^2$ не

зависит от θ. Действительно,

$$\mathbf{E}_\theta(\hat\theta-\theta)^2=\int_{\mathbb{R}^N}(\hat\theta(\vec x)-\theta)^2p(\vec x-\theta^{\rightarrow})d\vec x=$$
$$=\int_{\mathbb{R}^N}[\hat\theta(\vec y+\theta^{\rightarrow})-\theta]^2p(\vec y)d\vec y=\int_{\mathbb{R}^N}[\hat\theta(\vec y)]^2p(\vec y)d\vec y=\mathbf{E}_0(\hat\theta^2).$$

6.Докажем, наконец, утверждение 1 теоремы — эффективность. С учетом п.5 имеем

$$\mathbf{E}_\theta(\hat\theta-\theta)^2=\mathbf{E}_0(\hat\theta^2)=\mathbf{E}_0[(\hat\theta^0+\mathbf{E}_0(\hat\theta|S_0))^2]$$
$$=\mathbf{E}_0((\hat\theta^0)^2)+\mathbf{E}_0[(\mathbf{E}_0(\hat\theta|S_0))^2]+2\mathbf{E}_0[\hat\theta^0\mathbf{E}_0(\hat\theta|S_0)].$$

Проверим, что последнее слагаемое равно нулю. По формуле полного математического ожидания

$$\mathbf{E}_0[\hat\theta^0\mathbf{E}_0(\hat\theta|S_0)]=\mathbf{E}_0[\mathbf{E}_0[\hat\theta^0\mathbf{E}_0(\hat\theta|S_0)]].$$

Проверим, что

$$\mathbf{E}_0[\hat\theta^0\mathbf{E}_0(\hat\theta|S_0)|S_0]=0.$$

Действительно, "локально постоянный"множитель $\mathbf{E}_0(\hat\theta|S_0)$ выносится наружу, а

$$\mathbf{E}_0[\hat\theta^0|S_0]=\mathbf{E}_0[\hat\theta-\mathbf{E}_0(\hat\theta|S_0)|S_0]$$
$$=\mathbf{E}_0(\hat\theta|S_0)-\mathbf{E}_0(\mathbf{E}_0(\hat\theta|S_0)|S_0)=0.$$

Для завершения доказательства замечаем, что

$$\mathbf{E}_\theta(\hat\theta-\theta)^2=\mathbf{E}_0((\hat\theta^0)^2)+\mathbf{E}_0[(\mathbf{E}_0(\hat\theta|S_0))^2]$$
$$\geq\mathbf{E}_0((\hat\theta^0)^2)=\mathbf{E}_\theta(\hat\theta^0-\theta)^2$$

(последнее равенство следует из эквивариантности $\hat\theta^0$ (см. п.2) и п.5).

Рассмотрим теперь два примера.

Пример 3. Однопараметрическое семейство нормальных распределений $\mathbf{N}(a,1)$.

Для построения эффективной эквивариантной оценки $\hat a^0$ заметим, что

$$p(\vec X-a^{\rightarrow})=(2\pi)^{-N/2}\exp\left\{-\frac12\sum_{i=1}^N(X_i-a)^2\right\}=$$
$$=N^{-1/2}(2\pi)^{-N/2}\exp\left\{-\frac12\sum_{i=1}^N(X_i-\bar X)^2\right\}\cdot\sqrt{N}\varphi(\sqrt{N}(a-\bar X)).$$

Первый множитель при вычислении оценки Питмена сокращается, и мы получаем

$$\hat{a}^0 = \frac{\int a\sqrt{N}\varphi(\sqrt{N}(a-\bar{X}))da}{\int \sqrt{N}\varphi(\sqrt{N}(a-\bar{X}))da} = \bar{X}.$$

Действительно, по аргументу a функция

$$\sqrt{N}\varphi(\sqrt{N}(a-\bar{X}))$$

является плотностью нормального распределения $\mathbf{N}(\bar{X}, 1/N)$. Поэтому интеграл в знаменателе равен 1, а интеграл в числителе — среднему значению указанного нормального распределения.

Пример 5. Найдем эффективную эквивариантную оценку для параметра равномерного распределения на $\langle\theta, 1+\theta\rangle$. Имеем

$$p(\vec{X}-\theta) = \begin{cases} 1, & X_{\max}-1 \le \theta \le X_{\min}, \\ 0, & \text{иначе.} \end{cases}$$

Поэтому

$$\hat{\theta}^0 = \frac{\int_{X_{\max}-1}^{X\min} u du}{X_{\min} - X\max + 1}$$

$$= \frac{1}{2}\frac{X_{\min}^2 - (X_{\max}-1)^2}{X_{\min}-(X_{\max}-1)} = \frac{X_{\min}+X_{\max}-1}{2}.$$

2.10 Другие подходы к понятию оптимальной оценки

Мы рассмотрим два таких подхода, приводящие к байесовским и минимаксным оценкам.

Байесовский подход основан на предположении, что исследователю известны некоторые априорные предпочтения в множестве возможных значений параметра θ. Другими словами, предполагается, что фактическое значение параметра θ_{true} является реализовавшимся значением некоторой случайной величины θ с плотностью распределения $q(t)$.

Буквой t в этом параграфе мы далее будем обозначать конкретные значения параметра, а буквой θ — параметр как случайную величину.

Оценка θ^*, минимизирующая полное математическое ожидание

$$\mathbf{E}(\phi(\vec{X})-\theta)^2,$$

называется **байесовской**, отвечающей априорной плотности q. Здесь ϕ — переменная оценка, аргумент, по которому и производится минимизация. Слова "полное математическое ожидание"расшифровываются так:

$$\mathbf{E}(\phi(\vec{X})-\theta)^2 = \mathbf{E}_q(\mathbf{E}_t(\phi(\vec{X})-t)^2) = \int \mathbf{E}_t(\phi(\vec{X})-t)^2 q(t)dt,$$

т.е. как взвешенное с помощью априорной плотности q среднее значение мер эффективности $\mathbf{E}_t(\phi(\vec{X})-t)^2$. Другими словами, мы рассматриваем в пространстве $\mathbb{R}^{N+1}$ совместное распределение величин $X_1,\dots,X_N,\theta$, плотность которого задается формулой $p(\vec{x};t)q(t)$, и соответствующее математическое ожидание.

Из свойств условного математического ожидания $\mathbf{E}(\theta|\vec{X})$ (см. приложение D) вытекает, что именно оно дает нам байесовскую оценку. Для вычисления ее следует воспользоваться соответствующей условной плотностью

$$p(t|\vec{x}) = \frac{p(\vec{x};t)q(t)}{\int p(\vec{x};\tau)q(\tau)d\tau},$$

так что

$$\mathbf{E}(\theta|\vec{X}) = \frac{\int tp(\vec{X},t)q(t)dt}{\int p(\vec{X},t)q(t)dt} \tag{2.10}$$

При всей привлекательности предлагаемого в байесовском подходе усреднения, следует подчеркнуть, что убедительно мотивировать выбор того или иного априорного распределения обычно очень трудно. Впрочем, сторонники байесовского подхода считают предположение о существовании такого априорного распределения важнейшей частью своей теоретической концепции (см., например, [13]).

Заметим, что нормировка априорной плотности q несущественна — в формуле (2.10) нормирующие множители сокращаются. Поэтому в качестве (ненормированной) плотности априорного распределения можно взять, например, плотность вида

$$\exp(-t^2/2\sigma^2)$$

Тогда, например, в случае параметра сдвига при $\sigma \to \infty$ из формулы (2.10) в пределе получается эквивариантная оценка Питмена из параграфа 9.

Перейдем теперь к определению минимаксных оценок. Оценка θ^* называется **минимаксной**, если для любой другой оценки $\hat{\theta}$

$$\sup_t \mathbf{E}_t(\hat{\theta}-t)^2 \geq \sup_t \mathbf{E}_t(\theta^*-t)^2$$

(т.е. θ^* минимизирует супремум, стоящий в левой части этого неравенства).

Мы видим, что оба подхода — байесовский и минимаксный — предлагают свои способы сравнения оценок. Любые две оценки при этом становятся сравнимыми, но выбор того или иного способа сравнения остается открытым. Мы увидим ниже, что асимптотически все подходы дают примерно одно и то же.

Простейшая связь байесовости и минимаксности выражается следующей теоремой.

ТЕОРЕМА 1. Пусть θ^* — байесовская оценка, отвечающая некоторому априорному распределению q. Предположим, что для почти всех t, принадлежащих носителю $\mathbf{supp}\, q$ плотности q, математическое ожидание $\mathbf{E}_t(\theta^* - t)^2$ постоянно:

$$\mathbf{E}_t(\theta^* - t)^2 = c,$$

а для остальных t

$$\mathbf{E}_t(\theta^* - t)^2 \le c.$$

Тогда θ^* — минимаксная оценка.

Напомним, что носитель $\mathbf{supp}\, q$ по определению есть множество тех t, где $q(t) \neq 0$.

Докажем теорему 1. Пусть $\hat\theta$ — другая оценка. Тогда

$$\sup_t \mathbf{E}_t(\hat\theta - t)^2 \ge \int \mathbf{E}_t(\hat\theta - t)^2 q(t)dt$$

(взвешенное среднее не превосходит супремума). Правая часть написанного неравенства по предположению байесовости не меньше

$$\int \mathbf{E}_t(\theta^* - t)^2 q(t)dt = c = \sup_t \mathbf{E}_t(\theta^* - t)^2,$$

что и требовалось доказать.

Распределение q, отвечающее минимаксной оценке, называется наихудшим. К сожалению, оно не всегда существует — это может быть связано, в частности, с неограниченностью множества Θ изменения параметра θ. Приведем теорему, позволяющую обойти эту трудность.

ТЕОРЕМА 2. Предположим, что для оценки θ^* существует последовательность априорных плотностей q_k, такая, что при всех τ

$$\mathbf{E}_\tau(\theta^* - \tau)^2 \le \overline{\lim}_{r\to\infty} \int \mathbf{E}_t(\hat\theta^k - t)^2 q_k(t)dt$$

($\hat{\theta}^k$ — байесовская оценка, отвечающая q_k). Тогда θ^* минимаксна.

Доказательство почти не отличается от доказательства теоремы 1. Пусть $\hat{\theta}$ — другая оценка. Тогда

$$\sup_t \mathbf{E}_t(\hat{\theta}-t)^2 \geq \int \mathbf{E}_t(\hat{\theta}-t)q_k(t)dt \geq \int \mathbf{E}_t(\hat{\theta}^k-t)^2 q_k(t)dt.$$

Переходя к верхнему пределу при $k \to \infty$, получаем

$$\sup_t \mathbf{E}_t(\hat{\theta}-t)^2 \geq \overline{\lim} \int \mathbf{E}_t(\hat{\theta}^k-t)^2 q_k(t)dt \geq \mathbf{E}_\tau(\theta^*-\tau)^2.$$

Остается взять супремум по τ.

Рассмотрим теперь два примера.

Пример 3. Однопараметрическое семейство нормальных распределений $\mathbf{N}(a,1)$.

Возьмем в качестве априорного нормальное распределение $\mathbf{N}(0,\sigma^2)$ с (ненормированной) плотностью $\exp(-t^2/2\sigma^2)$ и найдем соответствующую байесовскую оценку. Апостериорная условная плотность $p(t|\vec{x})$ как функция аргумента t пропорциональна

$$\exp\left\{-\frac{t^2}{2\sigma^2}-\frac{1}{2}\sum_{i=1}^N (X_i-t)^2\right\},$$

т.е. является плотностью нормального распределения. Для нахождения параметров этого нормального распределения выделим полный квадрат в показателе:

$$\frac{t^2}{\sigma^2}+\sum_{i=1}^N (X_i-t)^2 = \left(\frac{1}{\sigma^2}+N\right)t^2 - 2t\sum_{i=1}^N X_i + \sum_{i=1}^N X_i^2$$
$$= \left(\frac{1}{\sigma^2}+N\right)\left[t-\frac{\sum_{i=1}^N X_i}{\frac{1}{\sigma^2}+N}\right]^2 + \cdots.$$

Таким образом, речь идет о нормальном распределении

$$\mathbf{N}\left(\frac{\bar{X}}{1+\frac{1}{N\sigma^2}}, \frac{\sigma^2}{1+N\sigma^2}\right).$$

Байесовская оценка — соответствующее математическое ожидание, т.е.

$$\hat{a}^\sigma = \frac{\bar{X}}{1+\frac{1}{N\sigma^2}}.$$

Докажем, что оценка

$$a^* = \bar{X} = \lim_{\sigma\to\infty} \hat{a}^\sigma$$

минимаксна (напомним, см. параграф 9, что она еще и эквивариантная эффективная). Для этого воспользуемся теоремой 2. Имеем

$$\mathbf{E}_\tau(a^* - \tau)^2 = \mathbf{V}_\tau \bar{X} = 1/N.$$

С другой стороны,

$$\overline{\lim}_{\sigma\to\infty} \int \mathbf{E}_t(\hat{a}^\sigma - t)^2 q_k(t)dt$$
$$= \overline{\lim}_{\sigma\to\infty} \mathbf{V}\hat{a}^\sigma = \overline{\lim}_{\sigma\to\infty} \frac{\sigma^2}{1+N\sigma^2} = \frac{1}{N}.$$

Условие теоремы 2 выполнено (со знаком равенства в неравенстве), так что a^* минимаксна.

Пример 1. Вероятность успеха.

Мы найдем минимаксную оценку с помощью теоремы 1, т.е. среди байесовских. В качестве априорного распределения для p естественно взять бета-распределение $B(\lambda_1, \lambda_2)$, подобрав его параметры надлежащим образом. Условная плотность $p(t|\vec{X})$ пропорциональна

$$t^S(1-t)^{N-S}t^{\lambda_1-1}(1-t)^{\lambda_2-1}$$

($S = X_1 + \cdots + X_N$ — суммарное число успехов), т.е. является плотностью бета-распределения $D(S+\lambda_1, N-S+\lambda_2)$. Байесовская оценка (т.е. соответствующее среднее значение) имеет вид

$$\hat{p}^{\lambda_1,\lambda_2} = \frac{S+\lambda_1}{N+\lambda_1+\lambda_2} = \frac{\bar{X} + \frac{\lambda_1}{N}}{1 + \frac{\lambda_1+\lambda_2}{N}}.$$

Тогда

$$\hat{p}^{\lambda_1,\lambda_2} - p = \frac{N}{N+\lambda_1+\lambda_2}\left[\bar{X} + \frac{\lambda_1}{N} - p(1 + \frac{\lambda_1+\lambda_2}{N})\right]$$

и

$$\begin{aligned}
&\mathbf{E}_p(\hat{p}^{\lambda_1,\lambda_2} - p)^2 \\
&\qquad = \frac{N^2}{(N+\lambda_1+\lambda_2)^2}\left[\mathbf{E}_p(\bar{X}-p)^2 + \left(\frac{\lambda_1}{N} - p\frac{\lambda_1+\lambda_2}{N}\right)^2\right] \\
&\qquad\quad = \frac{1}{(N+\lambda_1+\lambda_2)^2}\left[Np(1-p) + (\lambda_1 - p(\lambda_1+\lambda_2))^2\right] \\
&\qquad\qquad = \frac{1}{(N+\lambda_1+\lambda_2)^2}\left\{p^2[(\lambda_1+\lambda_2)^2 - N]\right. \\
&\qquad\qquad\qquad \left. -p[2\lambda_1(\lambda_1+\lambda_2) - N] + \lambda_1^2\right\}.
\end{aligned}$$

Последнее выражение не зависит от p при $\lambda_1 = \lambda_2 = \frac{1}{2}\sqrt{N}$. Таким образом, оценка

$$p^* = \frac{\bar{X} + \frac{1}{2\sqrt{N}}}{1 + \frac{1}{\sqrt{N}}}$$

удовлетворяет условиям теоремы 1 — она байесовская с априорным распределением $B(\sqrt{N}/2, \sqrt{N}/2)$ и

$$\mathbf{E}_p(p^* - p)^2 = \frac{N}{4(N+\sqrt{N})^2} = \frac{1}{4(1+\sqrt{N})^2}$$

не зависит от p. Поэтому p^* минимаксна. В то же время

$$\mathbf{E}_p(\bar{X} - p)^2 = \frac{p(1-p)}{N} < \mathbf{E}_p(p^* - p)^2$$

для всех p, удовлетворяющих неравенству

$$4p(1-p) < \frac{1}{(1+\frac{1}{\sqrt{N}})^2}.$$

Легко проверить, что дополнительная область представляет собой промежуток с центром в точке $1/2$, имеющий не такую уж малую длину $(4/N)^{1/4}(1+o(1))$. Даже при $N = 40000$ длина этого промежутка еще порядка 0.1.

В учебнике [1] определяются и изучаются асимптотически байесовские и асимптотически минимаксные оценки. В частности, оказывается, что при некоторых условиях регулярности оценки максимального правдоподобия являются асимптотически байесовскими (для любой априорной плотности q) и асимптотически минимаксными. Тем самым, в этих условиях все рассмотренные подходы к оптимальности "асимптотически совпадают".

2.11 Приближенное решение уравнения правдоподобия

Мы опишем сейчас практически приемлемую процедуру численного решения уравнения правдоподобия

$$\frac{dl(\theta)}{d\theta} = 0$$

(см. параграф 3). Обозначим для краткости через $f(\theta)$ левую часть этого уравнения и предположим, чтоf дифференцируема. Выберем некоторое начальное приближение t_0 к корню нашего уравнения (выбор t_0 обсуждается ниже) и линеаризуем уравнение в окрестности точки t_0, т.е. запишем

$$f(\theta) \approx f(t_0) + (\theta - t_0)f'(t_0).$$

Не следует забывать, что как корень $\hat{\theta}_{ML}$, так и последовательные приближения $\{t_k\}$ к нему, представляют собой случайные величины — функции от выборки. Это обстоятельство несколько изменит стандартную процедуру линеаризации.

Изменение (мы не пытаемся его мотивировать) состоит в том, что случайная величина $f'(t_0)$ заменяется "близкой"в некотором смысле к ней неслучайной величиной $-I(t_0)$ (мы знаем из параграфа 5, что при некоторых условиях регулярности $\mathbf{E}_\theta f'(\theta) = -I(\theta)$). В результате получаем "приближенное"равенство

$$f(\theta) \approx f(t_0) - I(t_0)(\theta - t_0)$$

Приравнивая правую часть к нулю и решая получающееся уравнение, находим корень

$$t_1 = t_0 + I^{-1}(t_0)f(t_0).$$

Затем аналогичным образом строим последовательные приближения

$$t_{k+1} = t_k + I^{-1}(t_k)f(t_k). \tag{2.11}$$

Полученный рецепт можно теперь "обосновать"следующим образом. Предположим, что последовательность $\{t_k\}$ сходится. Тогда из (2.11) следует, что $t_\infty = \lim_{k\to\infty} t_k$ удовлетворяет уравнению правдоподобия. На практике (см. ниже пример) поступают так. В качестве t_0 берется состоятельная оценка параметра θ. Оценки $t_1, t_2, \ldots$ трактуются как ее улучшения. Часто оказывается, что уже t_1 или t_2 асимптотически эффективна.

В качестве примера рассмотрим оценивание параметра сдвига распределения Коши с плотностью

$$p(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}.$$

Легко проверить, что уравнение правдоподобия здесь оказывается алгебраическим уравнением степени $2N-1$, где N — объем выборки. Решать это уравнение аналитически невозможно. В то же время последовательные приближения (2.11) строить легко. Некоторую трудность представляет лишь выбор начального приближения t_0 — распределение Коши не имеет математического ожидания, а $\bar{X}$, естественная оценка центра распределения, несостоятельна. Можно предположить, что связана эта несостоятельность со слишком большими весами крайних порядковых статистик — минимума, максимума и близких к ним по номеру в вариационном ряде. Уменьшая эти веса, мы, видимо, должны получить более подходящие линейные комбинации наблюдений. Самой естественной оценкой такого вида является эмпирическая медиана med. По определению для нечетного N она совпадает с центральной порядковой статистикой, а для четного N — с полусуммой двух центральных порядковых статистик. Можно доказать, что med — состоятельная оценка параметра θ. Кроме того, она асимптотически нормальна с коэффициентом разброса $\pi^2/4$. Для улучшения ее сделаем *первое приближение*. Фишеровская информация $I(\theta)$ для параметра сдвига постоянна — не зависит от θ. Вычисления, которые мы не приводим, показывают, что $I = N/2$. По формуле (2.11) получаем

$$t_1 = \mathrm{med} + \frac{4}{N}\sum_{i=1}^{N}\frac{\mathrm{med} - X_i}{1+(X_i-\mathrm{med})^2}.$$

Можно проверить, что t_1 асимптотически нормальна с коэффициентом разброса 2, т.е. асимптотически эффективна (так же, как и оценка максимального правдоподобия).

2.12 Уменьшение смещения методом “складного ножа”

Здесь мы рассмотрим один практический прием, часто позволяющий уменьшить смещение оценки, не ухудшая ее асимптотические

свойства. Этот прием называется методом "складного ножа"(jack-knife) первого порядка. Можно доказать, что применение его к оценке максимального правдоподобия при широких предположениях не нарушает асимптотическую эффективность.

Итак, пусть $\vec{X} = (X_1, \ldots, X_N)^T$ — исходная выборка, $\hat{\theta}(\vec{X})$ — оценка параметра θ. Будем обозначать $\vec{X}_{(i)}$ выборку, из которой удалено i-е наблюдение. Положим

$$\tilde{\theta}(\vec{X}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}(\vec{X}_{(i)})$$

и

$$\theta^*(\vec{X}) = N\hat{\theta}(\vec{X}) - (N-1)\tilde{\theta}(\vec{X}).$$

Оценка θ^* и называется оценкой "складного ножа". Предположим, что для смещения исходной оценки $\hat{\theta}$ имеется разложение вида

$$\mathbf{E}\hat{\theta} - \theta = \frac{\alpha_1}{N} + \frac{\alpha_2}{N^2} + \cdots .$$

Тогда

$$\mathbf{E}\tilde{\theta} - \theta = \frac{\alpha_1}{N-1} + \frac{\alpha_2}{(N-1)^2} + \cdots$$

и

$$\mathbf{E}\theta^* - \theta = \left(\alpha_1 + \frac{\alpha_2}{N} + \cdots\right) - \left(\alpha_1 + \frac{\alpha_2}{N-1} + \cdots\right) = \frac{\alpha_2}{N(N-1)} + \cdots .$$

Поэтому смещение порядка $O(1/N)$ исходной оценки $\hat{\theta}$ превращается в смещение порядка $O(1/N^2)$ новой оценки θ^*. В качестве иллюстрации рассмотрим эмпирическую дисперсию S^2. Не очень сложные вычисления показывают, что метод "складного ножа"преобразует ее в исправленную эмпирическую дисперсию $S^2_{\text{испр.}}$. Действительно,

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}^2 = \frac{s_2}{N} - \frac{s_1^2}{N^2},$$

где для удобства мы обозначили $s_1 = X_1 + \cdots + X_N$, $s_2 = X_1^2 + \cdots + X_N^2$.

Далее,

$$S^2(\vec{X}_{(i)}) = \frac{s_2 - X_i^2}{N-1} - \frac{(s_1 - X_i)^2}{(N-1)^2}$$
$$= \frac{s_2}{N-1} - \frac{s_1^2}{(N-1)^2} + 2X_i\frac{s_1}{(N-1)^2} - \frac{NX_i^2}{(N-1)^2}.$$

Поэтому

$$\tilde{S}^2 = \frac{1}{N}\sum_{i=1}^{N} S^2(\vec{X}_{(i)})$$
$$= \frac{s_2}{N-1} - \frac{s_1^2}{(N-1)^2} + 2\frac{s_1}{N}\cdot\frac{s_1}{(N-1)^2} - \frac{s_2}{(N-1)^2}$$

и

$$NS^2 - (N-1)\tilde{S}^2$$
$$= s_2 - \frac{s_2^2}{N} - s_2 + \frac{s_1^2}{N-1} - 2\frac{s_1^2}{N(N-1)} + \frac{s_2}{N-1}$$
$$= \frac{s_2}{N-1} - \frac{s_1^2}{N(N-1)} = \frac{N}{N-1}\left(\frac{s_2}{N} - \frac{s_1^2}{N^2}\right)$$
$$= \frac{N}{N-1}S^2 = S^2_{\text{испр.}}.$$

В справочнике [3] можно найти другие подобные приемы, а также ссылки на соответствующие оригинальные работы.

Глава 3

Доверительные интервалы

Эту небольшую главу можно рассматривать как связку, обеспечивающую переход от задач оценивания к задачам проверки гипотез. Кроме того, излагается очень важный результат, относящийся к нормально распределенным величинам — так называемая лемма Фишера.

3.1 Основные определения и асимптотическая теория доверительных интервалов

Предположим, что θ — некоторый функционал от теоретического распределения, ε — малое положительное число. Дополнительное число $1-\varepsilon$ будем называть доверительной вероятностью (и действительно, это число будет характеризовать надежность предполагаемого статистического вывода или уровень доверия к нему).

Доверительным интервалом (confidence interval) для θ, отвечающим доверительной вероятности $1-\varepsilon$, называется интервал $\langle \underline{\theta}, \overline{\theta} \rangle$ со случайными концами, обладающий двумя свойствами:

1. его концы $\underline{\theta}$ и $\overline{\theta}$ являются статистиками, т.е. функциями от выборки $\vec{X}$ (и ни от чего больше);

2. интервал $\langle \underline{\theta}, \overline{\theta} \rangle$ "ловит"значение функционала θ с вероятностью, не меньшей $1-\varepsilon$, т.е.

$$\mathsf{P}_\theta(\langle \underline{\theta}, \overline{\theta} \rangle \ni \theta) \geq 1-\varepsilon \tag{3.1}$$

(каково бы ни было априори допустимое теоретическое распределение).

Тип интервала (открытый, замкнутый, полуоткрытый) в большинстве задач принципиального значения не имеет. При необходимости его следует уточнить.

Мы использовали зеркально отраженный знак принадлежности, чтобы подчеркнуть важное обстоятельство: в записи $\langle\cdots\rangle \ni \theta$ меняющимся является интервал, в то время как θ — фиксированное (неизвестное статистику) число.

В многомерном случае вместо интервалов используются доверительные множества, определяющиеся аналогичным образом.

Как правило, построить доверительное множество (интервал) удается только в случае, когда совокупность априори допустимых теоретических распределений не слишком обширна, например, когда она описывается конечным набором параметров. В более общей ситуации может оказаться возможным построение асимптотического доверительного интервала (множества). Асимптотические доверительные интервалы определяются сходным образом: вместо выполнения неравенства (3.1) следует потребовать выполнение предельного соотношения

$$\underline{\lim}_{N\to\infty}\mathsf{P}_\theta(\langle\underline{\theta},\overline{\theta}\rangle \ni \theta) \geq 1-\varepsilon. \tag{3.2}$$

Доверительные множества указывают погрешность, с которой можно по выборке найти приближенное значение функционала θ при заданном уровне доверия $1-\varepsilon$ (к этой погрешности).

Аналогично основным характеристикам точечных оценок (состоятельность, несмещенность, эффективность) можно ввести характеристики доверительных интервалов или множеств. Мы ограничимся интервалами.

Будем говорить, что доверительный интервал **состоятелен**, если оба его конца, $\underline{\theta}$ и $\overline{\theta}$, сходятся по вероятности к оцениваемому функционалу (равносильная формулировка: $\overline{\theta}-\underline{\theta}\to 0$ по вероятности).

Назовем доверительный интервал **несмещенным**, если

$$\mathbf{E}_\theta[\underline{\theta}+\overline{\theta}] = 2\theta$$

(т.е. если его центр — несмещенная точечная оценка).

Эффективность доверительных интервалов естественно характеризовать величиной

$$\mathbf{E}_\theta(\overline{\theta}-\underline{\theta})^2.$$

Так же, как и в случае точечных оценок, доверительные интервалы могут оказаться в этом смысле несравнимыми.

Поскольку задача построения доверительных интервалов значительно сложнее задачи построения точечных оценок, свойствам оптимальности на практике редко уделяется большое внимание — чаще довольствуются интервалами, которые удается построить.

Перейдем теперь к описанию того, как можно строить доверительные интервалы в конкретных ситуациях. Прежде всего отметим, что уже в самом определении идет речь о событии нетривиальной вероятности (т.е. не равной ни 0, ни 1). Подсчитывать ее следует (см. определение) по теоретическому распределению, что довольно проблематично — в каждой задаче фигурирует целое семейство априори допустимых распределений. Рассмотрим простой учебный (малореалистичный с точки зрения практики) пример — нормальное распределение $\mathbf{N}(a,1)$. Будем строить доверительный интервал для параметра a, отталкиваясь от известной нам точечной оценки $\hat{a} = \bar{X}$. Как следует из параграфа 1.6, $\bar{X} \in \mathbf{N}(a, 1/N)$. Стандартизуем эту величину, написав

$$\frac{\bar{X} - a}{\sqrt{1/N}} = \sqrt{N}(\bar{X} - a) \in \mathbf{N}(0,1).$$

Стандартное нормальное распределение является одним из ходовых шаблонных распределений — для него много десятилетий назад были составлены подробные таблицы. Воспользовавшись, например, правилом "пяти процентов", мы получим, что

$$\mathsf{P}_a(|\sqrt{N}(\bar{X} - a)| < 1.96) \approx 0.95$$

(приблизительное равенство, поскольку мы округляем табличное значение до сотых). Решая неравенство, фигурирующее под знаком вероятности, мы получаем равносильное соотношение

$$\mathsf{P}\left(\left\langle \bar{X} - \frac{1.96}{\sqrt{N}}, \bar{X} + \frac{1.96}{\sqrt{N}} \right\rangle \ni a\right) \approx 0.95,$$

так что $\langle \bar{X} - \frac{1.96}{\sqrt{N}}, \bar{X} + \frac{1.96}{\sqrt{N}} \rangle$ — доверительный интервал, отвечающий доверительной вероятности 0.95. Подчеркнем, что к успеху в этом построении нас привело шаблонное распределение $\mathbf{N}(0,1)$. Дальше в этой главе мы познакомимся еще с несколькими подобными примерами, уже более полезными с точки зрения практического использования. Тем не менее, стоит сразу подчеркнуть, что более распространенным является появление шаблона в качестве предельного распределения. При этом строится только асимптотический доверительный интервал.

Проиллюстрируем построение асимптотических доверительных интервалов двумя примерами, а затем сформулируем и некоторый общий подход.

Пример 1. Доверительный интервал (асимптотический) для вероятности успеха.

Этот пример уже вкратце обсуждался в параграфе 1.4. Асимптотический шаблон дает нам предельная теорема Муавра-Лапласа:

$$\frac{N\bar{X} - Np}{\sqrt{Np(1-p)}} \approx \mathbf{N}(0,1).$$

Шаблоном снова является стандартное нормальное распределение. Выбирая (по доверительной вероятности $1-\varepsilon$) $z = z_\varepsilon$ из таблицы нормального распределения, получаем

$$\mathsf{P}\left(\left|\frac{N\bar{X} - Np}{\sqrt{Np(1-p)}}\right| < z\right) \approx 1 - \varepsilon.$$

Остается решить это неравенство относительно p:

$$\left|\frac{N\bar{X} - Np}{\sqrt{Np(1-p)}}\right| < z$$
$$\iff (1 + z^2/N)p^2 - (2\bar{X} + z^2/N)p + \bar{X}^2 < 0$$
$$\iff p_- < p < p_+,$$

где

$$p_\pm = \frac{2\bar{X} + \frac{z^2}{N} \pm \sqrt{4\bar{X}^2 + 4\frac{z^2}{N}\bar{X} + \frac{z^4}{N^2} - 4\bar{X}^4(1 + \frac{z^2}{N})}}{2(1 + \frac{z^2}{N})}$$
$$= \frac{\bar{X} + \frac{z^2}{N} \pm \sqrt{\frac{z^2}{N}\bar{X}(1-\bar{X}) + \frac{z^4}{4N^2}}}{1 + \frac{z^2}{N}}.$$

Оставляя в последнем выражении главные члены разложения по обратным степеням N, получим

$$p_\pm = \bar{X} \pm \frac{z}{\sqrt{N}}\sqrt{\bar{X}(1-\bar{X})} + O(1/N).$$

Оставлять следующие члены разложения, обратно пропорциональные N, вряд ли целесообразно, т.к. они пренебрежимо малы по сравнению с погрешностью, возникающей от нормальной аппроксимации, т.е. от применения теоремы Муавра-Лапласа.

Таким образом, асимптотический доверительный интервал для p имеет вид

$$\left\langle \bar{X} - \frac{z}{\sqrt{N}}\sqrt{\bar{X}(1-\bar{X})}, \bar{X} + \frac{z}{\sqrt{N}}\sqrt{\bar{X}(1-\bar{X})} \right\rangle.$$

Пример 3. Доверительный интервал (асимптотический) для параметра a нормального распределения $\mathbf{N}(a, \sigma^2)$ с неизвестным σ.

Здесь снова можно воспользоваться стандартным нормальным распределением $\mathbf{N}(0,1)$ в качестве шаблона:

$$\bar{X} \in \mathbf{N}(a, \sigma^2/N), \quad \frac{\bar{X} - a}{\sqrt{\frac{\sigma^2}{N}}} \in \mathbf{N}(0,1).$$

Теперь остается избавиться от мешающего параметра σ^2. Для этого заменим σ^2 на точечную оценку $\hat{\sigma}^2 = S^2_{\text{испр.}}$ и получим

$$\sqrt{N}\frac{\bar{X} - a}{\hat{\sigma}} \approx\in \mathbf{N}(0,1).$$

Отсюда получаем асимптотический доверительный интервал

$$\left\langle \bar{X} - z\frac{\hat{\sigma}}{\sqrt{N}}, \bar{X} + z\frac{\hat{\sigma}}{\sqrt{N}} \right\rangle,$$

где $z = z_\varepsilon$, как и в предыдущем примере, находится из таблиц стандартного нормального распределения. В параграфе 3.3 мы построим точный доверительный интервал для этого примера и сравним его с только что найденным асимптотическим.

Обобщая рассуждения этих двух примеров, можно сказать, что для получения асимптотического доверительного интервала следует исходить из функции от выборки и параметров, имеющей асимптотически шаблонное распределение. Из таблиц этого распределения определяется некоторое множество, из которого и строится доверительный интервал (в рассмотренных примерах это построение сводилось к решению неравенств). Мешающие параметры, если таковые имеются, заменяются состоятельными точечными оценками.

В параграфе 2.8 приведен результат об асимптотической нормальности оценок максимального правдоподобия, которым можно воспользоваться в нашем построении:

$$\sqrt{Ni(\theta)}(\hat{\theta}_{ML} - \theta) \approx\in \mathbf{N}(0,1).$$

Решая неравенство вида

$$|\sqrt{Ni(\theta)}(\hat{\theta}_{ML} - \theta)| < z$$

относительно θ, мы получим доверительное множество. Скорее всего, оно окажется интервалом.

В разобранных примерах доверительные интервалы оказываются состоятельными и, можно подозревать, асимптотически эффективными. Мы не останавливаемся на этом более детально.

3.2 Лемма Фишера

Для нормально распределенных выборок имеется точная (а не асимптотическая) теория доверительных интервалов, даже при наличии мешающего параметра. В этом параграфе излагается базис этой теории.

В формулировке приводимой ниже леммы Фишера участвует распределение хи-квадрат, уже обсуждавшееся в параграфах 1.5 и 1.6. Напомним, что сумма квадратов n независимых величин, распределенных по стандартному нормальному закону $\mathbf{N}(0,1)$, имеет распределение χ^2_n, при этом индекс n называется числом степеней свободы.

Теорема (лемма Фишера). Пусть $\vec{X}$ — выборка, имеющая нормальное распределение $\mathbf{N}(a, \sigma^2)$. Тогда

1. $\bar{X}$ и S^2 независимы;
2. $\frac{NS^2}{\sigma^2} \in \chi^2_{N-1}$.

Доказательство леммы Фишера чрезвычайно важно как образец, поскольку подобные рассуждения неоднократно будут появляться далее, в эконометрических главах, и мы будем ссылаться на аналогию с простейшим случаем — только что сформулированной теоремой.

Перейдем непосредственно к доказательству. Прежде всего заметим, что достаточно доказать теорему в случае $a = 0$, $\sigma = 1$. Действительно,

рассмотрим преобразованную выборку

$$X'_i = \frac{X_i - a}{\sigma} \in \mathbf{N}(0,1).$$

Ясно, что

$$\overline{X'} = \frac{\bar{X}}{\sigma}$$

и

$$S'^2 = \frac{S^2}{\sigma^2}.$$

Поэтому достаточно доказать, что

$1)'$ независимы $\overline{X'}$ и S'^2;

$2)'$ $NS'^2 \in \chi^2_{N-1}$.

Эти два утверждения как раз и составляют упомянутый частный случай.

Итак, далее считаем $X_1, \ldots, X_N \in \mathbf{N}(0,1)$. Величина NS^2, разумеется, является суммой квадратов нормально распределенных величин $X_i - \bar{X}$, но величины эти зависимые:

$$\sum_{i=1}^{N}(X_i - \bar{X}) = 0,$$

а к тому же и с неправильной ($\neq 1$) дисперсией. Зависимость, как мы увидим, уменьшает на единицу число степеней свободы (подобные соображения на эвристическом уровне иногда очень полезны, см. главу 6), а дисперсия "подправляется"сама собой.

Основная идея доказательства — воспользоваться инвариантностью распределения выборки при вращениях. Опишем эту инвариантность более точно. Для начала запишем плотность совместного распределения выборки в виде

$$p(\vec{x}) = (2\pi)^{-N/2} \exp\{-\frac{1}{2}\sum_{i=1}^{N} x_i^2\} = (2\pi)^{-N/2} \exp\{-\frac{1}{2}\vec{x}'\vec{x}\}.$$

Напомним, что $\vec{x}$ понимается как вектор-столбец, а штрих — знак транспонирования.

Вращениями (этот геометрический термин не обязателен, хотя и очень нагляден) мы называем ортогональные линейные преобразования вида

$$\vec{y} = A\vec{x}.$$

Матрица A, как известно, называется ортогональной, если $A^{-1} = A'$ (обратная совпадает с транспонированной). Для таких матриц $\det A = \pm 1$. При сделанном преобразовании

$$\vec{y}'\vec{y} = (A\vec{x})'A\vec{x} = \vec{x}'A'A\vec{x} = \vec{x}'\vec{x}$$

(сумма квадратов координат — квадрат расстояния до начала координат; мы доказали, что он сохраняется, это оправдывает термин "вращение"). Из доказанного соотношения вытекает, что $p(\vec{x}) = p(\vec{y})$ — инвариантность плотности при вращениях.

Рассмотрим теперь ортогональное преобразование выборки:

$$\vec{Y} = A\vec{X}$$

и докажем, что $p(\vec{y})$ — плотность распределения случайного вектора $\vec{Y}$. Для этого возьмем (измеримое) множество $B \subset \mathbb{R}^N$ и запишем

$$\mathsf{P}(\vec{Y} \in B) = \mathsf{P}(A\vec{X} \in B) = \mathsf{P}(\vec{X} \in A^{-1}B) = \int_{A^{-1}B} p(\vec{x})d\vec{x}.$$

Сделаем теперь замену переменных $\vec{y} = A\vec{x}$ и воспользуемся инвариантностью плотности при этом преобразовании ($p(\vec{x}) = p(\vec{y})$), а также инвариантностью элемента объема: $d\vec{y} = |\det A|d\vec{x} = d\vec{x}$. После указанной замены получим

$$\mathsf{P}(\vec{Y} \in B) = \int_B p(\vec{y})d\vec{y},$$

так что случайный вектор $\vec{Y}$ имеет ту же плотность p, что и исходный вектор $\vec{X}$. Другими словами, величины $Y_1, \ldots, Y_N$ независимы и распределены по стандартному нормальному закону $\mathbf{N}(0, 1)$.

Выберем теперь ортогональную матрицу A специальным образом — чтобы ее первая строка состояла из одинаковых элементов $\frac{1}{\sqrt{N}}$. Докажем, что такая матрица существует. Для этого заметим, что условие ортогональности $AA' = \mathbf{1}$ означает, что строки матрицы A, как векторы, ортогональны и нормированы (скалярное произведение i-й строки A и j-го столбца A' есть δ_{ij}, т.е. равно 1 при $i = j$ и 0 в остальных случаях). Иначе можно сказать, что строки ортогональной матрицы образуют ортогональный нормированный базис пространства векторов-строк размерности N.

Отметим теперь, что вектор (у нас — вектор-строка) из элементов $\frac{1}{\sqrt{N}}$ нормирован:

$$\sum_{j=1}^{N}\left(\frac{1}{\sqrt{N}}\right)^2 = 1.$$

Дополним этот вектор-строку произвольным образом до ортогонального нормированного базиса (это, очевидно, возможно) и составим матрицу A из полученных таким образом строк. Получим искомую матрицу.

Действуя построенной матрицей A на выборку $\vec{X}$, найдем

$$Y_1 = \sum_{j=1}^{N}\frac{1}{\sqrt{N}}X_j = \sqrt{N}\bar{X}.$$

Кроме того,

$$Y_1^2 + \cdots + Y_N^2 = \vec{Y}'\vec{Y} = \vec{X}'\vec{X} = X_1^2 + \cdots + X_N^2,$$

откуда

$$\begin{aligned} Y_2^2 + \cdots + Y_N^2 &= X_1^2 + \cdots + X_N^2 - (\sqrt{N}\bar{X})^2 \\ &= N\left[\frac{X_1^2 + \cdots + X_N^2}{N} - \bar{X}^2\right] = NS^2. \end{aligned}$$

Остается сделать необходимые выводы, опираясь на установленные ранее свойства величин Y_i — независимость и $\mathbf{N}(0,1)$-распределенность.

Во-первых, из формул

$$\bar{X} = \frac{Y_1}{\sqrt{N}}, \quad NS^2 = Y_2^2 + \cdots + Y_N^2$$

с очевидностью следует независимость $\bar{X}$ и S^2. Во-вторых, из представления NS^2 с такой же очевидностью следует, что эта величина имеет распределение χ^2_{N-1}.

Лемма Фишера доказана.

3.3 Точные доверительные интервалы для параметров нормального распределения

Рассмотрим сначала математическое ожидание a (при неизвестном σ) — пример, уже обсуждавшийся на асимптотическом уровне в параграфе

1. Нам потребуется новый шаблон — распределение Стьюдента $\mathbf{t}_n$. Символически оно определяется формулой

$$\mathbf{t}_n = \sqrt{n}\frac{\mathbf{N}(0,1)}{\sqrt{\chi^2_n}}.$$

Понимать ее следует так. Подставляем в знаменатель вместо символа χ^2_n случайную величину, имеющую это распределение χ^2_n, а в числитель вместо символа нормального распределения $\mathbf{N}(0,1)$ — случайную величину, имеющую это нормальное распределение и **не зависящую** от величины, подставленной в знаменатель. Тогда вся дробь, как случайная величина, будет иметь распределение $\mathbf{t}_n$ — распределение Стьюдента с n степенями свободы.

Лемма Фишера позволяет утверждать, что дробь

$$\sqrt{N-1}\frac{\frac{\bar{X}-a}{\sqrt{\sigma^2/N}}}{\sqrt{\frac{NS^2}{\sigma^2}}} = \sqrt{N-1}\frac{\bar{X}-a}{\sqrt{S^2}} = \sqrt{N}\frac{\bar{X}-a}{\sqrt{S^2_{\text{испр.}}}}$$

имеет распределение Стьюдента с $n = N-1$ степенями свободы. Главное достоинство стьюдентовской дроби — масштабная инвариантность — мешающий параметр σ благополучно сократился. Построим с помощью этой дроби доверительный интервал. Для этого по доверительной вероятности $1-\varepsilon$ найдем из таблиц распределения Стьюдента значение $z = z_\varepsilon$ так, чтобы

$$\mathsf{P}(|\mathbf{t}_{N-1}| < z) = 1-\varepsilon. \tag{3.3}$$

Решая теперь неравенство

$$\left|\sqrt{N}\frac{\bar{X}-a}{\sqrt{S^2_{\text{испр.}}}}\right| < z$$

относительно a, получим искомый интервал

$$\left\langle \bar{X} - z\frac{S_{\text{испр.}}}{\sqrt{N}}, \bar{X} + z\frac{S_{\text{испр.}}}{\sqrt{N}} \right\rangle$$

(здесь $S_{\text{испр.}} = \sqrt{S^2_{\text{испр.}}}$), который очень похож на асимптотический, полученный в параграфе 1. Отличие лишь в табличном значении, которое сейчас определяется по другой таблице. В асимптотическом смысле оба интервала совпадают, т.к. при $n \to \infty$ распределение

Стьюдента $\mathbf{t}_n$ слабо сходится к нормальному $\mathbf{N}(0,1)$. Это обстоятельство можно объяснить так. Распределение $\mathbf{t}_n$ — это распределение дроби вида

$$\frac{X_0}{\sqrt{\frac{X_1^2+\cdots+X_n^2}{n}}},$$

где $X_0, X_1, \ldots, X_n$ — независимые $\mathbf{N}(0,1)$-распределенные величины. Поскольку среднее арифметическое

$$\frac{X_1^2+\cdots+X_n^2}{n}$$

согласно закону больших чисел сходится к 1, общему значению математических ожиданий квадратов, то исходная дробь сходится к $X_0 \in \mathbf{N}(0,1)$. Отсюда несложно вывести и слабую сходимость распределений, но мы на этом не останавливаемся.

Обсудим теперь вопрос об оптимальности полученного доверительного интервала в следующем, довольно узком, смысле. Вместо выбора z из (3.3) можно было бы более общим образом взять z_1 и z_2 из соотношения

$$\mathsf{P}(z_1 < \mathbf{t}_{N-1} < z_2) = 1-\varepsilon. \tag{3.4}$$

Докажем, что интервал $\langle -z_\varepsilon, z_\varepsilon\rangle$, использовавшийся ранее, самый короткий из всех интервалов вида $\langle z_1, z_2\rangle$. Тогда и построенный по нему доверительный интервал, как легко видеть, будет кратчайшим из всех подобных интервалов.

Для доказательства воспользуемся тем, что плотность $\mathbf{t}_n(x)$ распределения Стьюдента — четная функция (это почти очевидно), монотонно убывающая на положительной полуоси (это свойство мы доказывать не будем, оно следует из явной формулы, которую можно найти во многих источниках, например, [1]). Обратимся к соотношению (3.4) и заметим, что вероятность представляется геометрически как площадь под графиком плотности.

Предположим для определенности, что $-z_\varepsilon < z_1 < 0 < z_\varepsilon$ (остальные варианты рассматриваются аналогично). Тогда, очевидно, $z_\varepsilon < z_2$. Площади под графиком плотности на промежутках $\langle -z_\varepsilon, z_1\rangle$ и $\langle z_\varepsilon, z_2\rangle$ должны совпадать. Однако на первом из них минимальное значение плотности есть $t_n(-z_\varepsilon)$, а на втором — максимальное значение плотности есть $t_n(z_\varepsilon) = t_n(-z_\varepsilon)$. Поэтому на всем первом промежутке $\langle -z_\varepsilon, z_1\rangle$ плотность t_n больше, чем на втором промежутке $\langle z_\varepsilon, z_2\rangle$. Из равенства

площадей вытекает, что длина первого промежутка меньше, чем второго. Таким образом, при переходе от $\langle -z_\varepsilon, z_\varepsilon \rangle$ к $\langle z_1, z_2 \rangle$ вычитается более короткий промежуток, чем добавляется (см. рис).

Перейдем теперь к построению доверительного интервала для σ при неизвестном a. Опять воспользуемся леммой Фишера:

$$\frac{NS^2}{\sigma^2} \in \chi^2_{N-1}.$$

В этом соотношении мешающий параметр a уже отсутствует. Поэтому берем χ^2_{N-1} в качестве шаблонного распределения, выбираем $\langle z_1, z_2 \rangle$ из соотношения

$$\mathsf{P}(z_1 < \chi^2_{N-1} < z_2) = 1 - \varepsilon \tag{3.5}$$

и, решая неравенство

$$z_1 < \frac{NS^2}{\sigma^2} < z_2$$

относительно σ^2, находим доверительный интервал

$$\left\langle \frac{NS^2}{z_2}, \frac{NS^2}{z_1} \right\rangle.$$

Плотность распределения χ^2_n, хотя и не симметрична, при $n \geq 3$ одновершинна — имеет единственный максимум — и монотонна с каждой стороны от него (см. параграф 1.5, где имеется явная формула плотности гамма-распределения, частным случаем которого является хи-квадрат). Поэтому соображения, аналогичные изложенным выше применительно к распределению Стьюдента, сразу же говорят нам, что кратчайший доверительный интервал получается, если значения плотности в точках z_1 и z_2 совпадают. Подбирать такие z_1 и z_2 с использованием вычислительной техники несложно. В литературе докомпьютерного

времени обычно приводятся упрощенные рекомендации, позволяющие обойтись двукратным заглядыванием в таблицу. Именно, предлагается (3.5) заменить парой соотношений

$$\mathsf{P}(\chi^2_{N-1} < z_1) = \mathsf{P}(\chi^2_{N-1} > z_2) = \frac{\varepsilon}{2}.$$

При не слишком малых N такой выбор z_1 и z_2 почти оптимален.

В заключение отметим, что при известном a можно несколько улучшить рецепт построения доверительного интервала. В качестве оценки дисперсии σ^2 естественно в этом случае брать

$$S^2_{\text{модиф.}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - a)^2.$$

Очевидно, что

$$\frac{NS^2_{\text{модиф.}}}{\sigma^2} \in \chi^2_N.$$

Доверительный интервал, основанный на этом соотношении, представляется явно более предпочтительным, т.к. формула явно учитывает информацию о математическом ожидании. Нетрудно сообразить, что увеличение на единицу числа степеней свободы укорачивает этот интервал по сравнению с интервалом, основанном на (3.5). Впрочем, случай известного a представляет, главным образом, академический, а не практический, интерес.

3.4 Двумерные доверительные множества для параметров нормального распределения

Продолжая обсуждение нормально распределенной выборки, рассмотрим построение доверительной области для пары параметров (a, σ^2). Для простоты мы ограничимся асимптотической доверительной областью (и даже в этом случае опустим громоздкие выкладки).

По лемме Фишера величины $\bar{X} \in \mathbf{N}(a, \sigma^2/N)$ и $NS^2/\sigma^2 \in \chi^2_{N-1}$ независимы. Аппроксимируем распределение хи-квадрат нормальным (см. параграф 1.5). Асимптотически при $N \to \infty$ можно написать

$$\frac{\frac{NS^2}{\sigma^2} - N}{\sqrt{2N}} \approx\in \mathbf{N}(0, 1).$$

Таким образом, случайные величины

$$\sqrt{N}\frac{\bar{X}-a}{\sigma}, \quad \sqrt{\frac{N}{2}}\left(\frac{S^2}{\sigma^2}-1\right)$$

независимы и имеют асимптотически распределение $\mathbf{N}(0,1)$. Для двумерного нормального распределения с плотностью

$$p(x,y)=\varphi(x)\varphi(y)=\frac{1}{2\pi}\exp\{-\frac{1}{2}(x^2+y^2)\}$$

по доверительной вероятности $1-\varepsilon$ легко найти круг

$$\{(x,y): x^2+y^2<c^2\},$$

имеющий именно эту вероятность. Можно сосчитать, что $c^2=2\ln\frac{1}{\varepsilon}$. Остается заменить x и y нашими случайными величинами и получить неявное описание асимптотического доверительного множества

$$\left[\sqrt{N}\frac{\bar{X}-a}{\sigma}\right]^2+\left[\sqrt{\frac{N}{2}}\left(\frac{S^2}{\sigma^2}-1\right)\right]^2<c^2. \tag{3.6}$$

Положим $t=\sqrt{N}\frac{\bar{X}-a}{S}$, $q=\frac{S^2}{\sigma^2}$. Тогда неравенство (3.6) можно переписать в виде

$$t^2q+\frac{N}{2}(q-1)^2<c^2$$

или

$$q^2-2(1-\frac{t^2}{N})q+1<\frac{2c^2}{N}.$$

Решая его относительно q, получаем

$$q_-<q<q_+,$$

где

$$q_\pm=1-\frac{t^2}{N}\pm\sqrt{\frac{2(c^2-t^2)}{N}+\frac{t^4}{N^2}}.$$

Несложные, но скучные выкладки показывают, что (с точностью до малых более высокого порядка)

$$q_\pm\approx 1\pm\frac{\sqrt{2(c^2-t^2)}}{\sqrt{N}} \text{ при } t^2<c^2.$$

Неравенство $t^2<c^2$ дает нам интервал

$$\bar{X}-\frac{cS}{\sqrt{N}}<a<\bar{X}+\frac{cS}{\sqrt{N}} \tag{3.7}$$

для тех a, для которых $q_\pm$ вещественны (и положительны). Для этих a интервал $\langle q_-, q_+\rangle$ изменения q записывается в виде

$$1 - \frac{\sqrt{2(c^2 - t^2)}}{\sqrt{N}} < q < 1 + \frac{\sqrt{2(c^2 - t^2)}}{\sqrt{N}}.$$

Для σ^2 при этом получаем (асимптотически)

$$S^2\left(1 - \frac{d(a)}{\sqrt{N}}\right) < \sigma^2 < S^2\left(1 + \frac{d(a)}{\sqrt{N}}\right),$$

где $d(a) = \sqrt{2(c^2 - t^2)}$ зависит через t от a, изменяющегося в промежутке (3.7).

3.5 Доверительные интервалы и гипотезы о параметрах

Перейдем, наконец, к обещанной связи с проверкой гипотез. Имеются в виду гипотезы вида $\theta = \theta_0$, где θ_0 — некоторое заданное конкретное значение параметра. Проверять их с помощью доверительных интервалов очень просто. Если гипотетическое значение θ_0 **не** попадает в доверительный интервал, гипотезу следует отвергнуть, в противном случае, с обычными оговорками, принять. Природу этих оговорок в рассматриваемом примере очень легко понять. Доверительный интервал всего лишь показывает, что, приняв гипотезу, мы не вступаем в отчетливо видимое противоречие с эмпирическими данными. Однако, приняв альтернативное, но близкое, предположение о θ, мы также не вошли бы в такое противоречие. Отличить гипотезу $\theta = \theta_0$ от близких гипотез, тем самым, невозможно. Отвержение же гипотезы производится как раз на основе явного (хотя и не абсолютного) противоречия с эмпирическими данными.

Отметим, что традиционно при проверке гипотез задается малое положительное число $\varepsilon > 0$ — уровень значимости. Получить из него доверительную вероятность можно вычитанием из 1 ("переходом к противоположному событию"). Впрочем, А.А.Боровков предлагает уровнем значимости называть прямо $1 - \varepsilon$ (см. [1]). Поскольку такое словоупотребление расходится с принятым, мы, с некоторым сожалением и колебанием, не принимаем его предложение.

Изложенный рецепт привлекает своей общностью, однако не стоит забывать о том, что нам при этом не потребовалось даже уточнить,

как выглядит альтернативная гипотеза. Надо думать, в различных более узких задачах возможны и более оптимальные рецепты.

Приведем три простых примера не столь прямолинейного использования доверительных интервалов для проверки гипотез.

Критерий знаков.

Предположим, что у нас имеются две независимые между собой выборки одинакового объема N — $X_1, \ldots, X_N$ и $Y_1, \ldots, Y_N$. Основная гипотеза состоит в том, что эти две выборки одинаково распределены. Критерий знаков дает грубый способ, который иногда позволяет сразу же отвергнуть основную гипотезу. Правда, если это не удается, доводов в пользу ее почти не появляется.

Способ этот состоит в рассмотрении последовательности знаков: пишем $+$, если $X_i > Y_i$, пишем -, если $X_i < Y_i$, ничего не пишем, если $X_i = Y_i$. Получаем последовательность знаков длины $N' \leq N$, т.е. выборку из успехов и неудач. Заметим, что если основная гипотеза справедлива, то обе вероятности

$$p_+ = \mathsf{P}(X_i > Y_i | X_i \neq Y_i)$$

и

$$p_- = \mathsf{P}(X_i < Y_i | X_i \neq Y_i)$$

равны $1/2$. Если это значение $1/2$ не попадает в доверительный интервал для вероятности успеха p_+, гипотезу можно отвергнуть. В противном случае рекомендуется продолжить исследование более точными методами.

Сравнение двух независимых нормально распределенных выборок с одинаковыми дисперсиями.

Пусть $X_1, \ldots, X_N$ — независимые величины, имеющие нормальное распределение $\mathbf{N}(a, \sigma^2)$, $X'_1, \ldots, X'_{N'}$ — независимые между собой и с $X_1, \ldots, X_N$ величины, имеющие нормальное распределение $\mathbf{N}(a', \sigma^2)$. Основная гипотеза заключается в совпадении двух теоретических распределений, т.е. в совпадении средних: $a = a'$. Подчеркнем, что равенство дисперсий предполагается, хотя само значение σ^2 считается неизвестным. Для проверки гипотезы образуем стьюдентовскую дробь

вида

$$\sqrt{N+N'-2}\frac{\left(\frac{\bar{X}-a}{\sigma}-\frac{\overline{X'}-a'}{\sigma}\right)\left(\frac{1}{N}+\frac{1}{N'}\right)^{-1/2}}{\sqrt{\frac{NS^2}{\sigma^2}+\frac{N'S'^2}{\sigma^2}}}$$
$$=\sqrt{N+N'-2}\sqrt{\frac{NN'}{N+N'}}\frac{(\bar{X}-\overline{X'})-(a-a')}{\sqrt{NS^2+N'S'^2}},$$

имеющую распределение $\mathbf{t}_{N+N'-2}$ (проверьте!), и построим с ее помощью доверительный интервал для разности $a-a'$. Если гипотетическое значение 0 для нее не попадает в доверительный интервал, гипотезу можно отвергнуть (на соответствующем уровне значимости).

Сравнение дисперсий двух независимых нормально распределенных выборок.

Предположим, что $X_1,\ldots,X_N$ — независимые величины, имеющие нормальное распределение $\mathbf{N}(a,\sigma^2)$, а $X'_1,\ldots,X'_{N'}$ — независимые между собой и с $X_1,\ldots,X_N$ величины, имеющие нормальное распределение $\mathbf{N}(a',\sigma'^2)$. Основная гипотеза заключается в том, что $\sigma=\sigma'$ (проверка этой гипотезы при определенных условиях может составить первый этап перед проверкой совпадения средних). Для проверки воспользуемся еще одним шаблонным распределением, так называемым $\mathbf{F}$-распределением Фишера (оно будет использоваться и в последующих главах). По определению, случайная величина

$$\frac{n_2Z_1}{n_1Z_2}=\frac{Z_1/n_1}{Z_2/n_2},$$

где Z_1 и Z_2 независимы, $Z_1\in\chi^2_{n_1}$, $Z_2\in\chi^2_{n_2}$, имеет распределение $\mathbf{F}_{n_1,n_2}$. Оба индекса называются числами степеней свободы (числителя и знаменателя соответственно).

По лемме Фишера (она избавляет нас от мешающих параметров a и a')

$$\frac{S^2_{\text{испр.}}/\sigma^2}{S'^2_{\text{испр.}}/\sigma'^2}\in\mathbf{F}_{n_1,n_2}.$$

С помощью таблиц распределения Фишера можно теперь построить доверительный интервал для отношения дисперсий σ'^2/σ^2. Если гипотетическое значение 1 для этого отношения не попадает в доверительный интервал, основная гипотеза отвергается на выбранном уровне значимости.

Глава 4

Проверка статистических гипотез

В этой главе мы рассмотрим только общую (классическую) часть теории, оставляя для следующих, эконометрических, глав более сложные и специальные вопросы. Там их обсуждение будет более естественным.

4.1 Ошибки двух родов и уровень значимости

Начнем даже не с ошибок, а с напоминания простейших определений. Статистической гипотезой называется предположительное высказывание о неизвестном теоретическом (оно же генеральное) распределении вероятностей. Гипотеза называется **простой**, если этому высказыванию удовлетворяет единственное априори допустимое распределение, и **сложной** — в остальных случаях. Тем самым, совокупность всех априори допустимых мер разбивается на две взаимно дополнительные части: H_0 — распределения, удовлетворяющие выдвинутой гипотезе (она часто называется основной или нулевой), и H_1 — остальные априори допустимые распределения, которые автоматически формируют альтернативную гипотезу.

Как правило, основная гипотеза представляет собой формулировку некоторой идеализации, которая, сама по себе, конечно, исследователя устроила бы (например, определенной конкретностью, или другими свойствами), но которая вызывает известные сомнения (ср. с комментариями в параграфе 3.5). Соответственно этому формируется и отношение исследователя к возможным ошибкам в статистическом выводе. Ошибка первого рода — отвергнуть основную гипотезу, в то время как "на самом деле"она справедлива — заботит его в первую очередь, а потому для вероятности этой ошибки устанавливается жесткая верхняя граница, называемая уровнем значимости (significance

level). К обсуждению допускаются только критерии (тесты), дающие ошибку первого рода, удовлетворяющую этому требованию. Таких тестов, вообще говоря, бесконечно много, и сравнивать их можно уже по вероятности ошибки второго рода — принять основную гипотезу, в то время как на самом деле она ложна. Как именно сравнивать, будет обсуждаться дальше. Такая постановка задачи (с фиксированным уровнем значимости) нарушает первоначальное видимое равноправие основной и альтернативной гипотез, но обычно согласуется со здравым смыслом. В некоторых случаях альтернативная гипотеза вообще представляет собой чисто формальное ("голое") отрицание основной гипотезы, а тогда и рассуждать о вероятностях ошибки второго рода почти бессодержательно. Напротив, находить тесты с заданным уровнем значимости обычно удается.

Вопрос о том, как задается уровень значимости, выходит за рамки статистики — фактически этот уровень характеризует надежность ожидаемого вывода, а желаемая надежность как-то связана с предметной интерпретацией статистических данных. Образно говоря, надежность (или уровень значимости) устанавливается заказчиком статистического исследования. Эконометристу в некоторой степени сложнее — он сам часто является и заказчиком собственного исследования.

Итак, вероятность ошибки первого рода представляет собой функцию на множестве H_0, ограниченную сверху уровнем значимости ε, а вероятность ошибки второго рода — функцию на дополнительном множестве H_1, состоящем из остальных априори допустимых распределений. В параметрическом случае область Θ изменения параметра θ разбивается на взаимно дополнительные части Θ_0 и Θ_1, имеющие аналогичный смысл, а вероятности ошибок становятся функциями от параметра на этих множествах Θ_0 и Θ_1.

В этой главе мы будем предполагать, что θ однозначно определяет априори допустимое распределение — возможные "мешающие"параметры включены в обозначение θ.

Для того чтобы выражения типа "вероятность ошибки первого рода"стали до конца определенными, следует еще уточнить, что статистическим критерием или тестом называется отображение, переводящее выборку $\vec{X}$ в статистический вывод. В простейшем случае одномерных наблюдений выборка $\vec{X}$ — точка N-мерного пространства, а статистических выводов всего два — либо принять H_0, либо отвергнуть

(т.е. принять Y_1). Поэтому тест представляет собой отображение из $\mathbb{R}^N$ в двухточечное множество $\{H_0, H_1\}$. Обычно такое отображение задают критической областью — подмножеством $\mathbb{R}^N$, на котором оно (отображение) принимает значение H_1 (основная гипотеза отвергается). Мы будем обозначать критическую область через K. Фактически часто удобно отождествлять тест с его критической областью. Запишем с помощью K вероятности ошибок, ограничиваясь для удобства параметрическим случаем. Вероятность ошибки первого рода есть

$$\alpha(\theta) = \mathcal{P}_\theta(K), \theta \in \Theta_0.$$

Вероятность ошибки второго рода есть

$$\beta(\theta) = 1 - \mathcal{P}_\theta(K), \theta \in \Theta_1.$$

Функция

$$m(\theta) = 1 - \beta(\theta) = \mathcal{P}_\theta(K), \theta \in \Theta_1,$$

часто называется мощностью критерия.

Легко понять, что ограничение $\alpha(\theta) \leq \varepsilon$ означает, что критическая область K "не очень велика". Напротив, уменьшить вероятность ошибки второго рода (т.е. увеличить мощность) можно, грубо говоря, лишь за счет увеличения критической области. Тем самым, уменьшать эту вероятность можно лишь до некоторой степени (при заданном уровне эначимости).

Тест K называется равномерно наиболее мощным критерием уровня значимости ε, если для всех $\theta \in \Theta_1$

$$m(\theta) \geq m'(\theta),$$

где $m'(\theta)$ — функция мощности любого другого критерия K' с тем же уровнем значимости (равносильное неравенство $\beta(\theta) \leq \beta'(\theta)$).

Поскольку не любые две функции сравнимы между собой, равномерно наиболее мощные критерии существуют лишь в некоторых особых случаях. Два таких случая — простая альтернатива и (более общий вариант) — односторонняя альтернатива — мы рассмотрим далее. Если равномерно наиболее мощного критерия нет, приходится модифицировать постановку задачи (здесь имеется довольно глубокая аналогия с теорией оценивания). Можно ограничить класс рассматриваемых тестов, что аналогично предположениям типа несмещенности или эквивариантности в теории оценивания,

а можно ввести какой-либо числовой функционал от функции мощности, посредством которого уже сравнивать тесты (байесовские и минимаксные критерии, см. о них в [1]).

В некоторых прикладных исследованиях, связанных с проверкой простой гипотезы, уровень значимости ε заранее не фиксируется. Вместо этого рассматривается все семейство вложенных друг в друга критических областей K_ε, отвечающих данному семейству тестов, и определяется то минимальное значение ε, ниже которого основная гипотеза уже не отвергается:

$$\inf\{\varepsilon : \vec{X} \in K_\varepsilon\}.$$

Это число называется P-значением (P-value).

4.2 Построение оптимального критерия в простейшем случае — теорема Неймана-Пирсона

Разумеется простейшей является задача проверки *простой* гипотезы при *простой* альтернативе. Реального практического значения подобная ситуация не имеет, однако служит стартовой позицией для важных обобщений.

Для простой гипотезы различие между уровнем значимости ε и вероятностью ошибки первого рода α практически исчезает — с одной стороны, $\alpha \leq \varepsilon$, а с другой стороны — критерий, для которого это неравенство строгое ($\alpha < \varepsilon$), обычно можно улучшить (т.е. заменить более мощным), не меняя уровня значимости. В предыдущей фразе мы намеренно использовали довольно неопределенный термин "обычно", смысл которого постепенно будет уточняться в этом и следующем параграфах.

Для формулировки теоремы Неймана-Пирсона, указывающей наиболее мощный (слово "равномерно"здесь излишне) критерий, нам потребуется функция, называемая отношением правдоподобия (подобная функция уже возникала в параграфе 2.7 и в логарифмической форме в параграфе 2.8). В теперешней ситуации отношение правдоподобия $Z(\vec{x})$ определяется так. Если основное и альтернативное теоретические распределения непрерывны и задаются плотностями $p_0(\vec{x})$ и $p_1(\vec{x})$ (для повторной выборки эти N-мерные плотности —

произведения одномерных), то

$$Z(\vec{x}) = \frac{p_1(\vec{x})}{p_0(\vec{x})}.$$

Если же теоретические распределения дискретны, то можно воспользоваться той же формулой, только понимая p_0 и p_1 как вероятности —

$$p_i(\vec{x}) = \mathcal{P}_i(\vec{X} = \vec{x}), \quad i = 0, 1.$$

Мы в дальнейшем, как обычно, будем рассматривать случай непрерывных распределений, упоминая о дискретных выборках по мере необходимости. Во избежание малоинтересных усложнений предположим, что носители плотностей p_0 и p_1, т.е. множества вида $\{\vec{x} : p_0(\vec{x}) \neq 0\}$ и $\{\vec{x} : p_1(\vec{x}) \neq 0\}$ совпадают или почти совпадают (отличаются на множество нулевого N-мерного объема) и что отношение правдоподобия $Z(\vec{x})$ — непрерывная функция на своей области определения $\{\vec{x} : p_0(\vec{x}) \neq 0\}$.

Теорема Неймана-Пирсона (предварительная формулировка). Наиболее мощные критерии любого уровня значимости задаются критическими областями вида

$$K(c) = \{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) > c\}. \tag{4.1}$$

При этом константа c определяется по уровню значимости ε из уравнения

$$\mathcal{P}_0(K(c)) = \varepsilon. \tag{4.2}$$

Даже для непрерывных распределений, не говоря уже о дискретных, эта формулировка а) недостаточна; б) не вполне корректна (поэтому мы и назвали ее предварительной). С другой стороны, более точная формулировка оказывается более сложной и требующей развернутых пояснений. Поэтому мы начнем доказательство прямо сейчас, комментируя проблемы по ходу рассуждений. Корректная формулировка будет дана в конце доказательства, а в следующем параграфе мы и ее обобщим, введя расширенное толкование статистических тестов — так называемые рандомизированные критерии.

Заметим, тем не менее, что сама идея критических областей вида (8.1) выглядит очень естественной — чем больше степень концентрации альтернативной вероятности около точки $\vec{x}$ по сравнению с такой же концентрацией основной вероятности, тем естественнее отвергать основную гипотезу.

Итак, предположим, что $K = K(c)$ — критическая область вида (8.1), выбранная по уровню значимости ε (позже мы обсудим, как быть, если уравнение (7.2) неразрешимо). Пусть K' — критическая область другого критерия с тем же уровнем значимости (т.е. $\mathcal{P}_0(K') \le \varepsilon$). Докажем, что $m \ge m'$, т.е. что критерий K не хуже K'. Для этого заметим, что

$$m - m' = \int_K p_1(\vec{x})d\vec{x} - \int_{K'} p_1(\vec{x})d\vec{x} = \int_{K-K'} p_1(\vec{x})d\vec{x} - \int_{K'-K} p_1(\vec{x})d\vec{x}$$

(из обоих интегралов мы вычли "общую часть— интеграл по пересечению множеств $K \cap K'$). На множестве $K - K' \subset K$ выполняется неравенство $p_1(\vec{x}) > cp_0(\vec{x})$, в то время как на множестве $K' - K \subset \mathbb{R}^N - K$ — противоположное неравенство $p_1(\vec{x}) \le cp_0(\vec{x})$ (мы пользуемся определением (8.1)). Подставляя эти неравенства, получаем

$$\begin{aligned} m - m' &\ge \int_{K-K'} cp_0(\vec{x})d\vec{x} - \int_{K'-K} cp_0(\vec{x})d\vec{x} \\ &= c\left[\int_{K-K'} p_0(\vec{x})d\vec{x} - \int_{K'-K} p_0(\vec{x})d\vec{x}\right] \\ &= c\left[\int_K p_0(\vec{x})d\vec{x} - \int_{K'} p_0(\vec{x})d\vec{x}\right] \end{aligned}$$

(теперь "общая часть"добавляется обратно, уже с новой подинтегральной функцией). Остается заметить, что последнее выражение в квадратных скобках равно

$$\mathcal{P}_0(K) - \mathcal{P}_0(K') = \varepsilon - \mathcal{P}_0(K') \ge 0.$$

Обсудим теперь слабые места этого рассуждения. Таких мест можно указать два. Первое из них уже упоминалось выше — разрешимость уравнения (8.1). Второе менее существенно, но все же заслуживает обсуждения — в какой степени можно менять критическую область K, сохраняя уровень значимости и мощность.

Итак, обратимся к уравнениям (8.1) и (7.2) и для начала отметим, что выбор знака строгого неравенства в (8.1) ничем не мотивирован. Если рассмотреть множества вида

$$\bar{K}(c) = \{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) \ge c\}$$

и уравнения

$$\mathcal{P}_0(\bar{K}(c)) = \varepsilon, \tag{4.3}$$

то с ними можно повторить то же рассуждение. Тем самым, множества $\bar{K}(c)$ также можно рассматривать в качестве кандидатов на роль критических областей наиболее мощных критериев. Положим

$$f(c) = \mathcal{P}_0(K(c)),$$
$$\bar{f}(c) = \mathcal{P}(\bar{K}(c)).$$

Очевидно, обе эти функции монотонно убывают (в широком смысле), причем

$$\bar{f}(c) \geq f(c),$$
$$\bar{f}(c) = \lim_{t \nearrow c} \bar{f}(t) = \lim_{t \nearrow c} f(t) = f(c-0),$$
$$f(c) = \lim_{t \searrow c} f(t) = \lim_{t \searrow c} \bar{f}(t) = \bar{f}(c+0).$$

Мы видим, что обе функции f и $\bar{f}$ имеют одни и те же точки разрывов и отличаются как раз в них. Каждый разрыв (если, конечно, таковые существуют) порождает открытый промежуток значений ε, для которых ни уравнение (7.2) ($f(c) = \varepsilon$), ни аналогичное уравнение $\bar{f}(c) = \varepsilon$ не имеют решений. Это связано с тем, что множество уровня

$$\bar{K}(c) - K(c) = \{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) = c\}$$

имеет ненулевой объем. Проще всего исключить эту возможность дополнительным условием в формулировке теоремы.

Прямо противоположная возможность — наличие многих решений у уравнения (7.2) (или (7.3)) — может реализоваться лишь для исключительных значений ε — когда функция $f(c)$ на каком-то промежутке постоянна и равна ε (так будет, если $Z(\vec{x})$ не принимает значений из этого промежутка). Чтобы предусмотреть эту возможность в формулировке, удобно еще обозначить

$$c_-(\varepsilon) = \min\{c : \mathcal{P}_0(K(c)) = \varepsilon\},$$
$$c_+(\varepsilon) = \max\{c : \mathcal{P}_0(\bar{K}(c)) = \varepsilon\}.$$

Корректная формулировка будет выглядеть следующим образом.

Теорема Неймана-Пирсона (уточненная формулировка).

Предположим, что каждое множество уровня

$$\{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) = c\}$$

отношения правдоподобия имеет нулевой N-мерный объем (меру Лебега). Тогда для каждого $\varepsilon > 0$ уравнения (7.2) и (7.3) разрешимы, причем любое измеримое множество K, такое, что

$$K(c_-(\varepsilon)) \subset K \subset \bar{K}(c_+(\varepsilon)), \tag{4.4}$$

(все эти множества почти совпадают), дает нам критическую область наиболее мощного критерия уровня значимости ε.

Подчеркнем еще раз, что может существовать лишь не более счетного числа исключительных значений ε, для которых $c_-(\varepsilon) \neq c_+(\varepsilon)$. Для остальных ε включения (7.4) упрощаются и записываются в виде

$$K(c(\varepsilon)) \subset K \subset \bar{K}(c(\varepsilon)). \tag{4.5}$$

Остающаяся после сделанного уточнения формулировки неопределенность в форме критической области K несущественна, т.к. в эту зону вектор наблюдений $\vec{X}$ с вероятностью 1 не попадет.

Более важным для применений этой теоремы является исключенный нашей уточненной формулировкой случай, когда оба уравнения (7.2) и (7.3) могут оказаться неразрешимыми. Выход из этого положения дает рандомизация, обсуждающаяся в следующем параграфе. Для дискретных распределений, которые мы еще подробно не обсуждали, именно этот случай неразрешимости становится главным (см. параграф 3).

4.3 Рандомизация

Как общая концепция, рандомизация достаточно важна, поэтому проиллюстрируем соответствующую идею небольшим отвлеченным примером, и лишь потом вернемся к проблеме, возникшей в предыдущем параграфе.

Классический пример "неразрешимой"логической ситуации — пример "буриданова осла", не сумевшего сделать выбор между двумя равноценными охапками сена. Рандомизация дает вполне приемлемый рецепт действий в подобных ситуациях — подбрось монетку и действуй в соответствии с ее "советом". Конечно, монетка должна быть симметричной (как и охапки сена у осла), а кроме того, случайный механизм, с этой монеткой связанный, не должен быть связан с прочими сторонами возникшей ситуации — нужно обеспечить некую

"беспристрастность". На язык теории вероятностей это переводится термином "независимость— монетка должна быть не зависящей от прошлого течения явления (а также и настоящего и будущего).

Перейдем теперь к задаче предыдущего параграфа. Мы видели, что отношение прпавдоподобия устанавливает некоторую иерархию предпочтений среди возможных значений $\vec{X}$ нашей выборки — чем больше это отношение, тем менее привлекательнее выглядит основная гипотеза. Проблема возникает в том случае, когда множество

$$K(c) = \{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) > c\}$$

еще "недостаточно велико": $\mathcal{P}_0(K(c)) < \varepsilon$, в то время как множество

$$\bar{K}(c) = \{\vec{x} \in \mathbb{R}^N : Z(\vec{x}) \geq \varepsilon\}$$

уже "слишком велико": $\mathcal{P}_0(\bar{K}(c)) > \varepsilon$. Хочется расширить $K(c)$, добавляя не все точки разности $\bar{K}(c) - K(c)$ — их слишком много, а только некоторые из них. Вопрос в том, какие? С точки зрения отношения правдоподобия все они равноправны, для них $Z(\vec{x}) = c$. Другая мотивировка отсутствует. Следовательно, см. начало параграфа, нужно создать вспомогательный случайный (random) механизм, не зависящий от наших наблюдений, который бы "за нас решил", какой статистический вывод делать, если реализовавшийся набор значений $\vec{X}_{\text{эмп.}}$ оказался в "пограничной области": $Z(\vec{X}_{\text{эмп.}}) = c$. Этот механизм, испытание с двумя исходами, должен обеспечить требуемый уровень значимости ε.

Легко сообразить, что вероятность успеха p (будем, для определенности, называть успехом вывод H_0 — принятие основной гипотезы) должна удовлетворять соотношению

$$\mathcal{P}_0(K(c)) + (1-p)[\mathcal{P}_0(\bar{K}(c)) - \mathcal{P}_0(K(c))] = \varepsilon.$$

Эквивалентным образом это можно переписать в виде

$$p\mathcal{P}_0(K(c)) + (1-p)\mathcal{P}_0(\bar{K}(c)) = \varepsilon.$$

В левой части этого равенства записана выпуклая линейная комбинация двух вероятностей, одна из которых меньше ε, а другая — больше ε. Очевидно, что найдется единственное p, обеспечивающее равенство.

Для дискретных распределений без такой рандомизации фактически не обойтись, т.к. наши функции

$$f(c) = \mathcal{P}_0(K(c))$$

и

$$\bar{f}(c) = \mathcal{P}_0(\bar{K}(c))$$

принимают (в объединении) лишь дискретное множество значений. Уровень значимости ε чаще всего не совпадает ни с одним из этих значений.

Подводя итог, мы можем дать окончательную формулировку теоремы.

Теорема Неймана-Пирсона. В задаче проверки простой гипотезы при простой альтернативе для любого уровня значимости существует наиболее мощный рандомизированный критерий. Этот критерий определяется при помощи множеств $K(c)$ и $\bar{K}(c)$ и рандомизации почти единственным образом.

Для краткости мы не включили в эту формулировку подробное описание множеств $K(c)$ и $\bar{K}(c)$, а также точное значение параметра p рандомизации.

Дадим для полноты общее определение рандомизированного критерия как правила получения статистического вывода.

Критической функцией назовем отображение

$$\pi : \mathbb{R}^N \longrightarrow [0, 1].$$

Эта функция определяет вероятность $\pi(\vec{x})$ принятия основной гипотезы при $\vec{X} = \vec{x}$. Сам статистический вывод определяется случайным розыгрышем между двумя возможностями с вероятностями $\pi(\vec{x})$ и $1 - \pi(\vec{x})$ соответственно. Критерий является нерандомизированным, если его критическая функция принимает только значения 1 и 0. Множество $\pi^{-1}(0)$ при этом называется критической областью.

Вероятностью ошибки первого рода можно теперь назвать функцию

$$\alpha(\theta) = \mathbf{E}_\theta \pi(\vec{X}), \theta \in \Theta_0,$$

а вероятностью ошибки второго рода — функцию

$$\beta(\theta) = 1 - \mathbf{E}_\theta \pi(\vec{X}), \theta \in \Theta_1.$$

Сделаем в заключение параграфа несколько замечаний общего характера. Прежде всего отметим, что рандомизированные критерии по своей идее аналогичны смешанным стратегиям в теории игр. Далее, ясно, что если уже в простейшей задаче проверки гипотезы они появились, неизбежно и их появление в более общих задачах. И, наконец, последнее. Если не стремиться к оптимальности, часто без рандомизированных критериев удается обойтись.

4.4 Пример наиболее мощного критерия

Проиллюстрируем теорему Неймана-Пирсона построением наиболее мощного критерия в случае выбора из двух нормальных распределений. Еще раз стоит подчеркнуть, что саму теорему (как и описанный ниже пример) следует рассматривать лишь как начальный этап в решении более сложных задач.

Пусть основная гипотеза H_0 утверждает, что неизвестное теоретическое распределение есть $\mathbf{N}(a_0, \sigma^2)$ (оба параметра — известные числа), альтернативная гипотеза H_1 — что теоретическое распределение есть $\mathbf{N}(a_1, \sigma^2)$ (дисперсия — та же, что и в H_0, среднее значение a_1 — известное число).

Для определенности будем считать, что $a_0 < a_1$. Как скоро выяснится, это предположение приведет к правосторонней критической области. Случай $a_0 > a_1$, приводящий к левосторонней критической области, рассматривается аналогично.

Запишем отношение правдоподобия и его логарифм, обозначая положительные постоянные множители, может быть, разные, но не имеющие существенной роли, единым символом const, а постоянное слагаемое – символом constant. Итак,

$$Z(\vec{x}) = \frac{\prod_{i=1}^N \left[\frac{1}{\sigma}\varphi\left(\frac{x_i - a_1}{\sigma}\right)\right]}{\prod_{i=1}^N \left[\frac{1}{\sigma}\varphi\left(\frac{x_i - a_0}{\sigma}\right)\right]},$$

$$\begin{aligned}\ln Z(\vec{x}) &= \text{const} \sum_{i=1}^N \left[(x_i - a_0)^2 - (x_i - a_1)^2\right] \\ &= \text{const} \sum_{i=1}^N (a_1 - a_0) x_i + \text{constant} = \text{const} \sum_{i=1}^N x_i + \text{constant}.\end{aligned}$$

Последний переход использовал предположение $a_0 < a_1$.

По теореме Неймана-Пирсона заключаем, что критическая область наиболее мощного критерия имеет вид

$$K = K(c) = \{\vec{x} \in \mathbb{R}^N : \bar{x} > c\}.$$

Другими словами, основную гипотезу следует отвергнуть, если $\bar{X} > c$. Как мы сейчас увидим, такое c определяется по уровню значимости ε однозначно, а рандомизации не требуется.

Для нахождения c заметим, что, в предположении справедливости основной гипотезы,

$$\bar{X} \in \mathbf{N}(a_0, \sigma^2/N),$$

так что

$$\sqrt{N}\frac{\bar{X} - a_0}{\sigma} \in \mathbf{N}(0, 1).$$

Выбирая $z = z_\varepsilon = \Phi^{-1}(1 - \varepsilon)$, мы получаем

$$\sqrt{N}\frac{c - a_0}{\sigma} = z,$$

т.е.

$$c = a_0 + z\frac{\sigma}{\sqrt{N}}.$$

Разумеется, осмысленный уровень значимости ε должен предполагаться меньшим, чем $1/2$, а тогда

$$z > \Phi^{-1}(1/2) = 0, \quad c > a_0.$$

В зависимости от соотношения между a_0, a_1, z, σ и N возможны два варианта:

основной: $a_0 < a_0 + z\frac{\sigma}{\sqrt{N}} < a_1$;

дополнительный: $a_0 + z\frac{\sigma}{\sqrt{N}} \geq a_1$.

При фиксированных a_0, a_1, z, σ и достаточно больших N выполняется основной вариант, не вызывающий каких-либо недоумений. В частности, если $\overline{X_{\text{эмп.}}} = a_1$, основная гипотеза отвергается. При малых N, больших σ или малой разности $a_1 - a_0$ может реализоваться дополнительный вариант, который, в частности, приводит к тому, что "рецепт"наиболее мощного критерия выглядит противоречащим здравому смыслу: если $\overline{X_{\text{эмп.}}} = a_1$, этот рецепт призывает отвергнуть альтернативу в пользу основной гипотезы!

Объяснение этого эффекта весьма прозаично: число наблюдений N слишком мало, чтобы отличить одну гипотезу от другой — за счет близости a_0 и a_1, или за счет большого разброса σ сделать это разумным образом невозможно: даже оптимальный (т.е. наиболее мощный) тест не позволяет эти гипотезы различить. Совет может быть один — увеличивать число наблюдений и получать из них дополнительную информацию.

Заключительный комментарий. Напомним, что сама постановка задачи о различении двух простых гипотез малореалистична, так

что буквального применения только что высказанные толкования и рекомендации не имеют. Однако они очень хорошо передают дух проблемы и описывают возможный (надо полагать, единственно возможный) путь разрешения трудностей.

4.5 Использование монотонности отношения правдоподобия

В этом параграфе обсуждаются идеи, позволяющие иногда находить равномерно наиболее мощные критерии при сложных альтернативах. Основой для надежд на получение подобных результатов может служить простое замечание, относящееся к примеру из предыдущего параграфа. Именно, критическая область оказалась одной и той же для всех $a_1 > a_0$.

Рассмотрим для начала параметрическую гипотезу H_0 вида $\theta \leq \theta_0$ при альтернативе H_1 вида $\theta > \theta_0$. Такую альтернативу естественно назвать односторонней. Мы сможем обсудить подобную задачу проверки при специальном предположении о параметрическом семействе априори допустимых распределений.

Для определенности будем считать, что совместное распределение выборки задается плотностью $p_\theta(\vec{x})$ (дискретный случай можно рассматривать аналогично). Будем, далее, предполагать, что существует одномерная достаточная статистика $T = T(\vec{X})$, так что (вспомним теорему факторизации из параграфа 2.6)

$$p_\theta(\vec{x}) = \psi(T(\vec{x}), \theta)h(\vec{x}).$$

Отношение правдоподобия в этом случае представляется в виде

$$Z(\vec{x}; \theta_1, \theta_2) = \frac{p_{\theta_2}(\vec{x})}{p_{\theta_1}(\vec{x})} = \frac{\psi(T(\vec{x}), \theta_2)}{\psi(T(\vec{x}), \theta_1)},$$

т.е. как функция от достаточной статистики.

Будем говорить, что семейство мер $\mathcal{P}_\theta$ имеет монотонное отношение правдоподобия, если при фиксированных $\theta_1 < \theta_2$ функция $Z(\vec{x}; \theta_1, \theta_2)$ является монотонно возрастающей функцией от достаточной статистики:

если $T(\vec{x}) \leq T(\vec{x}')$,

то $Z(\vec{x}; \theta_1, \theta_2) \leq Z(\vec{x}'; \theta_1, \theta_2)$.

Ограничение *возрастающими* функциями несущественно: убывающую функцию аргумента T можно истолковать и как

возрастающую функцию аргумента $-T$, а выбор варианта достаточной статистики (T или $-T$) зависит от нас.

Очевидно, что в примере из предыдущего параграфа отношение правдоподобия было монотонным.

Если условие монотонности выполнено, то неравенство вида

$$Z(\vec{x};\theta_1,\theta_2) > c$$

можно равносильным образом переписать в виде

$$T(\vec{x}) > c' \quad \text{или} \quad T(\vec{x}) \geq c',$$

где c' однозначно определяется по c,θ_1, θ_2 и N. Вторая возможность может реализоваться в точках разрыва отношения правдоподобия как функции достаточной статистики T.

Сформулируем теперь результат, относящийся к случаю односторонней альтернативы.

Теорема 1. Если семейство априори допустимых распределений $\mathcal{P}_\theta$ имеет монотонное отношение правдоподобия, то существует равномерно наиболее мощный рандомизированный критерий проверки гипотезы $H_0 = \{\theta \leq \theta_0\}$ при односторонней альтернативе $H_1 = \{\theta > \theta_0\}$. Этот критерий имеет вид:

если $T(\vec{X}) > c$, то H_0 отвергается;
если $T(\vec{X}) = c$, то H_0 отвергается
с некоторой вероятностью p;
если $T(\vec{X}) < c$, то H_0 не отвергается
(т.е. принимается).

Числа c и p определяются по уровню значимости ε и распределению $\mathcal{P}_{\theta_0}$ так же, как в теореме Неймана-Пирсона:

$$\mathcal{P}_{\theta_0}(T(\vec{X}) > c) \leq \varepsilon \leq \mathcal{P}_{\theta_0}(T(\vec{X}) \geq c),$$
$$p\mathcal{P}_{\theta_0}(T(\vec{X}) > c) + (1-p)\mathcal{P}_{\theta_0}(T(\vec{X}) \geq c) = \varepsilon.$$

При этом мощность критерия $m(\theta)$ строго возрастает по θ. Кроме того, при каждом $\theta < \theta_0$, указанный критерий минимизирует ошибку первого рода $\alpha(\theta)$.

В терминах критической функции $\pi(\vec{X})$ можно записать вид нашего критерия более коротким образом:

$$\pi(\vec{X}) = \begin{cases} 1, \text{ если } T(\vec{X}) > c, \\ p, \text{ если } T(\vec{X}) = c, \\ 0, \text{ если } T(\vec{X}) < c; \end{cases}$$

$$\mathbf{E}_{\theta_0}\pi(\vec{X}) = \varepsilon.$$

Эта теорема (кроме последнего утверждения) почти автоматически следует из теоремы Неймана-Пирсона. Для получения последнего утверждения нужно поменять местами гипотезы H_0 и H_1 и снова воспользоваться теоремой Неймана-Пирсона. Мы опускаем все формальные детали соответствующих рассуждений.

Частный случай сформулированной выше теоремы относится к экспоненциальным семействам

$$p_\theta(\vec{x}) = h(\vec{x})\exp\{\hat{\theta}(\vec{x})A(\theta) + B(\theta)\} \tag{4.6}$$

(см. параграф 2.4). Монотонность отношения правдоподобия при этом превращается в монотонность функции $A(\theta)$:

$$Z(\vec{x};\theta_1,\theta_2) = \exp\{\hat{\theta}(\vec{x})[A(\theta_2) - A(\theta_1)] + [B(\theta_2) - B(\theta_1)]\}$$

и знак разности $A(\theta_2) - A(\theta_1)$ должен быть постоянным при $\theta_1 < \theta_2$.

Пример экспоненциальных семейств (или даже пример нормальных распределений из предыдущего параграфа) почти очевидным образом показывает, что при двусторонней альтернативе (например, $H_0 = \{\theta = \theta_0\}$, $H_1 = \{\theta \neq \theta_0\}$) равномерно наиболее мощного критерия не существует (критическая область не может одновременно оказаться лево- и правосторонней).

Тем не менее, и в двустороннем случае удается при некоторых предположениях получить равномерно наиболее мощный критерий, поменяв ролями основную и альтернативную гипотезы. Сформулируем без доказательства соответствующий результат (см. [1])

Теорема 2. Предположим, что для однопараметрического экспоненциального семейства (4.6) функция $A(\theta)$ монотонна, а $\theta_1 < \theta_2$ — два значения параметра. Тогда для задачи проверки гипотезы $H_0 = \{\theta \notin]\theta_1,\theta_2[\}$ при альтернативе $H_1 = \{\theta \in]\theta_1,\theta_2[\}$ равномерно наиболее мощный критерий существует и имеет вид:

если $c_1 < T(\vec{X}) < c_2$, то H_0 отвергается;
если $T(\vec{X}) = c_1$, то H_0 отвергается с некоторой вероятностью p_1;
если $T(\vec{X}) = c_2$, то H_0 отвергается с некоторой вероятностью p_2;
если $T(\vec{X}) < c_1$ или $T(\vec{X}) > c_2$, то

H_0 не отвергается.

Числа c_1, c_2, p_1, p_2 определяются по уровню значимости ε и распределениям $\mathcal{P}_{\theta_1}$ и $\mathcal{P}_{\theta_2}$ так же, как в теореме Неймана-Пирсона:

$$\mathcal{P}_{\theta_1}(c_1 < T(\vec{X}) < c_2) \le \varepsilon \le \mathcal{P}_{\theta_1}(c_1 \le T(\vec{X}) \le c_2),$$
$$\mathcal{P}_{\theta_2}(c_1 < T(\vec{X}) < c_2) \le \varepsilon \le \mathcal{P}_{\theta_2}(c_1 \le T(\vec{X}) \le c_2),$$
$$p_1\mathcal{P}_{\theta_1}(c_1 < T(\vec{X}) < c_2) + (1-p_1)\mathcal{P}_{\theta_1}(c_1 \le T(\vec{X}) \le c_2),$$
$$p_2\mathcal{P}_{\theta_2}(c_1 < T(\vec{X}) < c_2) + (1-p_2)\mathcal{P}_{\theta_2}(c_1 \le T(\vec{X}) \le c_2).$$

Наиболее трудным техническим местом здесь является нахождение c_1 и c_2.

И эту формулировку можно сжато записать в терминах критической функции:

$$\pi(\vec{X}) = \begin{cases} 1, \text{ если } c_1 < T(\vec{X}) < c_2, \\ p_1, \text{ если } T(\vec{X}) = c_1, \\ p_2, \text{ если } T(\vec{X}) = c_2, \\ 0, \text{ если } T(\vec{X}) < c_1 \text{ или } T(\vec{X}) > c_2, \end{cases}$$

$$\mathbf{E}_{\theta_1}\pi(\vec{X}) = \mathbf{E}_{\theta_2}\pi(\vec{X}) = \varepsilon.$$

Аналогичный результат имеет место и для основной гипотезы вида $H_0 = \{\theta \notin [\theta_1, \theta_2]\}$ и соответствующей альтернативы.

4.6 Несмещенные и инвариантные критерии

Мы продолжаем обсуждение таких постановок задач, когда существуют равномерно наиболее мощные тесты. Еще один путь состоит в сужении класса критериев, из которых разрешается выбирать. В теории оценивания рассматривался аналогичный прием — выделение класса несмещенных оценок K_0, класса эквивариантных оценок K_{eq} и т.п. Прежде чем вводить класс несмещенных критериев, играющий сходную роль, напомним, см. параграф 3, что для рандомизированных критериев можно определить вероятности ошибок и мощность почти так же, как и для нерандомизированных:

$$\alpha(\theta) = \mathbf{E}_\theta\pi(\vec{X}),\ \theta \in \Theta_0,$$
$$m(\theta) = \mathbf{E}_\theta\pi(\vec{X}),\ \theta \in \Theta_1,$$
$$\beta(\theta) = 1 - m(\theta),$$

где $\pi(\vec{X})$ — критическая функция рандомизированного критерия. Дадим теперь нужное определение.

Критерий называется несмещенным, если

$$\inf_{\theta\in\Theta_1} m(\theta) \geq \sup_{\theta\in\Theta_0} \alpha(\theta).$$

Если ограничиваться критериями с заданным уровнем значимости ε, т.е. с

$$\sup_{\theta\in\Theta_0} \alpha(\theta) = \varepsilon,$$

условие несмещенности превращается в

$$m(\theta) \geq \varepsilon \text{ при всех } \theta \in \Theta_1.$$

Наглядный смысл условия несмещенности в том, что вероятность $m(\theta)$ отвергнуть основную гипотезу H_0 в том случае, когда она несправедлива, никогда не оказывается меньше вероятности $\alpha(\theta)$ отвергнуть основную гипотезу H_0 в том случае, когда она справедлива.

Следующие очень простые соображения отчасти мотивируют введение несмещенных тестов. Во-первых, легко видеть, что равномерно наиболее мощный тест, если он существует, обязан быть несмещенным. Действительно, тривиальный критерий с критической функцией $\pi(\vec{X}) \equiv \varepsilon$, вообще не использующий выборку, имеет мощность ε. Следовательно, мощность наиболее мощного критерия должна быть не меньше. Во-вторых, требование несмещенности исключает те односторонние критерии, которые препятствовали существованию наиболее мощного теста при двусторонней альтернативе.

Приведем без доказательства (см. [1]) результат, дополняющий теорему 2 из параграфа 5.

Теорема 1. Предположим, что для однопараметрического экспоненциального семейства (4.6) функция $A(\theta)$ монотонна, а $\theta_1 \leq \theta_2$ — два значения параметра. Тогда для задачи проверки гипотезы $H_0 = \{\theta \in [\theta_1, \theta_2]\}$ при альтернативе $H_1 = \{\theta \notin [\theta_1, \theta_2]\}$ в классе несмещенных критериев существует равномерно наиболее мощный. Его критическая функция в случае строго неравенства $\theta_1 < \theta_2$ задается формулой

$$\pi(\vec{X}) = \begin{cases} 0,\ c_1 < T(\vec{X}) < c_2, \\ p_1,\ T(\vec{X}) = c_1, \\ p_2,\ T(\vec{X}) = c_2, \\ 1,\ T(\vec{X}) < c_1 \text{ или } T(\vec{X}) > c_2. \end{cases}$$

Описание критической функции в случае $\theta_1 = \theta_2$ мы не приводим (см. [1]).

Еще одно возможное сужение класса рассматриваемых критериев — требование инвариантности. Рассмотрим соответствующие идеи на примере.

Пусть $X_1, \ldots, X_N$ — повторная выборка, имеющая распределение $\mathbf{N}(a, \sigma^2)$ (оба параметра неизвестны). Рассмотрим основную гипотезу $H_0 = \{\sigma^2 \in [\sigma_1^2, \sigma_2^2], a \in \mathbb{R}\}$ при альтернативе $H_1 = \{\sigma^2 \notin [\sigma_1^2, \sigma_2^2], a \in \mathbb{R}\}$. Очевидно, H_0 и H_1 "инвариантны относительно сдвига— о среднем значении ничего не предполагается. Достаточная статистика $(\bar{X}, S^2)$ имеет две компоненты с разным поведением при сдвиге: $\bar{X}$ эквивариантна, а S^2 — инвариантна (см. параграф 2.9). Естественно предположить, что проверка инвариантной гипотезы H_0 должна основываться на S^2. Такие критерии также называются инвариантными. Можно доказать, что наиболее мощный инвариантный критерий (притом нерандомизированный) существует и его критическая область имеет вид $\{S^2 \notin [c_1, c_2]\}$. Числа c_1 и c_2 выбираются так, чтобы по каждому из распределений $\mathbf{N}(a, \sigma_1^2)$ и $\mathbf{N}(a, \sigma_2^2)$ вероятность ошибки первого рода была равна ε. Сделать такой выбор не слишком сложно, т.к. по лемме Фишера

$$\frac{NS^2}{\sigma^2} \in \chi^2_{N-1}.$$

4.7 Критерий хи-квадрат

В предыдущих параграфах мы видели, что общая теория проверки гипотез весьма сложна. Поэтому, изложив некоторые основные ее идеи, мы посвятим остаток главы обсуждениям ряда популярных статистических тестов, первым из которых является критерий хи-квадрат.

Первый вариант критерия хи-квадрат.

Предположим, что случайные величины $X_1, \ldots, X_N$, составляющие выборку, принимают значения из конечного множества $E = \{e_1, \ldots, e_r\}$, а соответствующие вероятности $p_j = \mathsf{P}(X_i = e_j)$, составляющие в сумме единицу, образуют вектор параметров:

$$\vec{p} = (p_1, \ldots, p_r)^T \in \mathbb{R}^r, \ \sum_{i=1}^{r} p_j = 1.$$

Нам будет удобно предположить, что конечное множество E выбрано специальным образом:

$$E = \{e_1 = (1, 0, \ldots, 0)^T, e_2 = (0, 1, 0, \ldots, 0)^T, \\ \ldots, e_r = (0, 0, \ldots, 1)^T\},$$

т.е. состоит из векторов r-мерного пространства $\mathbb{R}^r$, одна из компонент которых равна единице, а остальные — нулю. При изучении испытаний Бернулли мы поступали похожим образом — кодировали успех числом 1, а неудачу — числом 0. Сейчас испытания Бернулли (они соответствуют $r = 2$) закодированы чуть иначе: успех — вектором $(1, 0)^T$, а неудача — вектором $(0, 1)^T$. Наше теперешнее векторное представление испытаний с r исходами немного избыточно, зато все исходы равноправны.

Итак, мы будем иметь дело с векторными наблюдениями $X_1, \ldots, X_N$ с очень простыми значениями из $\mathbb{R}^r$. Как обычно, будем рассматривать сумму

$$S_N = X_1 + \cdots + X_N.$$

Компоненты этого случайного вектора S_N имеют смысл кратностей появления отдельных исходов в нашей выборке. Они будут обозначаться $n_1, \ldots, n_r$:

$$S_N = (n_1, \ldots, n_r)'.$$

Очевидно,

$$\sum_{j=1}^{r} n_j = N.$$

Распределение случайного вектора S_N часто называется полиномиальным. Оно задается обобщенной формулой Бернулли:

$$\mathsf{P}(n_1 = k_1, \ldots, n_r = k_r) = \frac{N!}{k_1! \ldots k_r!} p_1^{k_1} \cdots p_r^{k_r} \\ (k_1, \ldots, k_r \geq 0, k_1 + \cdots + k_r = N).$$

Впрочем, эта формула далее не потребуется, поскольку мы будем действовать в рамках асимптотического подхода и заменим полиномиальное распределение аппроксимирующим его многомерным нормальным.

Согласно многомерному варианту центральной предельной теоремы Леви́ (см. одномерный ее вариант в параграфе 1.4) распределение

случайного вектора

$$\frac{S_N - \mathbf{E}S_N}{\sqrt{N}}$$

слабо сходится при $N \to \infty$ к r-мерному нормальному распределению с нулевым средним, матрица ковариаций которого совпадает с матрицей ковариаций C отдельного наблюдения X_i. Легко вычислить, что

$$\mathbf{E}S_N = N\mathbf{E}X_1 = N\sum_{j=1}^{r} p_j e_j = N\vec{p}.$$

Аналогично вычисляется матрица ковариаций C, а через нее и матрица ковариаций S_N. Для проведения этого вычисления обозначим через $(X_1)_j$ $(j = 1, \ldots, r)$ компоненты случайного вектора X_1. Каждая из них принимает два значения (1 и 0), причем $\mathsf{P}((X_1)_j = 1) = p_j$. Поэтому

$$c_{jj} = \mathbf{V}[(X_1)_j] = p_j(1 - p_j).$$

Для вычисления внедиагональных элементов (ковариаций) c_{j_1,j_2} $(j_1 \neq j_2)$ заметим, что $(X_1)_{j_1} \cdot (X_1)_{j_2} \equiv 0$. Поэтому

$$c_{j_1,j_2} = \mathsf{cov}[(X_1)_{j_1}, (X_1)_{j_2}] = -\mathbf{E}[(X_1)_{j_1}] \cdot \mathbf{E}[(X_1)_{j_2}] = -p_{j_1}p_{j_2}.$$

Кроме того, очевидно, что

$$\mathsf{cov}(S_N) = N\mathsf{cov}(X_1) = NC$$

(векторы $X_1, \ldots, X_N$ независимы, а при сложении независимых векторов их ковариационные матрицы также складываются).

Перейдем теперь к постановке задачи проверки гипотезы и к описанию критерия хи-квадрат, дающего асимптотическое решение этой задачи.

Простая гипотеза, которую мы будем проверять, имеет вид $H_0 = \{\vec{p} = \vec{p}^{\,0}\}$, где $\vec{p}^{\,0}$ — известный гипотетический вектор параметров, удовлетворяющий естественному условию $\sum_{j=1}^{r} p_j^0 = 1$. В качестве альтернативы берется отрицание основной гипотезы: $H_1 = \{\vec{p} \neq \vec{p}^{\,0}\}$.

Для построения статистического критерия Пирсон предложил статистику

$$\Pi = \sum_{j=1}^{r} \left(\frac{n_j - Np_j^0}{\sqrt{Np_j^0}} \right)^2.$$

Как мы видели чуть выше, случайные величины

$$\frac{n_j - Np_j}{\sqrt{Np_j}}$$

асимптотически нормальны, хотя и зависимы при разных j. Из проделанных вычислений следует, что

$$\mathbf{V}\left(\frac{n_j - Np_j}{\sqrt{Np_j}}\right) = \frac{1}{p_j}\mathbf{V}\left(\frac{n_j - Np_j}{\sqrt{N}}\right) = \frac{c_{jj}}{p_j} = 1 - p_j.$$

Тем не менее, замечательным образом оказывается, что сумма их квадратов имеет асимптотически распределение хи-квадрат. Этот результат, собственно, и привел Пирсона к успеху.

Теорема Пирсона. Предположим, что основная гипотеза $H_0 = \{\vec{p} = \vec{p}^{\,0}\}$ справедлива. Тогда при $N \to \infty$ распределение случайной величины Π слабо сходится к распределению хи-квадрат с $r-1$ степенями свободы:

$$\mathsf{P}^0(\Pi < z) \to \mathsf{P}(\chi^2_{r-1} < z).$$

Теорему Пирсона мы докажем в следующем параграфе, а сейчас перейдем к статистическим приложениям ее.

Заметим сначала, что Π — взвешенная сумма квадратов отклонений. Точнее,

$$\Pi = N\sum_{j=1}^{r}\frac{1}{p_j^0}\left(\frac{n_j}{N} - p_j^0\right)^2 = N\sum_{j=1}^{r}\frac{1}{p_j^0}(\hat{p}_j - p_j^0)^2,$$

где $\hat{p}_j = \frac{n_j}{N}$ — известные нам из главы 2 эффективные несмещенные оценки вероятностей p_j. В предположении справедливости гипотезы H_0 эти оценки сходятся именно к гипотетическим вероятностям p_j^0, а тогда Π, как сумма квадратов отклонений, с подавляющей вероятностью не слишком велика. Поэтому естественно рассмотреть критерий, имеющий критическую область вида

$$K = \{\Pi > z\}.$$

Теорема Пирсона дает нам возможность найти такое z по уровню значимости ε, пользуясь (асимптотическим) шаблоном χ^2_{r-1}. Выбирая z_ε из соотношения

$$\mathsf{P}(\chi^2_{r-1} < z_\varepsilon) = 1 - \varepsilon,$$

мы получаем критическую область

$$K_\varepsilon = \{\Pi > z_\varepsilon\},$$

имеющую асимптотически требуемый уровень значимости ε. Это и есть критерий хи-квадрат, предложенный Пирсоном. Можно доказать, что он асимптотически оптимален (см. [1]).

С практической точки зрения значительно бóльшую ценность имеют другие версии критерия хи-квадрат, основанные на рассмотренном простейшем варианте.

Второй вариант критерия хи-квадрат: простая гипотеза.

Рассмотрим задачу проверки одномерной простой гипотезы $H_0 = \{F = F^0\}$ при альтернативе $H_1 = \{F \neq F^0\}$. Здесь F^0 — известная гипотетическая функция распределения на прямой $\mathbb{R}$. Образуем вспомогательную гипотезу H_0^*, являющуюся следствием H_0, следующим образом. Разобьем числовую ось $\mathbb{R}$ на r дизъюнктных промежутков $\Delta_1, \ldots, \Delta_r$ и преобразуем выборку $X_1, \ldots, X_N$, положив

$$X_i^* = e_j, \text{ если } X_j \in \Delta_j \ (j = 1, \ldots, r).$$

Пусть, как и в первом варианте,

$$p_j = \mathsf{P}(X_i = e_j),$$

а p_j^0 — соответствующие гипотетические вероятности,

$$p_j^0 = \mathsf{P}^0(X_i = e_j).$$

Предположение

$$H_0^* = \{\vec{p} = \vec{p}^{\,0}\}$$

является простой гипотезой по отношению к преобразованной выборке $X_1^*, \ldots, X_N^*$ и сложной — по отношению к исходной выборке $\vec{X}$. Очевидно, что H_0^* действительно является следствием H_0: гипотеза H_0 утверждает, что **все** теоретические вероятности вычисляются по функции распределения F^0, а гипотеза H_0^* — что **некоторые** вероятности, именно, вероятности попадания в промежутки Δ_j, вычисляются по функции распределения F^0.

Гипотезу H_0^* можно проверять при помощи критерия хи-квадрат, описанного выше. Если она отвергается при некотором уровне значимости ε, то и H_0 следует отвергнуть — принятие H_0 влечет принятие и H_0^* как следствия. Сложнее обстоит дело в случае, когда H_0^* не отвергается. Формально при этом о справедливости или несправедливости H_0 мы не получаем никакого суждения. Единственное, что можно отметить, — что чем мельче промежутки Δ_j, тем

"ближе"становится H_0^* к H_0. Доводов в пользу H_0 оказывается меньше, чем в других задачах проверки гипотез. Для выборки очень большого объема (мы сейчас будем обсуждать этот вопрос подробнее), видимо, все-таки можно надеяться, что тест, основанный на статистике П, даст удовлетворительный результат. Впрочем, указать фактический уровень значимости (даже асимптотически) весьма проблематично.

Качество теста хи-квадрат в рассматриваемой ситуации достаточно сильно зависит от выбора промежутков $\Delta_1, \ldots, \Delta_r$ и от их числа. Как мы увидим в параграфе 8, теорема Пирсона и по форме, и по доказательству похожа на теорему Муавра-Лапласа. Как известно, качество нормальной аппроксимации теоремы Муавра-Лапласа определяется величиной Npq. Часто предлагают пользоваться ею при $Npq > 20$. Не обсуждая эту рекомендацию по существу (она заведомо имеет символический характер), перенесем ее догматично на теорему Пирсона: $Np_j^0(1-p_j^0) > 20$ при всех $j = 1, \ldots, r$. Такой подход даст нам хоть какой-то ориентир.

Прежде всего отметим, что добиться одновременного выполнения всех этих неравенств проще всего в случае $p_1^0 = \cdots = p_r^0 = \frac{1}{r}$. Грубая оценка дает тогда $r < N/20$. Примерно так обычно и рекомендуют действовать. Отметим одну потенциальную опасность, подстерегающую неосторожных исследователей. Может возникнуть желание подобрать интервалы Δ_j, определяющие "группировку эмпирических данных", опираясь на сами эти данные. Разумеется, этот прием является жульничеством, которое иногда может "обеспечить"значительно большее согласие с проверяемой гипотезой, чем фактическое.

Можно доказать, что асимптотически, т.е. при $N \to \infty$, критерий хи-квадрат имеет уровень значимости ε, хотя и не в состоянии отличить распределение F^0 от других распределений, имеющих те же вероятности интервалов $\Delta_1, \ldots, \Delta_r$.

Третий вариант критерия хи-квадрат: сложная параметрическая гипотеза.

Этот и следующий варианты мы рассмотрим очень бегло, отсылая за деталями к подробным учебникам математической статистики ([8], [1]).

Рассмотрим сложную параметрическую гипотезу вида $H_0 = \{F \in (F_\theta)\}$, где (F_θ) — некоторое семейство распределений, зависящее от параметра θ. Размерность параметра θ мы обозначим буквой s. Предлагается свести эту задачу к предыдущей, оценив предварительно параметр θ по той же выборке и взяв в качестве F^0 распределение с соответствующим значением параметра, т.е. взяв $F^0 = F_{\hat{\theta}}$. Можно

установить, что если оценка $\hat{\theta}$ асимптотически оптимальна (например, является оценкой максимального правдоподобия, построенной по частотам группировки, но не по исходной выборке!), то распределение статистики Π слабо сходится к χ^2_{r-s-1} (напомним, что s — размерность параметра). На широко распространенном жаргоне этот результат выражают словами: "каждый оцененный по выборке параметр съедает одну степень свободы".

Четвертый вариант критерия хи-квадрат: независимость признаков.

Этот вариант относится к двумерным выборкам вида (X_i, Y_i), где каждая из величин X_i принимает одно из r значений $e_1, \ldots, e_r$, а каждая из величин Y_i — одно из s значений $f_1, \ldots, f_s$. Проверяется гипотеза независимости признаков X и Y. Положим

$$p_{jk} = \mathsf{P}(X_i = e_j, Y_i = f_k),$$

$$p_{j\cdot} = \sum_k p_{jk}, \quad p_{\cdot k} = \sum_j p_{jk}.$$

Гипотеза независимости имеет вид

$$H_0 = \{p_{jk} = p_{j\cdot} \cdot p_{\cdot k} \text{ при всех } j \text{ и } k\}.$$

Для проверки ее предлагается рассмотреть соответствующие кратности n_{jk}, $n_{j\cdot}$, $n_{\cdot k}$ и образовать величину

$$\Pi = \sum_{j,k} \left(\frac{n_{jk} - N\hat{p}_{j\cdot}\hat{p}_{\cdot k}}{\sqrt{N\hat{p}_{j\cdot} \cdot \hat{p}_{\cdot k}}} \right)^2,$$

где $\hat{p}_{j\cdot} = \frac{n_{j\cdot}}{N}$ и $\hat{p}_{\cdot k} = \frac{n_{\cdot k}}{N}$ — оценки соответствующих вероятностей. Можно доказать, что распределение случайной величины Π слабо сходится к $\chi^2_{(r-1)(s-1)}$. Число степеней свободы согласуется с приведенным выше жаргонным тезисом:

$$(rs - 1) - (r + s - 2) = (r - 1)(s - 1)$$

($r+s-2 = (r-1)+(s-1)$ — количество вероятностей $p_{j\cdot}$ и $p_{\cdot k}$, оцененных по выборке).

4.8 Доказательство теоремы Пирсона.

На протяжении всего доказательства мы предполагаем, что основная гипотеза $H_0 = \{\vec{p} = \vec{p}^{\,0}\}$ справедлива.

Как было отмечено в предыдущем параграфе, случайный вектор

$$S^* = \frac{S_N - \mathbf{E}S_N}{\sqrt{N}}$$

асимтотически нормален — его распределение слабо сходится к r-мерному нормальному распределению $\mathbf{N}(0, C)$. Соответствующая матрица ковариаций C (она найдена в предыдущем параграфе) вырождена, поскольку компоненты n_j вектора S_N линейно зависимы:

$$n_1 + \cdots + n_r = N.$$

Для компонент вектора S^*, как следствие, выполняется соотношение

$$\sum_{j=1}^{r} \frac{n_j - Np_j^0}{\sqrt{N}} = 0.$$

Таким образом, его распределение сосредоточено в гиперплоскости

$$\{\vec{x} \in \mathbb{R}^r : x_1 + \cdots + x_r = 0\}$$

r-мерного пространства. В этой же гиперплоскости сосредоточено и предельное распределение $\mathbf{N}(0, C)$. Рассмотрим вспомогательный случайный вектор Z с компонентами

$$Z_j = \frac{1}{\sqrt{p_j^0}} \frac{n_j - Np_j^0}{\sqrt{N}},\ j = 1, \ldots, r.$$

Его распределение сосредоточено в гиперплоскости

$$\{\vec{x} \in \mathbb{R}^r : \sqrt{p_1^0}x_1 + \cdots + \sqrt{p_r^0}x_r = 0\}. \tag{4.7}$$

Очевидно, что вектор Z получается из S^* умножением на диагональную матрицу

$$D = \operatorname{diag}(\frac{1}{\sqrt{p_1^0}}, \cdots, \frac{1}{\sqrt{p_r^0}}).$$

Вычислим матрицу ковариаций вектора $Z = DS^*$:

$$\mathsf{cov}(Z) = \mathbf{E}(ZZ^T) = \mathbf{E}[(DS^*)(DS^*)^T] = D\mathbf{E}(S^*(S^*)^T)D^T = DCD.$$

Отсюда

$$\mathbf{V}(Z_j) = d_{jj}c_{jj}d_{jj} = 1 - p_j^0,$$
$$\mathsf{cov}(Z_{j_1}, Z_{j_2}) = d_{j_1j_1}c_{j_1j_2}d_{j_2j_2} = -\sqrt{p_{j_1}^0 p_{j_2}^0}\ (j_1 \neq j_2).$$

Обозначая $\tau_j = \sqrt{p_j^0}$, мы можем записать матрицу DCD в виде

$$DCD = \mathbf{1}_r - \vec{\tau}\vec{\tau}',$$

где $\vec{\tau}$ — вектор, составленный из компонент τ_j, $j = 1, \ldots, r$. Распределение вектора Z слабо сходится к нормальному распределению $\mathbf{N}(0, DCD)$. Мы сейчас проверим, что "если это нормальное распределение рассматривать в гиперплоскости (4.7), то его матрица ковариаций окажется единичной". Расшифруем заключенное в кавычки выражение. Пусть T — вспомогательный случайный вектор в $\mathbb{R}^r$ с нулевым средним ($\mathbf{E}T = 0$), имеющий матрицу ковариаций DCD, а $e_1, \ldots, e_{r-1}$ — произвольный ортонормированный базис в гиперплоскости (4.7). Введем одномерные случайные величины $\tilde{T}_j = e_j'T$ $(j = 1, \ldots, r-1)$ и составим из них вектор $\tilde{T}$ в $\mathbb{R}^{r-1}$. Тогда матрица ковариаций вектора $\tilde{T}$ — единичная.

Для доказательства рассмотрим

$$\mathsf{cov}(\tilde{T}_{j_1}, \tilde{T}_{j_2}) = \mathbf{E}(\tilde{T}_{j_1}\tilde{T}_{j_2}) = \mathbf{E}(\tilde{T}_{j_1}\tilde{T}_{j_2}') =$$
$$= \mathbf{E}(e_{j_1}'TT'e_{j_2}) = e_{j_1}'DCDe_{j_2} = e_{j_1}'e_{j_2} - e_{j_1}'\vec{\tau}\vec{\tau}'e_{j_2} = e_{j_1}'e_{j_2}$$

(поскольку векторы $e_1, \ldots, e_{r-1}$ лежат в гиперплоскости (4.7), они ортогональны вектору $\vec{\tau}$, т.е. выполняются равенства $\vec{\tau}'e_{j_2} = e_{j_1}'\vec{\tau} = 0$). Таким образом, для любого такого вектора $\tilde{T}$ и в любом ортонормированном базисе матрица ковариаций — единичная.

Рассматривая вектор Z как вектор $\tilde{Z}$ в гиперплоскости (4.7), мы видим, что его распределение слабо сходится к стандартному нормальному распределению в этой гиперплоскости.

Остается заметить, что

$$\Pi = \sum_{j=1}^{r} Z_j^2 = \sum_{j=1}^{r-1} \tilde{Z}_j^2.$$

Отображение, переводящее вектор $\tilde{Z}$ в сумму квадратов его компонент, непрерывно. Поэтому распределение величины Π слабо сходится к

распределению суммы квадратов независимых $\mathbf{N}(0,1)$-величин, т.е. к хи-квадрат. Число степеней свободы определяется размерностью гиперплоскости (4.7), т.е. равно $r-1$.

4.9 Непараметрический критерий Колмогорова

В этом параграфе снова пойдет речь о проверке простой гипотезы вида $H_0 = \{F = F^0\}$, где F^0 — конкретная непрерывная функция распределения. Включать F^0 в какое-либо параметрическое семейство не потребуется, поэтому и критерий называется непараметрическим. В основе его лежит максимальное расхождение эмпирической и гипотетической функций распределения:

$$D_N^0 = \sup_x |F_N^*(x) - F^0(x)|.$$

Для удобства мы рассмотрим еще и аналогичное отклонение эмпирической функции распределения от теоретической:

$$D_N = \sup_x |F_N^*(x) - F(x)|,$$

о котором мы можем рассуждать лишь умозрительно. Оба отклонения совпадают при выполнении гипотезы H_0.

Основные утверждения, приводящие к критерию Колмогорова, формулируются следующим образом:

Теорема 1. Пусть $X_1, \ldots, X_N$ — выборка, имеющая непрерывное распределение F. Тогда случайная величина $\sqrt{N}D_N$ имеет "универсальное"распределение K_N, не зависящее от F.

Теорема 2. При $N \to \infty$ распределения K_N слабо сходятся к предельному распределению K.

Предельное распределение K называется распределением Колмогорова. Оно выступает в качестве асимптотического шаблона в описываемой ниже статистической процедуре.

Фактически, теоремы 1 и 2 уточняют для выборок с непрерывным распределением сформулированную в параграфе 1.4 теорему Гливенко-Кантелли. В частности, из них следует, что величина D_N сходится к нулю со скоростью $1/\sqrt{N}$.

Теорема 1 будет доказана в конце параграфа. Теорему 2 можно рассматривать как утверждение о предельном поведении конкретной последовательности распределений. Доказательство ее, довольно

сложное технически, мало что дает пользователям. Мы не будем его приводить[1].

Опишем процедуру критерия Колмогорова, опирающуюся на теоремы 1 и 2. Для начала заметим, что "малые"значения величины D_N^0 свидетельствуют о хорошем согласии эмпирических данных с гипотетическим распределением. По универсальному распределению K_N выберем табличное $z = z_\varepsilon$, зависящее от уровня значимости ε, такое, что $K_N(z_\varepsilon) = 1 - \varepsilon$. Тогда

$$\mathsf{P}_0(\sqrt{N}D_N > z_\varepsilon) = \mathsf{P}_0(\sqrt{N}D_N^0 > z_\varepsilon) = \varepsilon.$$

Это соотношение приводит к тесту, имеющему уровень значимости ε: основная гипотеза H_0 отвергается, если

$$\sqrt{N}D_N^0 > z_\varepsilon.$$

Чаще всего ограничиваются более доступным с точки зрения наличия таблиц упрощенным асимптотическим вариантом, когда z_ε находится по таблице предельного распределения K: $K(z_\varepsilon) = 1 - \varepsilon$. Тогда критерий имеет уровень значимости ε асимптотически.

Предостережение. У некоторых пользователей может возникнуть желание аналогичным образом проверять и сложные параметрические гипотезы, оценивая предварительно параметры *по тем же эмпирическим данным*. При использовании критерия Колмогорова это **недопустимо.** На примере критерия хи-квадрат мы уже видели, что подобные манипуляции меняют предельное распределение. Там это изменение сводилось к уменьшению числа степеней свободы, но не выводило из семейства распределений хи-квадрат. В случае критерия Колмогорова предельное распределение меняется более сложным образом. Проконтролировать это изменение весьма трудно.

Перейдем теперь к доказательству теоремы 1, которое представляется весьма поучительным, поскольку объясняет само существование непараметрических критериев.

В основе доказательства лежит простое утверждение, сводящее, в некотором смысле, рассуждения с произвольным непрерывным распределением к аналогичным рассуждениям с равномерным распределением.

[1]Отметим впрочем, что имеются красивые связи этого утверждения с теорией слабой сходимости мер в функциональных пространствах.

Лемма. Пусть X — случайная величина с непрерывной функцией распределения F. Тогда случайная величина $F(X)$ равномерно распределена на $\langle 0,1\rangle$.

Доказательство леммы. Предположим сначала, что $F(x)$ строго возрастает в области $\{x : 0 < F(x) < 1\}$. Тогда существует обратная функция F^{-1}, определенная на $\langle 0,1\rangle$. С ее помощью получаем ($0 < y < 1$):

$$\mathsf{P}(F(X) < y) = \mathsf{P}(X < F^{-1}(y)) = F(F^{-1}(y)) = y.$$

Мы получили функцию распределения равномерного закона на $\langle 0,1\rangle$ и лемма доказана. Если же F не является строго монотонной, следует рассмотреть "обобщенную обратную"функцию $F^{-1}_{\text{обобщ.}}$, такую, что

$$F(F^{-1}_{\text{обобщ.}}(y)) = y,\ y \in]0,1[,$$

и повторить с ней то же рассуждение. Мы опустим технические подробности конструкции такой обобщенной обратной функции.

Доказательство теоремы 1. Положим $Y_i = F(X_i)$, $i = 1,\dots,N$. Тогда $Y_1,\dots,Y_N$ — выборка, имеющая равномерное на $\langle 0,1\rangle$ распределение. Пусть $G^*_N(y)$ — ее эмпирическая функция распределения. Докажем, что (с вероятностью единица)

$$F^*_N(x) = G^*_N(F(x)). \tag{4.8}$$

Действительно,

$$X_i < x \Longleftrightarrow Y_i < F(x). \tag{4.9}$$

Поэтому количество наблюдений, меньших x, в первой выборке совпадает с количеством наблюдений, меньших $F(x)$, во второй выборке. Для читателей, стремящихся к полной точности, отметим, что если F не является строго возрастающей, то события из формулы (4.9) могут не совпадать, однако в любом случае "неразличимы— отличаются на событие нулевой вероятности.

Соотношение (4.8) позволяет сделать вывод, что отклонения

$$D_N = \sup_x |F^*_N(x) - F(x)|$$

и

$$D^Y_N = \sup_y |G^*_N(y) - y|$$

совпадают (с вероятностью 1). Действительно, замена переменной $y = F(x)$ переводит D_N в D^Y_N.

Окончательно получаем, что распределение величины D_N (а также, разумеется, и $\sqrt{N}D_N$) для произвольной выборки, имеющей непрерывное распределение F, совпадает с распределением аналогичной величины для равномерно распределенной выборки. В нашем рассуждении равномерно распределенная выборка строилась специальным образом, поэтому мы получили более сильное утверждение — совпадение самих величин D_N и D_N^Y, а не совпадение их распределений. Теорема 1 доказана.

Отметим, что явный вид распределения K_N нам не потребовался. Это распределение некоторым не очень простым, но вполне определенным образом конструируется из равномерного (см. [1]). Для предельного распределения K, которое затабулировано во всех справочниках, явный вид также известен:

$$K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2},\ z > 0,$$

хотя требуется лишь в очень редких случаях.

4.10 Другие непараметрические критерии

В заключительном параграфе этой главы мы перечислим еще несколько непараметрических критериев. В основном, это делается для того, чтобы критерий Колмогорова не представлялся чем-то исключительным.

Критерий омега-квадрат Мизеса-Смирнова для простой гипотезы $H_0 = \{F = F^0\}$.

Этот тест несколько напоминает колмогоровский, он основан на другом функционале, измеряющем расхождение эмпирического и теоретического распределений

$$\omega_N^2 = N \int_{-\infty}^{\infty} [F_N^*(x) - F(x)]^2 dF(x). \tag{4.10}$$

Аналогично теореме 9.1 проверяется, что распределение величины ω_N^2 универсально (в классе непрерывных F). Доказано, что при $N \to \infty$ эти универсальные распределения слабо сходятся к некоторому предельному распределению Ω. Тест омега-квадрат использует это распределение Ω, как асимптотический шаблон: гипотеза H_0 отвергается, если $\omega_N^{2,0} > z_\varepsilon$. Здесь z_ε находится по таблицам распределения Ω, а $\omega_N^{2,0}$ — величина, аналогичная (4.10), которая получается заменой F на F^0.

Критерий Колмогорова-Смирнова для проверки однородности.

Рассматривается задача сравнения двух теоретических распределений (ср. с критерием знаков в параграфе 3.5). Имеются две независимые между собой выборки: $X_1, \ldots, X_N$ с непрерывным теоретическим распределением F и $Y_1, \ldots, Y_{N'}$ с непрерывным теоретическим распределением G. Основная (очевидно, сложная) гипотеза имеет вид $H_0 = \{F = G\}$. Для ее проверки предлагается использовать расхождение между двумя эмпирическими функциями распределения:

$$D_{N,N'} = \sup_x |F_N^*(x) - G_{N'}^*(x)|$$

или

$$D_{N,N'}^+ = \sup_x [F_N^*(x) - G_{N'}^*(x)].$$

Аналогично теореме 9.1 доказывается, что, в предположении справедливости H_0, распределение величины $D_{N,N'}$ (а также $D_{N,N'}^+$) универсально. Кроме того, можно доказать, что при $N, N' \to \infty$ универсальные распределения величин

$$\sqrt{\frac{NN'}{N+N'}} D_{N,N'}$$

сходятся к распределению Колмогорова K. На этих результатах основан тест Колмогорова-Смирнова, процедура которого вполне аналогична процедурам тестов Колмогорова и омега-квадрат. Она использует K в качестве асимптотического шаблона. Можно также доказать, что универсальные распределения величин

$$\sqrt{\frac{NN'}{N+N'}} D_{N,N'}^+$$

слабо сходятся к распределению с плотностью

$$p(z) = 4ze^{-2z^2},\ z > 0.$$

Этот результат дает еще один тест проверки однородности.

В учебниках по математической статистике ([8], [1], [2] и др.) можно найти много других критериев (как параметрических, так и непараметрических).

Глава 5

Эконометрика и статистика

Читатель, впервые открывающий учебник по эконометрике, видимо, прежде всего замечает обилие статистической терминологии. Здесь и параметры, которые надлежит оценивать, и гипотезы, которые следует проверять, и доверительные интервалы, и корреляция, и прочее, и прочее, и прочее .... Лишь постепенно ему становится ясно, что эконометрика — это нечто бóльшее, чем приложения статистических методов к экономическим задачам (хотя и это также имеет место).

По-настоящему прочувствовать отличие от статистики можно лишь в процессе изучения эконометрики. В этой главе мы только намечаем некоторые узловые моменты, отсылая интересующихся читателей к другим источникам.

Довольно нестандартное, хотя отчасти субъективное, описание предмета эконометрики можно найти в [9].

В книге [22] дано детально структурированное описание эконометрического исследования.

В [15] традиционная эконометрическая методология сравнивается с современными подходами. Приведены интересные цитаты, выражающие точки зрения известных специалистов.

Большой интерес представляет книга [21], написанная одним из крупнейших современных эконометристов.

5.1 Специфика моделей и эмпирических данных в экономике

Каждое эконометрическое исследование проводится в рамках некоторой модели — умозрительной конструкции, выделяющей главные, существенные стороны интересующего исследователя фрагмента

окружающего экономического мира и отбрасывающей те, которые представляются незначимыми. В процессе исследования модель может претерпевать определенные изменения. Взаимоотношения модели и моделируемого явления могут быть довольно деликатными, и часто именно в них кроется успех (или неудача) исследования. Язык описания модели чаще всего — математика.

Экономическая наука, как одна из наук о человеческом обществе, обладает рядом особенностей, отличающих ее от многих других областей применения математических методов (в частности, от физики, где такие методы развиты в наибольшей степени).

Прежде всего следует отметить, что в экономических исследованиях практически нет места активному эксперименту. Если, скажем, физик-экспериментатор сам создает условия для проведения опыта — готовит аппаратуру, приводит в нужное состояние изучаемую субстанцию и т.д., а физик-теоретик старается объяснить или предсказать результат такого целенаправленного эксперимента, то экономист-исследователь на первом этапе лишь наблюдает за ходом событий и фиксирует происходящее. Последующие задачи, конечно, будут, как и в любой другой науке, стандартными — объяснить и предсказать.

Далее, человек, как существо, обладающее сознанием, способен в той или иной степени влиять на общественные процессы (неважно, опираясь на экономическую теорию, вопреки ей или же вне связи с ней). Некоторые стороны подобного влияния можно условно обозначить как "политические"факторы — бóльшая часть экономических и эконометрических моделей рассматривает их как заданные извне — экзогенно. В других ситуациях возникают так называемые коллективные эффекты (термин часто используется и в физике). Первый и наиболее известный пример такого эффекта в экономической сфере — "теорема о невидимой руке"Адама Смита.

Коллективные эффекты постоянно в той или иной форме проявляются в эконометрике. Обычно это выражается в присутствии стохастических характеристик (подробнее см. ниже). Заметим, впрочем, что это далеко не единственная причина их появления. Здесь следует отметить одну важную особенность. Статистические методы, развивавшиеся в течение многих десятилетий, начиная со второй половины XIX века (кинетическая теория газов Людвига Больцмана), были ориентированы на использование именно в физике, где масштабы "коллективности"явлений выражаются огромными числами — из

школьного курса физики известно так называемое число Авогадро: $6 \cdot 10^{23}$ молекул в одном моле вещества. Соответственно, и физические закономерности выполняются с большой точностью. Мы не можем, например, допустить, что весь воздух в комнате вдруг соберется в одной ее половине, хотя теоретические шансы и имеются (пример, конечно, сильно утрированный).

Напротив, коллективные эффекты в экономической области связаны с совсем другими числами, в том числе и весьма скромными. Так, число фирм, работающих на рынке, может исчисляться тысячами, сотнями или быть еще меньше. Число покупателей, принимаемых во внимание в рассматриваемой модели, редко будет превышать несколько миллионов (а миллион — это всего лишь 10^6).

Поэтому экономические соотношения, особенно в микроэкономических моделях, выполняются весьма приблизительно, часто даже лучше сказать — в тенденции (скорее качественно, чем количественно). В макроэкономике также количество доступных наблюдений может исчисляться десятками — как при изучении двадцатилетнего интервала между двумя мировыми войнами. Сами модели, использующиеся в эконометрических исследованиях, вынужденно (из соображений целесообразности) являются простыми, обычно линейными (см. ниже). Только в редких случаях, как в теории финансовых временных рядов, где исследователю могут оказаться доступными миллионы данных, имеет смысл конструировать более замысловатые и утонченные модели. Сами статистические методы во многих аспектах приходится переосмысливать и даже менять при переходе от физики к новым областям исследования.

5.2 Начальное описание предмета эконометрики и ее задач

Эконометрика есть ветвь экономической науки, связанная с количественным оцениванием и проверкой экономических закономерностей. Эконометрическое исследование основывается на экономической теории и на фактах, относящихся к событиям, имевшим место в реальном экономическом мире.

Экономическая теория дает исследователю модель интересующих его явлений. Эту экономическую модель эконометрист приспосабливает

к своим методам, трансформирует в эконометрическую. Основные эконометрические модели имеют алгебраический характер, т.е. представляются в виде совокупности уравнений, связывающих принимаемые во внимание характеристики и включающих неопределенные ("свободные") параметры, которые оцениваются на основе эмпирических данных. Эмпирические данные представляют собой количественно выраженные факты, относящиеся к изучаемой задаче. Как правило, предварительно они подвергаются различным процедурам проверки и уточнения, которых мы здесь не касаемся.

Большинство моделей рассматривает относительно замкнутый фрагмент экономического мира, взаимоотношения которого с остальной частью этого мира удается описать при помощи небольшого числа связей (экзогенных величин).

Важной особенностью эконометрических моделей является их стохастический характер — некоторые экономические показатели трактуются как случайные величины. Можно выделить два источника этой случайности (хотя отделить их друг от друга и не всегда удается). Некоторые показатели принято считать случайными по концептуальным причинам (можно сказать, генетически). Другие описываются как случайные вынужденно — ввиду неполноты модели и наличия неучтенных факторов, создающих так называемые стохастические ошибки.

Рассматриваемые ниже модели в большинстве своем являются линейными в двух отношениях. Во-первых, по параметрам, т.е. эти параметры входят в уравнения модели линейно, как коэффициенты в отдельных слагаемых. Во-вторых, по стохастическим ошибкам (см. ниже) — они включаются в уравнения аддитивно, как дополнительные слагаемые, описывающие флуктуации вокруг некоторых "главных", например, средних, значений. К линейным моделям иногда удается сводить и некоторые другие.

Для оценивания параметров модели, проверки гипотез о них, выявления ошибок спецификации и решения прочих сопутствующих вопросов используется эконометрическая техника, включающая в себя различные методы и приемы математической и прикладной статистики, во многих случаях специально приспособленные для этих целей.

Оцененная эконометрическая модель может использоваться как для структурного анализа, включая обратное влияние на экономическую теорию, так и для прогнозирования и связанной с ним выработки

экономической политики.

Основные величины, входящие в уравнения модели, подразделяются на внутренние (эндогенные) и внешние (экзогенные). Внутренние величины совместно определяются моделью; можно сказать, что в некотором смысле модель объясняет их. Напротив, экзогенные величины, хотя и входят в модель существенным образом (см. выше), определяются отдельными механизмами вне ее рамок и выступают, в зависимости от ситуации, как объясняющие величины, управляющие величины, начальные или граничные условия и т.д., и т.п. Особую, в определенной степени промежуточную, роль играют лаговые значения внутренних величин, см. пример 2 ниже.

Стохастические слагаемые, входящие в уравнения линейной модели, отличаются от основных величин прежде всего тем, что они принципиально не наблюдаемы (заметим, что основные величины также могут быть случайными; эндогенные — практически всегда). Часто их называют ошибками (errors) или возмущениями. Подобные члены обычно включаются во все уравнения модели, кроме условий равновесия и тождеств (тождества часто можно еще трактовать как определения). Присутствие стохастических ошибок в уравнениях мотивируется комплексом причин — влиянием неучтенных факторов, непредсказуемостью человеческих реакций, неточностями наблюдений и измерений и т.д.

Приведем несколько учебных примеров (подобные примеры в разных вариантах присутствуют практически во всех учебниках). В отличие от реальных эконометрических моделей, которые могут включать значительное (иногда десятки и сотни) число уравнений и величин, упрощенные учебные примеры (часто они называются моделями-прототипами) включают минимальное число уравнений — для понимания основных принципов эконометрического исследования этого достаточно. С точки зрения эконометрической техники значительная часть проблем отчетливо проявляется уже для модели, включающей одно единственное уравнение. Часто таким уравнением оказывается уравнение линейной (множественной) регрессии, которое подробно обсуждается дальше.

Подчеркнем важное обстоятельство, связанное с формированием эконометрической модели. Не любой фрагмент экономического мира поддается подобному моделированию. Набор интересующих исследователя величин, которые он надеется описать внутренним

образом, должен оказаться в некотором смысле полным. Нельзя, скажем, разделить спрос и предложение в примере 1 ниже. Если модель сконструирована неудачно, известные методы исследования могут оказаться неприменимыми, а сделанные с их помощью выводы — ошибочными. К этой проблеме мы будем неоднократно возвращаться.

Пример 1. Микроэкономическая модель-прототип спроса и предложения.

Будем представлять себе, что речь идет о производстве некоторого сельскохозяйственного продукта. Такое производство во многих случаях обладает естественной цикличностью. Мы предположим, что в пределах одного цикла устанавливается равновесие между спросом и предложением и формируется равновесная цена. Поэтому модель будет иметь статический характер, а время явным образом не появится.

Запишем уравнение спроса, уравнение предложения и условие равновесия в виде

$$\begin{aligned} q^D &= \beta_1 + \beta_2 p + \gamma_1 I + \varepsilon^D, \\ q^S &= \beta_3 + \beta_4 p + \gamma_2 r + \varepsilon^S, \\ q^D &= q^S. \end{aligned}$$

Здесь q^D — количество (quantity) продукта, выражающее спрос (Demand), q^S — количество продукта, выражающее предложение (Supply), p — цена (price), I — доход (Income), r — количество осадков (rainfall). Слагаемые ε^D и ε^S — стохастические ошибки, соответствующие необъясняемым нашими уравнениями частям спроса и предложения. Условие равновесия не содержит стохастической ошибки.

Нетрудно догадаться, что внутренними величинами в модели примера 1 являются цена p и количество продукта $q = q^D = q^S$, в то время как доход I и осадки r целесообразно трактовать внешним, экзогенным, образом.

Нет нужды подробно останавливаться на слабых местах выбранного модельного представления — каждый может сделать это самостоятельно. Подчеркнем однако, что при всей своей простоте модель выражает (если угодно, в карикатурной форме) некоторые теоретические представления: доход входит именно в уравнение спроса, а осадки, влияющие на урожай, — в уравнение предложения. Подобные системы уравнений называются структурными.

Пример 2. Макроэкономическая модель-прототип определения национального дохода.

Эта модель задается уравнениями

$$\begin{aligned} C_t &= \beta_1 + \beta_2 Y_t + \varepsilon_t^C, \\ I_t &= \beta_3 + \beta_4 Y_t + \gamma_1 Y_{t-1} + \varepsilon_t^I, \\ Y_t &= C_t + I_t + G_t. \end{aligned}$$

Здесь внутренними являются величины C_t, I_t, Y_t, описывающие, соответственно, потребление (Consumption), инвестиции (Investment) и доход (Yield) в году t, а внешней — G_t — правительственные расходы (Government spending). Запаздывающее (лаговое, lagged) значение Y_{t-1} национального дохода вместе с G_t составляет набор предопределенных (predetermined) величин. Последнее уравнение является тождеством и не содержит стохастического слагаемого.

Отметим, что пример 2, в отличие от примера 1, имеет отчетливо выраженный динамический характер. При решении этой структурной системы уравнений помимо "граничных"("сопровождающих") условий, определяемых правительственными расходами G_t, скорее всего, появится еще и "начальное"условие (скажем, Y_0, если время t изменяется, начиная с 1).

Приведенные выше описания моделей в примерах 1 и 2 являются неполными. Следует еще уточнить предположения о характере стохастических слагаемых ε. Анализ и проверка этих предположений — важная часть эконометрического исследования. Подобных вопросов мы будем неоднократно касаться в последующих главах.

5.3 Несколько комментариев к последующим главам

Наши обсуждения приблизились к той точке, когда нужно покинуть (относительно) гладкую равнину статистики повторных выборок и перейти к задачам более сложного характера. В некоторых местах предыдущих глав мы намеренно упоминали об этом, а при возможности и подгоняли формулировки и/или доказательства под возможные обобщения. Примеры предыдущего параграфа дают первый толчок к этим обобщениям. Первое из них, довольно безобидное, — переход к разнораспределенным наблюдениям, но с очень специальной формой этой разной распределенности, — будет обсуждаться в главе 6.

Даже это минимальное изменение приводит к другой расстановке акцентов. Так, обсуждение асимптотических свойств, начиная с состоятельности, отходит на второй план. Действительно,

любая форма неоднородности должна экстраполироваться на "дополнительные"наблюдения, появляющиеся при увеличении объема выборки. Удобно вводить соответствующие усложненные модели постепенно. В главе 6 асимптотический подход практически даже не упоминается.

Более серьезные обобщения излагаются в дальнейших главах. Они включают различные варианты неоднородности наблюдений, корреляцию между ними и другие обстоятельства, учет которых становится существенным при построении моделей с конкретной экономической интерпретацией. Мы будем изредка упоминать о таких интерпретациях. Конкретика обычно помогает прояснить содержательный смысл формальных конструкций.

Обобщения, о которых пойдет речь, возникают по содержательным причинам (некоторые из этих причин также будут обсуждаться). Поскольку используемые в обобщенных моделях приемы в ряде случаев оказываются более сложными, а иногда принципиально иными (даже несовместимыми с ранее рассмотренными), непременно возникает задача выбора разумной спецификации модели (мы впервые столкнемся с подобной проблемой в параграфе 6.12).

В конечном счете эконометрическое исследование включает целый комплекс задач, а статистические рецепты составляют далеко не единственную, хотя и важную, часть их решения.

Глава 6

Линейная регрессионная модель

6.1 Спецификация модели. Соглашения об обозначениях и терминологии

Спецификацией модели называют ее концептуальную функциональную форму. В этой главе будет рассматриваться модель, имеющая спецификацию

$$Y = \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon. \tag{6.1}$$

В уравнении (6.1) Y — объясняемая величина, $X_1, \ldots, X_k$ — объясняющие величины, или регрессоры, ε — стохастическая ошибка. Коэффициенты $\beta_1, \ldots, \beta_k$ — неопределенные (свободные) параметры, подлежащие оцениванию.

Спецификация (6.1) подразумевает некоторую теоретическую концепцию — мы считаем, что существуют "истинные"значения коэффициентов $\beta_{1,true}, \ldots, \beta_{k,true}$, но они неизвестны и могут обсуждаться лишь умозрительно. (Конечно, это замечание относится к любой задаче оценивания, однако в литературе по статистике этот нюанс редко упоминается.) Следуя установившейся традиции, мы в дальнейшем изложении будем часто использовать обозначение β и для "истинных"коэффициентов.

С практической точки зрения исследователь располагает данными N совместных наблюдений величин $Y, X_1, \ldots, X_k$, так что для i-го наблюдения $(i = 1, \ldots, N)$ может представлять себе соотношение

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{6.2}$$

(представление данных), вытекающее из спецификации модели. Подчеркнем, что первый индекс из двух в нашей системе обозначений

всегда — номер наблюдения. Если же индекс всего один, то он обозначает номер наблюдения у Y и ε, но номер регрессора у X.

Отличие формул (6.1) и (6.2) в том, что спецификация (6.1) может обсуждаться вне всякой связи с эмпирическими данными, т.е. концептуально, при этом $Y, X_1, \ldots, X_k, \varepsilon$ оказываются обозначениями для типов объектов. Напротив, $Y_i, X_{ij}, \varepsilon_i$ в формуле (6.2) понимаются как величины, отвечающие i-му наблюдению, т.е. как конкретные объекты, а не типы объектов. С точки зрения пользователя Y_i и X_{ij} можно также трактовать как числа — "реализовавшиеся"значения соответствующих величин. Для ε_i такого утилитарного понимания быть не может — коэффициенты модели свободны, т.е. неизвестны исследователю, а потому и ошибка ненаблюдаема.

Удобно использовать также сокращенные векторно-матричные обозначения. При этом значения Y_i объединяются в вектор-столбец Y размерности N; аналогично, значения X_{ij} объединяются в матрицу X, имеющую N строк и k столбцов, а ε_i — в вектор-столбец ε. Столбцы матрицы X удобно обозначать $X_1, \ldots, X_k$ — они состоят из значений соответствующих регрессоров. В этих обозначениях формула (6.1) приобретает второй смысл — смысл соотношения между N-мерными векторами $Y, X_1, \ldots, X_k$ и ε. Полностью сокращенную его запись

$$Y = X\beta + \varepsilon \tag{6.3}$$

мы получим, если введем еще и вектор-столбец β коэффициентов. Размерность вектора β, очевидно, равна k.

Заготовим сразу же еще одно соглашение об обозначениях. Среднее арифметическое компонент некоторого вектора (неважно, случайного или нет) будет обозначаться традиционной для статистики чертой сверху, например,

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \bar{X}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij},$$

а отклонения от этого среднего значения — соответствующей малой буквой:

$$y_i = Y_i - \bar{Y}, x_{ij} = X_{ij} - \bar{X}_j$$

и т. д. Аналогичные отклонения для вектора ошибок будут записываться подробно: $\varepsilon_i - \bar{\varepsilon}$.

Используя обозначение $d^{\rightarrow}$ для вектора, все компоненты которого равны d, можно записать отклонения в векторной форме

$$y = Y - \bar{Y}^{\rightarrow}, \quad x_j = X_j - \bar{X}_j^{\rightarrow}, \quad \varepsilon - (\bar{\varepsilon})^{\rightarrow}$$

6.2 Классическая линейная модель — обсуждение предположений

В этом параграфе мы дополняем спецификацию (6.1) простейшими предположениями о регрессорах и ошибках и получаем полное описание так называемого классического варианта линейной регрессионной модели.

Предположения о регрессорах включают два разноплановых свойства. Во–первых, регрессоры предполагаются неслучайными. Примерами таких регрессоров являются:

1. Константа; этот регрессор обычно включается в модель под первым номером: $X_1 = 1^{\rightarrow}$ (константу, отличную от единицы, можно включить множителем в соответствующий коэффициент β_1).

2. "Время": $X_{i2} = i$.

3. Любая "управляющая", т. е. подконтрольная исследователю величина.

С точки зрения экономической теории неслучайность регрессоров (особенно всех!) не очень частое явление, так что сделанное предположение довольно ограничительно. В дальнейшем (глава 7) мы будем обсуждать обобщения классической модели, в которых это предположение заменяется более реалистичными.

Второе предположение о регрессорах имеет прозаический характер: столбцы $X_1, \ldots, X_k$ регрессионной матрицы X предполагаются линейно независимыми векторами. Это свойство означает, что нельзя уменьшить количество регрессоров, выразив некоторые из них (хотя бы один) через остальные.

Предположение о линейной независимости столбцов регрессоров может выполняться лишь в случае, когда число наблюдений N не меньше числа регрессоров. Это вполне укладывается в обычные статистические рамки — оценить много параметров по малому числу наблюдений почти

никогда не удается осмысленным образом. Конечно, желательно, чтобы N было значительно больше k.

Перейдем теперь к предположениям об ошибках. В классической модели они формулируются наиболее жестким и не всегда реалистичным образом:

- предполагается, что ошибки ε_i $(i = 1, \ldots, N)$ образуют так называемый слабый белый шум — последовательность центрированных ($\mathbf{E}\varepsilon_i = 0$) и некоррелированных ($\mathbf{E}(\varepsilon_{i_1}\varepsilon_{i_2}) = 0$ при $i_1 \neq i_2$) случайных величин с одинаковыми дисперсиями $\mathbf{E}(\varepsilon_i^2) = \sigma^2$.

Свойство центрированности практически не является ограничением, т. к. при наличии постоянного регрессора среднее значение ошибки можно было бы включить в соответствующий коэффициент ($\beta_1 + \varepsilon = \beta_1 + \mathbf{E}\varepsilon + (\varepsilon - \mathbf{E}\varepsilon)$).

Обобщения классической модели, включающие автокорреляцию ошибок и/или неоднородность дисперсий, будут рассмотрены дальше (глава 7).

В ряде случаев сделанные предположения об ошибках будут дополняться свойством нормальности (гауссовости) — случайный вектор ε имеет нормальное распределение (гауссовский белый шум). Такую модель мы будем называть классической моделью с нормально распределенными ошибками. Как хорошо известно, многомерное нормальное распределение задается своим вектором математических ожиданий (в нашем случае это нулевой вектор) и матрицей ковариаций — здесь она имеет вид $\sigma^2\mathbf{1}$, где $\mathbf{1}$ — единичная матрица. Если компоненты нормально распределенного вектора некоррелированы, они автоматически оказываются независимыми, так что в классической модели с нормально распределенными ошибками эти ошибки образуют последовательность независимых одинаково нормально распределенных случайных величин $\mathbf{N}(0, \sigma^2)$.

Отметим еще одну тонкость, относящуюся к определению многомерного нормального распределения — если каждая из величин ε_i нормально распределена, то вектор ε, из них составленный, не обязан быть нормально распределенным (даже если величины ε_i не коррелируют!). К сожалению, в литературе иногда встречаются неаккуратные формулировки, игнорирующие эту тонкость.

6.3 Оценивание коэффициентов регрессии — метод наименьших квадратов

Классическая модель линейной регрессии имеет своими параметрами $\beta_1, \ldots, \beta_k$ и σ. Подчеркнем, что все они, включая σ, входят в модель линейно (параметр σ можно было бы явным образом выделить, записывая ошибку ε в виде $\sigma \cdot (\varepsilon/\sigma)$ и учитывая, что случайная величина ε/σ стандартизована — имеет нулевое математическое ожидание и единичную дисперсию). Отметим, впрочем, что из наших "слабых"предположений не следует, что величины ошибок ε_i одинаково распределены — это предполагается лишь на уровне второго порядка, а информация о моментах более высоких порядков отсутствует.

В этом параграфе мы рассматриваем первый этап процедуры оценивания — построение оценок коэффициентов регрессии $\beta_1, \ldots, \beta_k$ методом наименьших квадратов (МНК; английская аббревиатура OLS — ordinary least squares). Идею этого метода, предложенного К.Гауссом в начале XIX века, удобнее всего излагать геометрически — на языке векторов N-мерного пространства. В ходе этого обсуждения коэффициенты $\beta_1, \ldots, \beta_k$ будут трактоваться как свободно меняющиеся параметры. "Истинные"их значения $\beta_{1,true}, \ldots, \beta_{k,true}$ в ходе рассуждений явно появляться почти не будут.

Итак, в нашем распоряжении имеются векторы значений регрессоров $X_1, \ldots, X_k$ и вектор значений объясняемой величины Y. Мы стремимся найти такую линейную комбинацию $X\beta = \beta_1 X_1 + \cdots + \beta_k X_k$ регрессоров, которая "лучше всего"объясняла бы Y, т.е. "с наименьшим отклонением". Естественнее всего представляется измерять отклонение $Y - X\beta$ длиной соответствующего вектора и подбирать коэффициенты β так, чтобы эта длина (или, что равносильно, ее квадрат) была минимальна. Квадрат длины отклонения $Y - X\beta$ равен

$$(Y - X\beta)'(Y - X\beta) = \sum_{i=1}^{N} (Y_i - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2, \qquad (6.4)$$

так что предложение Гаусса сводится к поиску точки минимума $\hat{\beta}$ этой квадратичной функции коэффициентов и объявлению ее оценкой вектора "истинных"коэффициентов β_{true}.

Хотя возможны и другие меры отклонения, например, сумма модулей вместо суммы квадратов, однако они не получили широкого

распространения. Отчасти это связано с наличием у суммы квадратов ряда удобных свойств (см. ниже), а отчасти, по-видимому, с тем, что мы привыкли к евклидову способу измерения расстояний, и он нам кажется самым естественным. Определенную роль играют и установившиеся традиции.

Для нахождения точки минимума $\hat{\beta}$ мы снова воспользуемся геометрическими рассуждениями. Рассмотрим в N-мерном пространстве $\mathbb{R}^N$ взаимное положение вектора Y и подпространства $\mathcal{L}(X_1, \ldots, X_k)$, порожденного векторами $X_1, \ldots, X_k$ регрессоров (его размерность, очевидно, равна k). Пусть $\hat{Y}$ — ортогональная проекция вектора Y на подпространство $\mathcal{L}(X_1, \ldots, X_k)$. Тогда вектор-разность $Y - \hat{Y}$ перпендикулярен этому подпространству. Если $X\beta = \beta_1 X_1 + \cdots + \beta_k X_k$ — какая-то другая точка подпространства $\mathcal{L}(X_1, \ldots, X_k)$, то разность $Y - X\beta$ можно трактовать как наклонную, в то время как $Y - \hat{Y}$ — перпендикуляр. Так как перпендикуляр короче наклонной, получаем

$$(Y - \hat{Y})'(Y - \hat{Y}) < (Y - X\beta)'(Y - X\beta).$$

Поэтому $\hat{Y}$ доставляет минимум сумме квадратов (6.4).

Поскольку векторы регрессоров $X_1, \ldots, X_k$ линейно независимы, проекция $\hat{Y}$ единственным образом разлагается в линейную комбинацию их:

$$\hat{Y} = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k = X\hat{\beta}.$$

Вектор $\hat{\beta}$ коэффициентов — искомый.

От геометрической интерпретации точки минимума перейдем к соответствующим формулам. Запишем условие ортогональности

$$Y - \hat{Y} \perp \mathcal{L}(X_1, \ldots, X_k)$$

в виде

$$(X\beta)'(Y - X\hat{\beta}) = 0. \tag{6.5}$$

Здесь $X\beta$ — произвольный вектор пространства $\mathcal{L}(X_1, \ldots, X_k)$.

Перепишем теперь равенство (6.5) в виде

$$\beta' \cdot X'(Y - X\hat{\beta}) = 0$$

и заметим, что геометрически оно может быть истолковано как еще одно условие ортогональности

$$\beta \perp X'(Y - X\hat{\beta})$$

(теперь уже для векторов k-мерного пространства $\mathbb{R}^k$). Таким образом, k-мерный вектор $X^{'}(Y - X\hat{\beta})$ ортогонален произвольному вектору β пространства $\mathbb{R}^k$. Отсюда следует (даже равносильно), что он нулевой:

$$X^{'}(Y - X\hat{\beta}) = 0.$$

Записывая это равенство в виде

$$X^{'}X\hat{\beta} = X^{'}Y, \tag{6.6}$$

получаем для $\hat{\beta}$ так называемое нормальное уравнение МНК. Легко сообразить, что оно имеет единственное решение. Действительно, по предположению, ранг матрицы X равен k. Из свойств ранга матрицы следует, что тогда и ранг $X^{'}X$ равен k. Поскольку $X^{'}X$ — квадратная матрица порядка k, заключаем, что она обратима.

Окончательно, получаем выражение для оценок метода наименьших квадратов

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y. \tag{6.7}$$

Важно подчеркнуть, что вектор оценок $\hat{\beta}$ получается линейным преобразованием случайного вектора Y.

Образованный с помощью этих оценок вектор $\hat{Y} = X\hat{\beta}$ можно назвать вектором прогнозных (предсказываемых моделью) значений величины Y (английский термин — predicted values или fitted values).

Обозначим через P оператор ортогонального проектирования на подпространство регрессоров $\mathcal{L}(X_1, \ldots, X_k)$ (и соответствующую матрицу). Из формулы (6.7) следует, что

$$P = X(X^{'}X)^{-1}X^{'}. \tag{6.8}$$

Эта матрица, а также матрица $P^{\perp} = \mathbf{1} - P$, соответствующая проектированию на подпространство $\mathcal{L}^{\perp}(X_1, \ldots, X_k)$ векторов, ортогональных регрессорам, будут часто использоваться в последующих обсуждениях. Выпишем некоторые их свойства, легко вытекающие как из геометрического смысла проекций, так и из формального определения (6.8). Проверка этих свойств оставляется читателю.

$$P = P^{'}, \qquad P^{\perp} = (P^{\perp})^{'}, \qquad \text{(симметричность)}$$

$$P = P^2, \qquad P^{\perp} = (P^{\perp})^2, \qquad \text{(идемпотентность)}$$

$$PP^{\perp} = P^{\perp}P = 0, \qquad P + P^{\perp} = \mathbf{1},$$

$$PX_j = X_j, \qquad P^{\perp}X_j = 0,$$
$$PX = X, \qquad P^{\perp}X = 0.$$

Вектор

$$\hat{\varepsilon} = Y - \hat{Y} = P^{\perp}Y$$

называется вектором остатков (residuals). Для него можно записать также другое выражение

$$\hat{\varepsilon} = P^{\perp}(X\beta + \varepsilon) = P^{\perp}\varepsilon$$

($P^{\perp}X = 0$, как указано ранее). Остатки можно интерпретировать как "оцененные ошибки". Очевидно, $P\hat{\varepsilon} = 0$.

Подставляя в формулу (6.7) спецификацию (6.3), получаем еще одну полезную формулу

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon. \tag{6.9}$$

В то время как формула (6.7) содержит лишь наблюдаемые значения и потому может использоваться для расчетов, формула (6.9) играет важную теоретическую роль (см. дальше параграф 6.5).

6.4 Частный случай — парная регрессия

Полезно выписать явно два простейших случая формулы (6.7).

Случай 1 ($k = 1$). Очевидно, имеем

$$X'X = \sum_{i=1}^{N} X_{i1}^2, \qquad X'Y = \sum_{i=1}^{N} X_{i1}Y_i,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} X_{i1}Y_i}{\sum_{i=1}^{N} X_{i1}^2} = \frac{\overline{X_1 Y}}{\overline{X_1^2}}.$$

Если дополнительно предположить, что $X_1 = 1^{\rightarrow}$ (регрессия на константу), получаем

$$\hat{\beta}_1 = \bar{Y},$$

так что прогнозные значения $\hat{Y}_i$ равны $\bar{Y}$ при всех i, что можно записать также в виде $\hat{Y} = \bar{Y}^{\rightarrow}$.

Случай 2 ($k = 2$). Аналогично предыдущему случаю получаем

$$\frac{1}{N}X'X = \begin{pmatrix} \overline{X_1^2} & \overline{X_1 X_2} \\ \overline{X_1 X_2} & \overline{X_2^2} \end{pmatrix}, \quad \frac{1}{N}X'Y = \begin{pmatrix} \overline{X_1 Y} \\ \overline{X_2 Y} \end{pmatrix},$$

$$\hat{\beta}_1 = \frac{\overline{X_2^2} \cdot \overline{X_1 Y} - \overline{X_1 X_2} \cdot \overline{X_2 Y}}{\overline{X_1^2} \cdot \overline{X_2^2} - \overline{X_1 X_2}^2},$$

$$\hat{\beta}_2 = \frac{\overline{X_1^2} \cdot \overline{X_2 Y} - \overline{X_1 X_2} \cdot \overline{X_1 Y}}{\overline{X_1^2} \cdot \overline{X_2^2} - \overline{X_1 X_2}^2}.$$

При дополнительном предположении $X_1 = 1^{\rightarrow}$ (модель парной регрессии) формулы можно несколько упростить:

$$\hat{\beta}_1 = \frac{\overline{X_2^2} \cdot \bar{Y} - \bar{X}_2 \cdot \overline{X_2 Y}}{\overline{X_2^2} - \bar{X}_2^2} = \bar{Y} - \bar{X}_2 \hat{\beta}_2,$$

$$\hat{\beta}_2 = \frac{\overline{X_2 Y} - \bar{X}_2 \bar{Y}}{\overline{X_2^2} - \bar{X}_2^2} = \frac{\overline{x_2 y}}{\overline{x_2^2}}. \tag{6.10}$$

Для вектора $\hat{Y}$ прогнозных значений из формул (6.10) получаем

$$\hat{Y} = \bar{Y}^{\rightarrow} + \frac{\overline{x_2 y}}{\overline{x_2^2}}(X_2 - \bar{X}_2^{\rightarrow}) = \bar{Y}^{\rightarrow} + \frac{\overline{x_2 y}}{\overline{x_2^2}} x_2. \tag{6.11}$$

Очевидно, $\bar{x}_2 = 0$, поэтому, усредняя (6.11), находим

$$\overline{\hat{Y}} = \bar{Y}.$$

Перенося теперь в (6.11) вектор $\bar{Y}^{\rightarrow}$ в левую часть, находим

$$\hat{y} = \frac{\overline{x_2 y}}{\overline{x_2^2}} x_2 \tag{6.11$'$}$$

— прогнозный вектор в отклонениях.

Сопоставляя между собой полученные формулы, можно обнаружить еще и такую двухступенчатую процедуру построения оценки коэффициента парной регрессии $\hat{\beta}_2$ (см. (6.10)): сначала строятся регрессии величин Y и X_2 на константу и находятся векторы остатков y и x_2. Затем строится регрессия величины y на x_2 — формула (6.11$'$). Сходная процедура для линейной модели с произвольным числом регрессоров будет обсуждаться в параграфе 6.9.

Упражнение. Показать, что регрессия с двумя произвольными регрессорами может быть получена аналогичной двухступенчатой процедурой.

6.5 Свойства оценок наименьших квадратов

В этом параграфе рассматриваются статистические свойства оценок МНК, поэтому предположение о том, что регрессоры неслучайны, будет играть важную роль (до сих пор оно не использовалось).

Первое свойство — несмещенность вектора оценок $\hat{\beta}$. Оно является, как сейчас будет видно, следствием линейности по Y. Действительно, с помощью формулы (6.9) получаем

$$\mathbf{E}\hat{\beta} = \beta + \mathbf{E}(X^{'}X)^{-1}X^{'}\varepsilon$$
$$= \beta + (X^{'}X)^{-1}X^{'}\mathbf{E}\varepsilon = \beta.$$

Здесь мы в чистом виде пользуемся линейностью — постоянные множители, в том числе и матричные, выносятся за знак математического ожидания. Сходное вычисление дает нам матрицу ковариаций вектора $\hat{\beta}$:

$$\mathsf{cov}(\hat{\beta}) = \mathbf{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^{'}] = \mathbf{E}[(X^{'}X)^{-1}X^{'}\varepsilon\cdot((X^{'}X)^{-1}X^{'}\varepsilon)^{'}]$$
$$= \mathbf{E}[(X^{'}X)^{-1}X^{'}\varepsilon\varepsilon^{'}X(X^{'}X)^{-1}] = (X^{'}X)^{-1}X^{'}\mathbf{E}(\varepsilon\varepsilon^{'})X(X^{'}X)^{-1}$$
$$= \sigma^2(X^{'}X)^{-1}X^{'}\cdot X(X^{'}X)^{-1} = \sigma^2(X^{'}X)^{-1}.$$

Нелишним будет подчеркнуть, что в матричных вычислениях порядок сомножителей должен выдерживаться (левый множитель — налево, правый — направо).

Теорема Гаусса-Маркова. Оценка $\hat{\beta}$ метода наименьших квадратов является эффективной в классе линейных несмещенных оценок.

Уточним сначала, что понимается под эффективностью векторной несмещенной оценки. Пусть $\tilde{\beta}$ — другая линейная несмещенная оценка вектора β. Тогда эффективность означает, что матрица

$$\mathsf{cov}(\tilde{\beta}) - \mathsf{cov}(\hat{\beta})$$

неотрицательно определена. Это означает, что для любого вектора $\gamma \in \mathbb{R}^k$ величина

$$\gamma^{'}[\mathsf{cov}(\tilde{\beta}) - \mathsf{cov}(\hat{\beta})]\gamma \quad (= \mathsf{V}(\gamma^{'}\tilde{\beta}) - \mathsf{V}(\gamma^{'}\hat{\beta}))$$

неотрицательна.

Доказательство теоремы. Запишем линейную оценку $\tilde{\beta}$ в виде

$$\tilde{\beta} = CY.$$

Тогда условие несмещенности $\mathbf{E}\tilde{\beta} = \beta$ записывается в виде $CX\beta = \beta$, причем последнее равенство должно выполняться тождественно по β (ведь β — это неизвестный параметр). Таким образом, матрица C должна удовлетворять условию $CX = \mathbf{1}$. Представим ее в виде

$$C = (X^{'}X)^{-1}X^{'} + D.$$

Через вспомогательную матрицу D условие несмещенности записывается как $DX = 0$. Матрица ковариаций $\mathsf{cov}(\tilde{\beta})$ выражается формулой

$$\begin{aligned}\mathsf{cov}(\tilde{\beta}) &= \mathbf{E}[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^{'}] \\ &= \mathbf{E}[C\varepsilon(C\varepsilon)^{'}] = \sigma^2 CC^{'} \\ &= \sigma^2[(X^{'}X)^{-1} + DD^{'} + (X^{'}X)^{-1}X^{'}D^{'} + D((X^{'}X)^{-1}X^{'})^{'}] \\ &= \sigma^2[(X^{'}X)^{-1} + DD^{'}].\end{aligned}$$

Здесь мы воспользовались условием несмещенности $DX = 0$. Остается проверить неотрицательную определенность матрицы $DD^{'}$:

$$\gamma^{'}DD^{'}\gamma = (D^{'}\gamma)^{'}(D^{'}\gamma) \geq 0$$

как квадрат длины вектора $D^{'}\gamma$. Теорема доказана.

Из теоремы Гаусса-Маркова вытекает, в частности, что $\mathsf{V}(\tilde{\beta}_j) \geq \mathsf{V}(\hat{\beta}_j)$, так что скалярные оценки $\hat{\beta}_j$ эффективны в аналогичном классе линейных несмещенных оценок.

Повторяя почти дословно доказательство теоремы Гаусса-Маркова, можно доказать, что для любой матрицы Γ, имеющей k строк, эффективной линейной несмещенной оценкой вектора $\Gamma\beta$ является оценка $\Gamma\hat{\beta}$. Это утверждение оставляется читателю для самостоятельной проверки.

В частности, линейные комбинации оценок МНК эффективно оценивают аналогичные линейные комбинации коэффициентов регрессии.

6.6 Оценивание дисперсии ошибок

Дисперсия σ^2 является квадратичной характеристикой ошибок — моментом второго порядка, поэтому оценивать ее, видимо, следует также квадратичным образом. При этом естественным эмпирическим

объектом, ассоциирующимся с ошибками, является вектор остатков $\hat{\varepsilon} = P^{\perp}\varepsilon$. Очевидно, $\mathbf{E}\hat{\varepsilon} = 0$. Найдем матрицу ковариаций

$$\mathsf{cov}(\hat{\varepsilon}) = \mathbf{E}[P^{\perp}\varepsilon(P^{\perp}\varepsilon)^{'}] = P^{\perp}\mathbf{E}(\varepsilon\varepsilon^{'})P^{\perp} = \sigma^2 P^{\perp}.$$

Рассмотрим теперь сумму квадратов

$$\hat{\varepsilon}^{'}\hat{\varepsilon} = \mathsf{tr}(\hat{\varepsilon}\hat{\varepsilon}^{'}).$$

Соответствующее математическое ожидание равно

$$\mathbf{E}(\hat{\varepsilon}^{'}\hat{\varepsilon}) = \mathbf{E}\mathsf{tr}(\hat{\varepsilon}\hat{\varepsilon}^{'}) = \mathsf{tr}\mathbf{E}(\hat{\varepsilon}\hat{\varepsilon}^{'}) = \sigma^2\mathsf{tr}P^{\perp}.$$

Остается вспомнить, что $P^{\perp}$ — ортогональный проектор на подпространство $\mathcal{L}^{\perp}(X_1, \ldots, X_k)$, имеющее размерность $N - k$, дополнительную к размерности подпространства регрессоров, и его след (как и любого проектора) равен этой размерности.

Альтернативное доказательство равенства $\mathsf{tr}P^{\perp} = N - k$ можно провести прямым вычислением

$$\begin{aligned} \mathsf{tr}P^{\perp} &= \mathsf{tr}[\mathbf{1}_N - X(X^{'}X)^{-1}X^{'}] = N - \mathsf{tr}[X(X^{'}X)^{-1}X^{'}] \\ &= N - \mathsf{tr}[(X^{'}X)^{-1}X^{'}X] = N - \mathsf{tr}\mathbf{1}_k = N - k \end{aligned}$$

(мы пользуемся тем, что при циклической перестановке сомножителей след произведения матриц не меняется).

Из проведенных вычислений следует, что статистика

$$s^2 = \frac{\hat{\varepsilon}^{'}\hat{\varepsilon}}{N - k} \tag{6.12}$$

является несмещенной оценкой дисперсии σ^2. Этот результат эвристически объясняется тем, что после оценивания k коэффициентов регрессии в эмпирических данных остается $N - k$ неиспользованных степеней свободы.

В модели со слабым белым шумом, оперирующей только с моментами первого и второго порядка, обсуждать эффективность оценки s^2 (в каком-либо подходящем классе) невозможно, т. к. отсутствуют предположения о старших моментах. Единственное, что остается еще получить в рамках этого подхода — это матрицу перекрестных ковариаций векторов $\hat{\beta}$ и $\hat{\varepsilon}$:

$$\begin{aligned} \mathsf{cov}(\hat{\beta}, \hat{\varepsilon}) &= \mathbf{E}((\hat{\beta} - \beta)\hat{\varepsilon}^{'}) = (X^{'}X)^{-1}X^{'}\mathbf{E}(\varepsilon\varepsilon^{'})P^{\perp} \\ &= \sigma^2(X^{'}X)^{-1}X^{'}P^{\perp} = 0 \end{aligned} \tag{6.13}$$

(опять используем равенство $P^{\perp}X = 0$ из параграфа 6.3).

Оценка s^2 позволяет оценить и матрицу ковариаций вектора $\hat{\beta}$. В выражении

$$\mathsf{cov}(\hat{\beta}) = \sigma^2(X^{'}X)^{-1}$$

надо лишь заменить σ^2 на s^2:

$$\hat{\mathsf{cov}}(\hat{\beta}) = s^2(X^{'}X)^{-1}.$$

Эта матричная оценка, очевидно, оказывается несмещенной.

6.7 Модель с нормально распределенными ошибками

Предположение о нормальности распределения вектора ошибок позволяет уточнить и усилить ряд свойств, выведенных в предыдущих параграфах. Во-первых, появляется возможность включить оценки наименьших квадратов в общую схему метода максимального правдоподобия и сравнивать их не только с линейными оценками. Во-вторых, с нормальным распределением связаны другие, хорошо известные в статистике, распределения — хи-квадрат, Стьюдента, Фишера, которые сразу начинают работать.

Начнем с обсуждения метода максимального правдоподобия. В сделанных предположениях наблюдаемый вектор Y имеет нормальное распределение $\mathbf{N}(X\beta, \sigma^2\mathbf{1})$. Соответствующая функция правдоподобия имеет вид

$$L(\beta, \sigma^2) = \prod_{i=1}^{N} \left[\frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(Y_i - (X\beta)_i)^2}{2\sigma^2}} \right]$$
$$= (2\pi)^{-N/2}\sigma^{-N} \exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)^{'}(Y - X\beta) \right].$$

Поэтому максимизировать ее по β — то же самое, что минимизировать сумму квадратов $(Y - X\beta)^{'}(Y - X\beta)$. Таким образом, оценка $\hat{\beta}$ метода наименьших квадратов оказывается одновременно и оценкой максимального правдоподобия. Далее,

$$L(\hat{\beta}, \sigma^2) = (2\pi)^{-N/2}\sigma^{-N} \exp\left[-\frac{\hat{\varepsilon}^{'}\hat{\varepsilon}}{2\sigma^2} \right].$$

Отсюда находится оценка максимального правдоподобия для σ^2:

$$\sigma^2_{ML} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{N}.$$

Как и следовало ожидать, она смещенная (см. предыдущий параграф). Ее исправление дает несмещенную оценку s^2, обсуждавшуюся выше.

С помощью многомерного неравенства Рао–Крамера можно доказать, что $\hat{\beta}$ — эффективная оценка в классе всех (не обязательно линейных) несмещенных оценок вектора β. Утверждение о том, что s^2 — эффективная несмещенная оценка дисперсии σ^2, тоже верно, но для его доказательства приходится применять более сложные методы — теорию достаточных статистик (достаточная статистика в нашей ситуации имеет вид $(Y'Y, X'Y)$). Мы не приводим деталей соответствующих рассуждений, оставляя их для самостоятельного исследования наиболее подготовленными читателями.

Перейдем теперь к свойствам оценок $\hat{\beta}$ и s^2. Прежде всего, заметим, что они независимы. Действительно, случайный вектор $(\hat{\beta}', \hat{\varepsilon}')'$ нормально распределен. Согласно формуле (6.13) подвекторы $\hat{\beta}$ и $\hat{\varepsilon}$ не коррелируют. Следовательно, они независимы. А тогда и $s^2 = \hat{\varepsilon}'\hat{\varepsilon}/(N-k)$ не зависит от $\hat{\beta}$.

Докажем теперь, что случайная величина $\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2$ распределена по хи-квадрат с $N-k$ степенями свободы. Мы уже проверяли в параграфе 6.3, что $\hat{\varepsilon} = P^{\perp}\varepsilon$. Выберем ортогональный нормированный базис $e_1, \ldots, e_{N-k}$ в подпространстве $\mathcal{L}^{\perp}(X_1, \ldots, X_k)$, где принимает значения $\hat{\varepsilon}$. Пусть e — матрица, составленная из столбцов $e_1, \ldots, e_{N-k}$. Тогда $e'\hat{\varepsilon}$ — вектор размерности $N-k$, составленный из координат вектора $\hat{\varepsilon}$ в базисе $e_1, \ldots, e_{N-k}$. Очевидно, $e'\hat{\varepsilon}$ нормально распределен и центрирован. Вычислим его матрицу ковариаций

$$\mathsf{cov}(e'\hat{\varepsilon}) = \mathbf{E}[e'\hat{\varepsilon}(e'\hat{\varepsilon})'] = e'\mathbf{E}(\hat{\varepsilon}\hat{\varepsilon}')e = \\ = \sigma^2 e' P^{\perp} e = \sigma^2 e' e = \sigma^2 \mathbf{1}_{N-k}$$

(мы воспользовались вычисленным в параграфе 6.6 значением $\mathbf{E}(\hat{\varepsilon}\hat{\varepsilon}') = \sigma^2 P^{\perp}$, а также тем, что $P^{\perp}$ действует тождественно на векторы базиса $e_1, \ldots, e_{N-k}$). Заметим теперь, что суммы квадратов $\hat{\varepsilon}'\hat{\varepsilon}$ и $(e'\hat{\varepsilon})' \cdot (e'\hat{\varepsilon})$ дают одну величину — квадрат длины вектора $\hat{\varepsilon}$. Отсюда получаем, что

$$\sigma^{-2}\hat{\varepsilon}'\hat{\varepsilon} = \sigma^{-2} \sum_{j=1}^{N-k} (e_j'\hat{\varepsilon})^2$$

имеет распределение χ^2_{N-k}. Действительно, величины $\sigma^{-1}e_j'\hat{\varepsilon}$ имеют стандартное нормальное распределение и независимы.

Теперь мы получаем возможность построения доверительных интервалов для коэффициентов регрессии β_j и совместных доверительных областей для них. Ограничимся пока описанием конструкции доверительных интервалов. Мы знаем, что

$$\hat{\beta}_j \in \mathbf{N}(\beta_j, \sigma^2[(X'X)^{-1}]_{jj}),$$

$$\frac{(N-k)s^2}{\sigma^2} \in \chi^2_{N-k},$$

и эти величины независимы. Поэтому

$$\sqrt{N-k}\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\mathsf{V}(\hat{\beta}_j)}}}{\sqrt{\frac{(N-k)s^2}{\sigma^2}}} = \frac{\hat{\beta}_j-\beta_j}{s\sqrt{[(X'X)^{-1}]_{jj}}}$$

имеет распределение Стьюдента $\mathbf{t}_{N-k}$. Выбирая по доверительной вероятности $1-\alpha$ соответствующее табличное значение z_α ($(1-\alpha/2)$–квантиль распределения Стьюдента), мы получаем доверительный интервал вида $\hat{\beta}_j \pm z_\alpha s\sqrt{[(X'X)^{-1}]_{jj}}$ для коэффициента β_j. При большом числе степеней свободы распределение Стьюдента, как обычно, может быть заменено нормальным.

Доверительный интервал позволяет проверять гипотезу вида $\beta_j = \beta_{j0}$. Для этого достаточно лишь выяснить, попадает ли гипотетическое значение β_{j0} в построенный доверительный интервал. Гипотеза отвергается на уровне α, если гипотетическое значение β_{j0} не попадает в доверительный интервал.

Проверка более сложных гипотез, включающих линейные комбинации коэффициентов регрессии, обсуждается в следующем параграфе.

Доверительный интервал для σ^2 строится непосредственно по χ^2-распределенной дроби $(N-k)s^2/\sigma^2$. Мы предполагаем, что читатель может проделать это самостоятельно.

Без предположения о нормальности ошибок оба специальных распределения — Стьюдента и хи-квадрат — исчезают, однако часто предполагают, что при больших N изложенные рецепты дают "приближенные"доверительные интервалы.

6.8 Проверка линейных гипотез общего вида

Простейшие гипотезы вида $\beta_j = \beta_{j0}$ о коэффициентах регрессии, рассмотренные выше, составляют лишь малую часть содержательных линейных гипотез. Обозначим на уровне идей ряд примеров, в которых появляются гипотезы другого вида.

Гипотеза $\beta_2+\beta_3 = 1$ появляется в связи с производственной функцией Кобба–Дугласа.

Гипотеза $\beta_2 + \beta_3 = 0$ может проверяться в модели, где X_2 — ставка банковского процента, а X_3 — уровень инфляции.

Гипотеза $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ появляется при выяснении вопроса о значимости всей регрессионной связи.

Общая формулировка линейной гипотезы о коэффициентах имеет следующий вид:

$$H_0 : \quad R\beta = \gamma.$$

Здесь R — матрица коэффициентов, имеющая k столбцов. Каждая ее строка (будем считать, что число строк равно r) задает линейное ограничение

$$R_{l1}\beta_1 + \cdots + R_{lk}\beta_k = \gamma_l, \qquad l = 1, \ldots, r.$$

Без ограничения общности можно считать, что строки матрицы ограничений R линейно независимы, так что $r \le k$ (как правило число ограничений значительно меньше k).

Как и в предыдущем параграфе, мы будем предполагать, что ошибки нормально распределены. Для построения теста проверки гипотезы H_0 воспользуемся тем, что случайный вектор $R\hat{\beta}$ распределен по нормальному закону с математическим ожиданием $R\beta$ и матрицей ковариаций

$$\mathsf{cov}(R\hat{\beta}) = \mathbf{E}(R\hat{\beta} - R\beta)(R\hat{\beta} - R\beta)' = \\ = R\mathsf{cov}(\hat{\beta})R' = \sigma^2 R(X'X)^{-1}R'.$$

Легко проверить, что эта матрица невырождена. Действительно, она представляется в виде R_*R_*', где $R_* = R(X'X)^{-1/2}$ — матрица полного ранга r. Отсюда следует, что нормально распределенный вектор

$$(R(X'X)^{-1}R')^{-1/2}(R\hat{\beta} - R\beta)$$

центрирован и имеет матрицу ковариаций $\sigma^2\mathbf{1}_r$. Поэтому нормализованная сумма квадратов его компонент

$$\sigma^{-2}(R\hat{\beta}-R\beta)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-R\beta)$$

распределена по закону χ^2_r. В предыдущем параграфе установлено, что случайная величина

$$\frac{(N-k)s^2}{\sigma^2}$$

также распределена по хи-квадрат (с $N-k$ степенями свободы) и что она не зависит от вектора оценок $\hat{\beta}$. Вспоминая, что отношение независимых хи-квадрат величин, деленных на соответствующие числа степеней свободы, имеет $\mathbf{F}$-распределение Фишера, получаем, что в предположении H_0 дробь

$$\frac{(R\hat{\beta}-\gamma)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-\gamma)/r}{s^2}$$

распределена по закону $\mathbf{F}_{r,N-k}$. Большие значения этой дроби образуют критическую область искомого теста. Точно так же, неравенства вида

$$(R\hat{\beta}-\gamma)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta}-\gamma)\leq \text{const}$$

задают совместные доверительные области для компонент вектора $R\beta$, ограниченные эллипсоидами (поверхностями второго порядка). В обоих случаях используются процентные точки $\mathbf{F}$-распределения.

Описанные в предыдущем параграфе доверительные интервалы укладываются в нашу теперешнюю схему в качестве частного случая, т.к. имеет место "символическое"равенство:

$$(\mathbf{t}_{N-k})^2=\mathbf{F}_{1,N-k}.$$

Тестирование вызывающей особый интерес гипотезы $\beta_2=\cdots=\beta_k=0$ детально обсуждается в параграфе 6.10.

6.9 Блочная регрессия

Рассмотрим модель, в которой регрессоры разбиты на два непересекающихся блока:

$$X=(X_{(1)},X_{(2)}),$$

содержащих, соответственно, k_1 и k_2 регрессоров ($k_1 + k_2 = k$). Для определенности будем предполагать, что $X_{(1)}$ состоит из первых k_1 регрессоров.

Вектор коэффициентов β при этом также разбивается на подвекторы $\beta_{(1)}$ и $\beta_{(2)}$. Мы получим двухэтапную процедуру построения подвектора $\hat{\beta}_{(2)}$ оценок наименьших квадратов, обобщающую схему, изложенную в параграфе 6.4. Важнейший частный случай (ср. с §6.4) — $X_{(1)} = X_1 = 1^{\rightarrow}$, $X_{(2)} = (X_2, \ldots, X_k)$, однако мы увидим в дальнейшем, что блочная структура оказывается полезной и совсем в других контекстах.

Запишем формулу (6.6) для оценок наименьших квадратов в блочной форме:

$$\begin{pmatrix} X_{(1)}'X_{(1)} & X_{(1)}'X_{(2)} \\ X_{(2)}'X_{(1)} & X_{(2)}'X_{(2)} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{(1)} \\ \\ \hat{\beta}_{(2)} \end{pmatrix} = \begin{pmatrix} X_{(1)}'Y \\ \\ X_{(2)}'Y \end{pmatrix},$$

так что

$$X_{(1)}'X_{(1)}\hat{\beta}_{(1)} + X_{(1)}'X_{(2)}\hat{\beta}_{(2)} = X_{(1)}'Y,$$
$$X_{(2)}'X_{(1)}\hat{\beta}_{(1)} + X_{(2)}'X_{(2)}\hat{\beta}_{(2)} = X_{(2)}'Y.$$

Поскольку регрессоры первой группы линейно независимы, матрица $X_{(1)}'X_{(1)}$ обратима. Выражая $\hat{\beta}_{(1)}$ из первого уравнения и подставляя во второе, получаем

$$X_{(2)}'X_{(1)}(X_{(1)}'X_{(1)})^{-1}[X_{(1)}'Y - X_{(1)}'X_{(2)}\hat{\beta}_{(2)}] + X_{(2)}'X_{(2)}\hat{\beta}_{(2)} = X_{(2)}'Y.$$

Производя перегруппировку, запишем это равенство в виде

$$[X_{(2)}'X_{(2)} - X_{(2)}'X_{(1)}(X_{(1)}'X_{(1)})^{-1}X_{(1)}'X_{(2)}]\hat{\beta}_{(2)}$$
$$= X_{(2)}'Y - X_{(2)}'X_{(1)}(X_{(1)}'X_{(1)})^{-1}X_{(1)}'Y.$$

Вводя естественные обозначения

$$P_{(1)} = X_{(1)}(X_{(1)}'X_{(1)})^{-1}X_{(1)}', \qquad P_{(1)}^{\perp} = \mathbf{1} - P_{(1)},$$

получаем

$$X_{(2)}'P_{(1)}^{\perp}X_{(2)}\hat{\beta}_{(2)} = X_{(2)}'P_{(1)}^{\perp}Y,$$

откуда

$$(P_{(1)}^{\perp}X_{(2)})'(P_{(1)}^{\perp}X_{(2)})\hat{\beta}_{(2)} = (P_{(1)}^{\perp}X_{(2)})'(P_{(1)}^{\perp}Y). \tag{6.14}$$

Вектор $P_{(1)}^{\perp}Y$ можно рассматривать как вектор остатков от проектирования Y на подпространство $\mathcal{L}(X_{(1)}) = \mathcal{L}(X_1, \ldots, X_{k_1})$.

Обозначим его Y_*. Точно так же, столбцы матрицы $P^{\perp}_{(1)}X_{(2)}$ можно рассматривать как остатки от проектирования регрессоров второй группы на $\mathcal{L}(X_{(1)})$. Обозначим эту матрицу остатков X_*. Тогда (6.14) приобретает вид, сходный с (6.6):

$$X_*'X_*\hat{\beta}_{(2)} = X_*'Y_*. \tag{6.14$'$}$$

Матрица X_* имеет линейно независимые столбцы, в чем легко убедиться, выражая эти столбцы через первоначальные регрессоры $X_1, \ldots, X_k$. Действительно,

$$X_* = X_{(2)} - P_{(1)}X_{(2)} = X_{(2)} - X_{(1)}L,$$

т. к. столбцы матрицы $P_{(1)}X_{(2)}$ — линейные комбинации регрессоров первой группы, т. е. представляются в виде $X_{(1)}L_j$, где L_j — некоторые векторы коэффициентов — столбцы матрицы L. Рассмотрим линейную комбинацию $X_*\gamma$ столбцов матрицы X_*. Она представляется в виде $X_{(2)}\gamma - X_{(1)}L\gamma$ и равна нулю только при $\gamma = 0$ (регрессоры $X_1, \ldots, X_k$ линейно независимы).

Из доказанной линейной независимости столбцов X_* следуют обратимость матрицы $X_*'X_*$ и возможность разрешить уравнение (6.14$'$):

$$\hat{\beta}_{(2)} = (X_*'X_*)^{-1}X_*'Y_*. \tag{6.15}$$

В неявном виде эта разрешимость, конечно, следует из разрешимости системы (6.6) для полного набора оценок $\hat{\beta}$.

Теперь, подводя итог, мы можем интерпретировать изложенную схему следующим образом. На первом шаге процедуры строятся регрессии Y на $X_{(1)}$ и каждого столбца матрицы $X_{(2)}$ на $X_{(1)}$. На втором шаге строится регрессия остатков Y_* регрессии первого шага на X_* — матрицу остатков остальных регрессий первого шага. Полученные на втором шаге оценки $\hat{\beta}_{(2)}$ — искомые оценки коэффициентов регрессии из второй группы.

Возвращаясь к первой группе коэффициентов, мы можем теперь написать

$$\hat{\beta}_{(1)} = (X_{(1)}'X_{(1)})^{-1}X_{(1)}'(Y - X_{(2)}\hat{\beta}_{(2)}) \tag{6.15$_1$}$$

Рассмотрим теперь частный случай, упомянутый в начале параграфа — $X_{(1)} = X_1 = 1^{\rightarrow}$. Тогда на первом шаге строятся регрессии на константу, остатками от которых являются векторы отклонений $y = Y - \bar{Y}^{\rightarrow}$, $x_j = X_j - \bar{X}_j^{\rightarrow}$ $(j = 2, \ldots, k)$. На втором шаге строится регрессия

вектора y на укороченный набор новых регрессоров $x_2, \ldots, x_k$. Формулу (6.15) можно записать в виде (x — матрица, составленная из столбцов $x_2, \ldots, x_k$)

$$\hat{\beta}_{(2)} = (x'x)^{-1}x'y \tag{6.16}$$

— оценка коэффициентов линейной регрессии в отклонениях. Для оставшегося коэффициента β_1 теперь легко получаем

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \cdots - \hat{\beta}_k\bar{X}_k. \tag{6.16$_1$}$$

Очевидно, (6.15) и (6.16) обобщают ранее полученные формулы (6.10).

Из формул (6.16) получаем также

$$\hat{Y} = \hat{\beta}_1^{\rightarrow} + \hat{\beta}_2X_2 + \cdots + \hat{\beta}_kX_k = \bar{Y}^{\rightarrow} + \hat{\beta}_2x_2 + \cdots + \hat{\beta}_kx_k. \tag{6.17}$$

Отсюда следует, что $\overline{\hat{Y}} = \bar{Y}$ (для парной регрессии это было получено в параграфе 6.4). Действительно, нужное соотношение непосредственно вытекает из очевидных равенств $\bar{x}_2 = \cdots = \bar{x}_k = 0$.

Мы будем использовать блочную регрессию при обсуждении проблем спецификации (см. параграф 6.12).

6.10 Коэффициент детерминации и качество прогноза

В этом параграфе мы предполагаем, что $X_1 = 1^{\rightarrow}$.

Наиболее короткое определение коэффициента детерминации — квадрат выборочного коэффициента корреляции между фактическими (Y) и прогнозными ($\hat{Y}$) значениями объясняемой величины. Отсюда происходят обозначение R^2 и соответствующая формула. Для вычисления,впрочем, используется несколько иная формула

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}, \tag{6.18}$$

которая получается несложными преобразованиями.

Запишем сначала по определению

$$R^2 = \frac{(y'\hat{y})^2}{y'y \cdot \hat{y}'\hat{y}} \quad .$$

Поскольку $\overline{\hat{Y}} = \bar{Y}$, имеем

$$y = Y - \bar{Y}^{\rightarrow} = \hat{Y} + \hat{\varepsilon} - \bar{Y}^{\rightarrow} = \hat{y} + \hat{\varepsilon}.$$

Поэтому

$$y'\hat{y} = (\hat{\varepsilon} + \hat{y})'\hat{y} = \hat{y}'\hat{y}$$

(мы воспользовались ортогональностью остатков $\hat{\varepsilon}$ с прогнозным вектором $\hat{Y}$ и регрессором $X_1 = 1^{\rightarrow}$). Теперь из определения коэффициента детерминации получаем

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{(y-\hat{\varepsilon})'(y-\hat{\varepsilon})}{y'y} = \frac{y'y - \hat{\varepsilon}'\hat{\varepsilon}}{y'y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y},$$

что и требовалось доказать.

Если вспомнить, что разложение $Y = \hat{Y} + \hat{\varepsilon}$ определяется не набором регрессоров, а порожденным ими подпространством $\mathcal{L}(X_1, \ldots, X_k)$, определение коэффициента детерминации (в любой форме) без изменения переносится на чуть более общий случай — когда $1^{\rightarrow}$ лежит в этом подпространстве (но не обязательно является регрессором).

Из определения R^2 непосредственно вытекает неравенство

$$0 \le R^2 \le 1.$$

Можно еще отметить, что коэффициент корреляции R между Y и $\hat{Y}$ неотрицателен и сам по себе (без возведения в квадрат), т.к. прогноз $\hat{Y}$ не хуже прогноза без использования регрессоров — посредством $\hat{Y}^{\rightarrow}$. Крайнее значение $R^2 = 1$ означает совпадение $Y = \hat{Y}$, ожидать этого равенства вряд ли целесообразно. Другое крайнее значение $R^2 = 0$ свидетельствует о незначимом вкладе регрессоров $X_2, \ldots, X_k$ в объяснение — см. ниже обсуждение проверки соответствующей гипотезы.

При добавлении в модель новых регрессоров коэффициент детерминации может лишь увеличиться — сумма квадратов остатков уменьшается.

Принято считать, что выражение

$$y'y = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

(оно иногда называется вариацией) характеризует изменчивость величины Y. В этих терминах R^2 показывает, какую часть вариации $y'y$ составляет объясненная моделью часть вариации $\hat{y}'\hat{y}$. Хотя традиционная эконометрика считает коэффициент детерминации достаточно важной характеристикой модели (скажем, его значение

вычисляется эконометрическими пакетами), роль коэффициента R^2 не следует преувеличивать. Все авторы учебников подробно объясняют проблемы, возникающие в связи с его использованием.

Во-первых, различные варианты определения перестают совпадать, если константа не лежит в подпространстве регрессоров. Приемлемого определения в этом случае дать не удается.

Во-вторых, R^2 не инвариантен относительно выбора объясняемой величины. Действительно, возьмем в качестве новой объясняемой величины $Y_* = Y - X\alpha$, где α — некоторый (известный) вектор коэффициентов. Тогда наша модель приобретет вид

$$Y_* = X\beta_* + \varepsilon,$$

причем, очевидно, $\beta_* = \beta - \alpha$. Вектор остатков $\hat{\varepsilon} = P^{\perp}\varepsilon$ в обоих случаях один и тот же (матрица $P^{\perp}$ не связана с выбором объясняемой величины). Однако вектор $y_* = y - x\alpha$ совсем не обязан иметь ту же длину, что и y. Поэтому и

$$R^2_* = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'_* y_*}$$

не обязан совпадать с R^2. В то же время прогнозные свойства обеих моделей одинаковы:

$$\hat{Y}_* = PY_* = PY - PX\alpha = \hat{Y} - X\alpha.$$

По-существу, мы имеем дело с двумя представлениями одной модели, а не с двумя моделями.

В-третьих, несмотря на кажущуюся объективность этой характеристики качества модели (мы имеем в виду безразмерность R^2), коэффициент детерминации можно сделать сколь угодно близким к единице (или даже равным ей), если присоединить к модели дополнительные регрессоры в достаточном числе. При этом совершенно не требуется, чтобы эта операция имела какой-нибудь содержательный экономический смысл, главное — линейная независимость регрессоров. В учебной литературе обсуждается так называемый подправленный или скорректированный (adjusted) на число регрессоров коэффициент:

$$1 - R^2_{adj} = \frac{N-1}{N-k}(1 - R^2),$$

который далее использоваться не будет. Убедительного объяснения именно такой формулы для R^2_{adj} мы не нашли.

Наиболее важным применением коэффициента детерминации является использование его при тестировании значимости регрессионной модели в целом — при проверке гипотезы $H_0 : \beta_2 = \cdots = \beta_k = 0$. Опишем это применение более подробно.

Как уже было отмечено выше, о малой значимости регрессии свидетельствуют малые значения R^2. Остается (предполагая ошибки нормально распределенными) связать с R^2 одно из традиционных шаблонных распределений. Формулы (6.18) позволяют сделать это без труда. Действительно,

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y}, \quad 1 - R^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}.$$

Деля первое равенство на второе, получаем

$$\frac{R^2}{1 - R^2} = \frac{\hat{y}'\hat{y}}{\hat{\varepsilon}'\hat{\varepsilon}} = \frac{\hat{y}'\hat{y}/\sigma^2}{\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2}. \tag{6.19}$$

При этом для модели с нормально распределенными ошибками числитель и знаменатель последней дроби независимы и распределены по закону χ^2. Действительно, мы уже проверяли в параграфе 6.7 независимость $\hat{\beta}$ и $\hat{\varepsilon}$, откуда следует независимость $\hat{y}$ и $\hat{\varepsilon}$, а, тем самым, и желаемая независимость числителя и знаменателя. Там же установлено, что $\hat{\varepsilon}'\hat{\varepsilon}/\sigma^2$ распределена по закону χ^2_{N-k}. Остается разобраться с числителем.

Заметим сначала, что согласно формуле (6.17)

$$\hat{y} = \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k,$$

так что $\mathbf{E}\hat{y} = \beta_2 x_2 + \cdots + \beta_k x_k$. Последнее выражение равно нулю в предположении справедливости H_0. Кроме того, очевидно, вектор $\hat{y}$ нормально распределен. Вычислим, снова в предположении справедливости H_0, его матрицу ковариаций

$$\mathsf{cov}(\hat{y}) = \mathbf{E}(\hat{y}\hat{y}').$$

Будем при этом использовать обозначение $P_{(2)} = x(x'x)^{-1}x'$ в духе параграфа 6.9. Геометрический смысл матрицы $P_{(2)}$ фактически уже был получен в 6.9 — это матрица проектирования на $(k-1)$-мерное подпространство в $\mathcal{L}(X_1, \ldots, X_k)$, состоящее из векторов, ортогональных $X_1 = 1^{\rightarrow}$.

Заметим еще, что согласно формулам (6.17) и (6.16) из параграфа 6.9

$$\hat{y} = x\hat{\beta}_{(2)} = x(x^{'}x)^{-1}x^{'}y = P_{(2)}y.$$

Кроме того,

$$y - \mathbf{E}y = (Y - \mathbf{E}Y) - (\bar{Y} - \mathbf{E}\bar{Y})^{\rightarrow} = \varepsilon - (\bar{\varepsilon})^{\rightarrow}.$$

Легко сообразить, что $P_{(2)}(\bar{\varepsilon})^{\rightarrow} = 0$. Поэтому

$$\begin{aligned}\mathsf{cov}(\hat{y}) &= \mathbf{E}[P_{(2)}(\varepsilon - (\bar{\varepsilon})^{\rightarrow})(\varepsilon - (\bar{\varepsilon})^{\rightarrow})^{'}P_{(2)}] \\ &= \mathbf{E}[P_{(2)}\varepsilon\varepsilon^{'}P_{(2)}] = P_{(2)}\mathbf{E}[\varepsilon\varepsilon^{'}]P_{(2)} = \sigma^2 P_{(2)}.\end{aligned}$$

Теперь утверждение о том, что величина $\hat{y}^{'}\hat{y}/\sigma^2$ распределена по закону χ^2_{k-1}, доказывается тем же рассуждением, что и аналогичное утверждение для $\hat{\varepsilon}^{'}\hat{\varepsilon}/\sigma^2$ в параграфе 6.7 (напомним, что мы рассуждаем в предположении справедливости гипотезы H_0, так что $\mathbf{E}\hat{y} = 0$).

Возвращаясь, наконец, к (6.19), заключаем, что дробь

$$\frac{R^2/(k-1)}{(1-R^2)/(N-k)} = \frac{N-k}{k-1}\frac{R^2}{1-R^2}$$

имеет распределение Фишера $\mathbf{F}_{k-1,N-k}$. Остается взять нужную процентную точку $\mathbf{F}$-распределения и зафиксировать критическую область теста вида

$$\frac{R^2}{1-R^2} \geq \text{const}.$$

Упражнение. Используя блочную регрессию общего вида, обобщить проведенное рассуждение и доказать, что в предположении справедливости гипотезы $\beta_{(2)} = 0$ дробь

$$\frac{(R^2 - R^2_{(1)})/k_2}{(1-R^2)/(N-k)} = \frac{(\hat{\varepsilon}^{'}_{(1)}\hat{\varepsilon}_{(1)} - \hat{\varepsilon}^{'}\hat{\varepsilon})/k_2}{\hat{\varepsilon}^{'}\hat{\varepsilon}/(N-k)}$$

имеет распределение Фишера $\mathbf{F}_{k_2,N-k}$.

6.11 Индикаторные величины в линейной модели

Индикаторными или сигнальными мы называем величины, принимающие только два значения — 0 и 1 (английский термин — dummy; в русскоязычных текстах можно встретить крайне неудачный

перевод "фиктивная— и неверно по сути, и бессмысленно). Величины такого сорта появляются во многих случаях, когда неоднородность эмпирических данных имеет "групповой"характер, и мы пытаемся учесть ее, не выходя за рамки классической модели. Рассмотрим несколько стандартных примеров.

Пример 1. Индикатор военного времени. Если эмпирические данные представляют собой временной ряд (например, годичные данные), включающий, скажем, показатели, относящиеся к промежутку между двумя мировыми войнами, к периоду второй мировой войны и к послевоенному периоду, то может оказаться важным выделение военного времени. Это можно сделать следующим образом. Рассмотрим индикаторную величину I, принимающую значение $I_i = 1$ для военных лет, и значение $I_i = 0$ для остальных. С ее помощью каждый регрессор X_j, для которого различия мирного и военного времени кажутся нам существенными, порождает парную величину IX_j, которая включается в линейную модель со своим коэффициентом γ_j. Таким образом, модель включает слагаемые $\beta_j X_j$ и $\gamma_j IX_j$, которые учитывают различия мирного и военного времени на уровне коэффициентов. Для мирных лет в модели присутствует слагаемое $\beta_j X_j$, а для военных — слагаемое $(\beta_j + \gamma_j)X_j$. Тем самым, некоторым образом показатель X_j "переключается"с одного режима на другой.

Пример 2. Сезонные колебания. Аналогично примеру 1 можно учесть колебания коэффициентов по месяцам или другим естественным периодам. Для каждого месяца можно ввести свой индикатор: $I_1, I_2, \ldots, I_{12}$. По очевидным причинам сумма этих двенадцати индикаторов тождественно равна единице, так что они линейно зависимы. Поэтому, вводя величины $I_1 X_j, \ldots, I_{12} X_j$, мы должны опустить исходную величину X_j. Конечно, в примере 1 можно было бы поступить аналогичным образом.

Общая черта рассмотренных примеров — моменты переключения режимов известны. В примере 1 это не вполне очевидно, т. к. определенные факторы могут иметь последействие. Попытки обобщения вывели бы нас за рамки классической модели, и мы не будем сейчас их обсуждать.

Дискретные величины более чем с двумя значениями, обобщающие индикаторы, практически не используются, т. к. их удобнее заменять более простыми индикаторами, увеличивая при необходимости их число (как в примере 2). Выигрыша в числе параметров, заменяя один способ другим, не добиться.

В качестве иллюстрации использования индикаторных величин рассмотрим так называемый тест Чоу (Chow) проверки совпадения моделей. Предположим, что мы имеем дело с двумя сериями из N_1 и N_2 однотипных наблюдений:

$$Y_{(1)} = X_{(1)}\beta_{(1)} + \varepsilon_{(1)}, \quad Y_{(2)} = X_{(2)}\beta_{(2)} + \varepsilon_{(2)}.$$

Однотипность понимается как совпадение множеств регрессоров в двух сериях (в содержательном смысле — если в первой серии X_2 — процентная ставка, то и во второй серии X_2 — процентная ставка). Будем предполагать также, что дисперсии ошибок одинаковы (это предположение, вообще говоря, сомнительно, но отказ от него снова выведет нас за рамки классической модели, и потому это обобщение сейчас обсуждаться не будет). Рассмотрим проверку гипотезы $\beta_{(1)} = \beta_{(2)}$. Для этого введем индикатор второй серии I и расмотрим объединенную систему данных

$$Y = \begin{pmatrix} Y_{(1)} \\ \\ Y_{(2)} \end{pmatrix}, \quad X = \begin{pmatrix} X_{(1)} & IX_{(1)} \\ \\ X_{(2)} & IX_{(2)} \end{pmatrix} = \begin{pmatrix} X_{(1)} & 0 \\ \\ X_{(2)} & X_{(2)} \end{pmatrix}.$$

Соответствующая спецификация имеет вид

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

где

$$\varepsilon = \begin{pmatrix} \varepsilon_{(1)} \\ \\ \varepsilon_{(2)} \end{pmatrix}, \gamma = \begin{pmatrix} \beta_{(1)} \\ \\ \beta_{(2)} - \beta_{(1)} \end{pmatrix}.$$

Наша гипотеза $\beta_{(1)} = \beta_{(2)}$, или, эквивалентно, $\gamma_{(2)} = 0$, имеет вид, обсуждавшийся ранее, и проверяется (это и есть тест Чоу) с использованием распределения $\mathbf{F}_{k,N_1+N_2-2k}$ — см. упражнение в конце параграфа 6.10. При этом коэффициент детерминации R^2 и вектор остатков $\hat{\varepsilon}$ вычисляются по полной регрессионной матрице X, а коэффициент детерминации $R^2_{(1)}$ и вектор остатков $\hat{\varepsilon}_{(1)}$ — по

уменьшенной (restricted) матрице

$$X_{restr} = \begin{pmatrix} X_{(1)} \\ \\ X_{(2)} \end{pmatrix}.$$

6.12 Замечания о спецификации модели

На практике исследователь **выбирает** спецификацию модели. Сделать сразу окончательный выбор, как правило, не удается. Так, если речь идет о прогнозировании спроса на депозитные сертификаты, можно предполагать, что среди регрессоров окажутся ставка процента по этим сертификатам, ставка процента по каким-либо конкурирующим ценным бумагам и т. д. С уверенностью включать или не включать тот или иной регрессор в модель вряд ли возможно. Поэтому рассматриваются различные варианты модели, с тем чтобы в конечном итоге остановиться на одном из них. В примере с депозитными сертификатами можно попытаться учесть, скажем, разность между ставками процента по краткосрочным и долгосрочным вложениям. Но целесообразно ли это — является ли соответствующий фактор существенным (статистически значимым)? Ответы на подобные вопросы можно получить, только анализируя эмпирические данные и сравнивая разные модификации модели. При этом может оказаться, что некоторые регрессоры — лишние, а некоторые, наоборот, пропущены. Мы обсудим в этом параграфе часть подобных вопросов, связанных с выбором спецификации модели.

Начнем с замечаний концептуально-философского характера. Как понимать высказывание о том, что данная модель правильна (true model)? И существует ли вообще таковая? Вопросы "взаимоотношений"между моделью и моделируемым явлением достаточно деликатны. Обсуждаемые нами линейные регрессионные модели включают стохастическую ошибку ε, концентрирующую в себе всю совокупность неучтенных факторов, и потому в само́м линейном представлении $Y = X\beta + \varepsilon$ еще нет потенциальных трудностей. Проблемы появляются, когда мы начинаем постулировать какие-либо свойства стохастической ошибки. Проверить (тестировать) постулируемые свойства удается не всегда, надежность соответствующего вывода может быть невысокой. Надежный же вывод, скорее всего, окажется отрицательным. Таким образом, представление о том, что имеется некоторая "правильная"модель, является (еще

одной) идеализацией, появляющейся в процессе моделирования. В этом параграфе мы только начинаем обсуждение проблем спецификации, поэтому будем, все-таки, считать, что "правильную"модель можно представить себе, и для нее выполнены классические предположения.

Будем записывать правильную модель в виде

$$Y = X_t\beta_t + \varepsilon_t; \tag{6.20}$$

здесь индекс t является сокращением от true. Помимо модели (6.20), имеющей только умозрительный характер, исследователь имеет дело с фактической спецификацией $Y = X\beta + \varepsilon$, которая меняется в процессе работы.

Рассмотрим сначала относительно безобидный (как будет видно дальше) случай, когда в спецификацию включены дополнительные ("лишние") регрессоры, так что

$$X = (X_t, X_c),$$

и

$$Y = X_t\beta_{(1)} + X_c\beta_{(2)} + \varepsilon,$$

где $\beta_{(1)}$ и $\beta_{(2)}$ — частичные векторы коэффициентов. Отметим, что правильная модель получается при $\beta_{(2)} = 0$, но нам это неизвестно. Мы, надо думать, считаем, что вектор β, подразумеваемый нашей спецификацией, и есть правильный вектор коэффициентов β_t, что не совсем точно (они имеют разные размерности), и что вектор ошибок ε есть правильный вектор ошибок ε_t — это похоже на истину, впрочем, с оговоркой, что ошибки все-таки не наблюдаемы.

С практической точки зрения мы можем оценить коэффициенты β нашей спецификации стандартным образом, т. е. найти по выборке их оценки $\hat{\beta}$, а также соответствующие остатки $\hat{\varepsilon}$. На самом-то деле наша спецификация ошибочна (точнее, избыточна), так что таковы же и выражения для $\hat{\beta}$ и $\hat{\varepsilon}$. Точнее, частичный вектор $\hat{\beta}_{(1)}$ оценивает вектор β_t правильных коэффициентов, а $\hat{\beta}_{(2)}$ "оценивает"нулевой вектор. При обсуждении блочной регрессии в параграфе 6.9 мы получили формулы (6.15), из которых следует

$$\hat{\beta}_{(2)} = (X_c' P_t^{\perp} X_c)^{-1} X_c' P_t^{\perp} Y = (X_c' P_t^{\perp} X_c)^{-1} X_c' P_t^{\perp} \varepsilon_t,$$

$$\hat{\beta}_{(1)} = \beta_t + (X_t' P_c^{\perp} X_t)^{-1} X_t' P_c^{\perp} \varepsilon_t.$$

Эти оценки несмещенные —

$$\mathbf{E}\hat{\beta}_{(1)} = \beta_t, \quad \mathbf{E}\hat{\beta}_{(2)} = 0,$$

но неправильный выбор спецификации привел к потере в эффективности:

$$\mathsf{cov}(\hat{\beta}_{(1)}) = \sigma^2(X_t' P_c^{\perp} X_t)^{-1} \geq \sigma^2(X_t' X_t)^{-1} = \mathsf{cov}(\hat{\beta}_t),$$

$$\mathsf{cov}(\hat{\beta}_{(2)}) = \sigma^2(X_c' P_t^{\perp} X_c)^{-1} \geq 0 = \mathsf{cov}(0).$$

Первое неравенство вытекает из того, что

$$X_t' X_t - X_t' P_c^{\perp} X_t = X_t' P_c X_t \geq 0,$$

а второе — самоочевидно.

Эффективность — это важное свойство, так что злоупотреблять включением в модель лишних регрессоров не следует. Выявить наличие их поможет проверка гипотезы вида $\hat{\beta}_{(2)} = 0$ — она обсуждалась в параграфе 6.10.

Рассмотрим теперь оценку дисперсии σ_t^2 в рамках выбранной спецификации. Такой оценкой является

$$s^2 = \hat{\varepsilon}'\hat{\varepsilon}/(N-k),$$

где $k = k_1 + k_2$ — полное число коэффициентов. Она, естественно, отличается от

$$s_t^2 = \hat{\varepsilon}_t'\hat{\varepsilon}_t/(N-k_1),$$

но, как это ни парадоксально, обе оценки s_t^2 и s^2 являются несмещенными. Это следует из общих соображений — обе они получаются одной и той же процедурой, только в разных спецификациях. Первая — в правильной спецификации (6.20), а вторая — в фактически выбранной. Отметим, впрочем, что несмещенность s^2 можно проверить и непосредственно, используя несложно проверяемые соотношения

$$\hat{\varepsilon}_t = P_t^{\perp} Y = P_t^{\perp} X_c \hat{\beta}_{(2)} + \hat{\varepsilon},$$

$$\hat{\varepsilon}_t'\hat{\varepsilon}_t = \hat{\varepsilon}'\hat{\varepsilon} + \hat{\beta}_{(2)}' X_c' P_t^{\perp} X_c \hat{\beta}_{(2)},$$

$$\mathbf{E}[\hat{\beta}_{(2)}' X_c' P_t^{\perp} X_c \hat{\beta}_{(2)}] = \sigma_t^2 k_2.$$

Поскольку оценки различаются, а s_t^2 в модели с нормально распределенными ошибками — эффективная несмещенная оценка (см. параграф 6.7), то и здесь происходит потеря в эффективности.

Перейдем теперь к более печальной ситуации, когда выбранная спецификация не включает часть регрессоров из правильной модели, т.е. $X_t = (X, X_c)$. Теперь вектор β_t правильных коэффициентов разбивается на два подвектора $\beta_{t(1)}$ и $\beta_{t(2)} \neq 0$. Коэффициенты $\beta_{t(1)}$ отвечают регрессорам, включенным в нашу спецификацию $Y = X\beta + \varepsilon$. Оценки $\hat{\beta}$, которые мы можем построить, предназначаются для оценивания $\beta_{t(1)}$, что же касается коэффициентов $\beta_{t(2)}$, то мы, видимо, и не подозреваем о соответствующих объясняющих факторах, или не считаем их важными.

К сожалению, оценка $\hat{\beta}$, вообще говоря, смещена:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta_{t(1)} + X_c\beta_{t(2)} + \varepsilon_t) \\ &= \beta_{t(1)} + (X'X)^{-1}X'X_c\beta_{t(2)} + (X'X)^{-1}X'\varepsilon_t, \\ \mathbf{E}\hat{\beta} &= \beta_{t(1)} + (X'X)^{-1}X'X_c\beta_{t(2)}. \end{aligned}$$

Несмещенной оценка $\hat{\beta}$ оказывается в исключительном случае $X'X_c = 0$, когда столбцы дополнительных регрессоров ортогональны столбцам использованных регрессоров (исключение и есть исключение).

Рассмотрим теперь оценку дисперсии σ_t^2. Имеем

$$\begin{aligned} \hat{\varepsilon} = Y &- X\hat{\beta} = X\beta_{t(1)} + X_c\beta_{t(2)} + \varepsilon_t \\ &- X(\beta_{t(1)} + (X'X)^{-1}X'X_c\beta_{t(2)} + (X'X)^{-1}X'\varepsilon_t) \\ &= P^{\perp}X_c\beta_{t(2)} + P^{\perp}\varepsilon_t. \end{aligned}$$

Поэтому

$$\hat{\varepsilon}'\hat{\varepsilon} = \beta'_{t(2)}X'_cP^{\perp}X_c\beta_{t(2)} + \varepsilon'_tP^{\perp}\varepsilon_t + 2\beta'_{t(2)}X'_cP^{\perp}\varepsilon_t.$$

Второе слагаемое имеет требуемое математическое ожидание $(N-k)\sigma_t^2$, третье — вклада не дает, т.к. $\mathbf{E}\varepsilon_t = 0$. Наконец, первое слагаемое, очевидно, практически всегда положительно (даже в случае $X'X_c = 0$, когда оно обращается в $(X_c\beta_{t(2)})'X_c\beta_{t(2)}$). Таким образом,

$$s^2 = \hat{\varepsilon}'\hat{\varepsilon}/(N-k)$$

— смещенная вправо оценка дисперсии.

Обсудим, в завершение параграфа, вопрос о том, как выявить пропуски регрессоров в модели с нормально распределенными ошибками. Для этого заметим, что

$$\mathbf{E}\hat{\varepsilon} = P^{\perp}X_c\beta_{t(2)}.$$

Предположим сначала, что этот вектор отличен от нуля. В этом случае можно (как и в параграфе 6.7) выбрать некоторый ортонормированный базис $e_1, \ldots, e_{N-k}$ в подпространстве $\mathcal{L}^{\perp}(X_1, \ldots, X_k)$, где лежит $\hat{\varepsilon}$, составить из этих векторов–столбцов матрицу e и рассмотреть вектор $e^{'}\hat{\varepsilon}$ с координатами $e_j^{'}\hat{\varepsilon}$. Очевидно,

$$\mathbf{E}(e^{'}\hat{\varepsilon}) = e^{'}P^{\perp}X_c\beta_{t(2)} = e^{'}X_c\beta_{t(2)}.$$

Кроме того, доказанная в параграфе 6.7 формула

$$\mathsf{cov}(e^{'}\hat{\varepsilon}) = \sigma^2\mathbf{1}_{N-k},$$

очевидно, справедлива и сейчас (единственное отличие, нецентрированность $e^{'}\hat{\varepsilon}$, не играет роли при вычислении ковариаций).

Среднее арифметическое случайных величин $e_j^{'}\hat{\varepsilon}$

$$\frac{1}{N-k}(1^{\rightarrow})^{'}e^{'}\hat{\varepsilon} \tag{6.21}$$

представляется в виде суммы своего математического ожидания

$$\frac{1}{N-k}(1^{\rightarrow})^{'}e^{'}X_c\beta_{t(2)} \tag{6.22}$$

и среднего арифметического $N-k$ независимых величин с распределением $\mathbf{N}(0, \sigma^2)$. Поэтому можно построить доверительный интервал для (6.22) — среднего значения нормально распределенной с дисперсией $\sigma^2/(N-k)$ случайной величины (6.21).

Конечно, дисперсия σ^2 нам не известна, но если воспользоваться завышенной (см. выше) оценкой s^2, то мы будем лишь несколько реже отвергать гипотезу $\beta_{t(2)} = 0$ и следствия из нее, но, все-таки, при удачном стечении обстоятельств сможем выявить отсутствие центрированности для вектора $e^{'}\hat{\varepsilon}$.

Если математическое ожидание (6.22) обращается в 0, наш прием непригоден. В этом случае можно попытаться сменить базисную матрицу e.

Большим недостатком указанного метода является необходимость строить матрицу e — это трудоемкая вычислительная задача. В некоторых случаях можно не доводить построение базиса $\{e_j\}$ до конца и ограничиться несколькими первыми базисными векторами.

Вернемся теперь к случаю $\mathbf{E}\hat{\varepsilon} = 0$, когда не поможет никакое изменение матрицы e. В этой ситуации можно попытаться уменьшить

выборку, отбрасывая одно или несколько наблюдений. В любом случае, можно надеяться, что возможный пропуск регрессоров вскроется после нескольких попыток. А на "нет", как говорят в статистике[1], и суда нет.

И, наконец, последнее замечание. Предположим, что мы выявили нечто, похожее на пропуск регрессоров. Ведь это всего-лишь сигнал о том, что "что-то не в порядке". Предположений в нашей модели довольно много, и, может быть, нарушается одно из свойств ошибок. На этой вопросительной ноте мы заканчиваем параграф и главу.

[1]И не только в статистике!

Глава 7

Анализ регрессионных предположений

Классические предположения, на основе которых в предыдущей главе излагалась статистическая техника исследования линейной регрессионной модели, удовлетворяют эконометриста лишь в редких случаях. Чаще всего он вынужден отказываться от части этих предположений. Ниже обсуждаются связанные с этим проблемы.

Удобно еще раз повторить в сжатом, но явном, виде весь список использовавшихся в главе 6 свойств.

Регрессоры X_j неслучайны и линейно независимы.

Ошибки ε_i случайны, центрированы, не коррелируют, имеют одинаковые дисперсии.

Во многих местах дополнительно предполагалось, что ошибки совместно нормально распределены.

Несколько первых параграфов настоящей главы посвящены изменению отдельных предположений этого перечня. Остальные предположения при этом чаще всего предполагаются справедливыми, может быть, в слегка уточненном виде. Более решительные обобщения классической модели по мере возможности представлены во второй части главы.

7.1 Стохастические регрессоры

Как уже упоминалось в параграфе 6.2, неслучайность регрессоров — довольно специфическое и редкое обстоятельство. Объявить их стохастическими (т.е. случайными) — дело нехитрое. Сложнее уточнить подобную декларацию осмысленными предположениями о характере этой случайности и о взаимоотношениях вводимых в

модель дополнительных случайных величин с уже имеющимися, т.е. с ошибками. Не следует забывать и о том, что некоторые регрессоры (константа, время,...) принципиально неслучайны.

До тех пор, пока рассматриваемая модель включает **одно** уравнение (т.е. до тех пор, пока мы предполагаем, что смогли выделить фрагмент экономической действительности, допускающий осмысленное описание посредством **одного** уравнения), регрессоры (объясняющие величины) мы вынуждены трактовать экзогенным (внешним) образом. Это относится и к их законам распределения (в самом общем варианте — к совместному распределению величин $X_{ij}, i = 1, \ldots, N, j = 1, \ldots, k$). Константу из нашего списка можно, разумеется, убрать, однако возможность влияния других неслучайных регрессоров (например, времени) следует, вообще говоря, предусмотреть. В любом случае совместное распределение регрессоров задается экзогенно. Предполагать некоррелированность или независимость вдоль последовательности наблюдений (т.е. при разных i) без должной мотивировки, проведенной в содержательных экономических терминах, не следует. Временные ряды и пространственные данные в этом отношении чаще всего различаются.

Перейдем теперь к подробному и точному описанию модели со стохастическими регрессорами. Остальные предположения в этом параграфе мы лишь уточняем, не меняя их сути.

Итак, пусть регрессионная матрица X случайна, а ее распределение задано экзогенно, причем с вероятностью 1 столбцы регрессоров $X_1, \cdots, X_k$ линейно независимы.

Предположим далее, что при фиксированной матрице X условное распределение вектора ошибок[1] ε удовлетворяет всем основным классическим предположениям:

центрированность: $\mathbf{E}(\varepsilon|X) = 0$;

отсутствие корреляций: $\mathbf{E}(\varepsilon_{i_1}\varepsilon_{i_2}|X) = 0$ $(i_1 \neq i_2)$;

однородность дисперсий: $\mathbf{E}(\varepsilon_i^2|X) = \sigma^2$ (значение дисперсии **не зависит** от условия X).

[1]Точнее, при почти любом выборе условия X (все утверждения об условных распределениях и условных математических ожиданиях, в соответствии с определениями, выполняются при почти всех условиях, т.е. с вероятностью 1 — см. приложение D).

При желании (или необходимости) можно дополнительно предположить условную нормальность вектора ε. В этом случае вектор ошибок ε и регрессионная матрица X оказываются стохастически независимыми, а условные характеристики ошибок становятся безусловными.

Для сформулированной таким образом модели можно (в условном смысле) воспользоваться многими формулами и утверждениями главы 6. После этого усреднение по условиям дает нам и безусловные результаты. Опишем эту схему рассуждений более подробно.

Прежде всего, при фиксированном условии X получаем формулу (6.7) для оценок метода наименьших квадратов:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Она остается осмысленной и в ситуации, когда регрессионная матрица X трактуется как случайная.

Это же относится и к выражению $\hat{Y} = X\hat{\beta}$ для вектора прогнозных значений.

Перейдем к свойствам оценок метода наименьших квадратов. Почти очевидно, что $\hat{\beta}$ — несмещенная оценка вектора коэффициентов. Действительно, в соответствии со свойствами условного математического ожидания

$$\mathbf{E}\hat{\beta} = \mathbf{E}(\mathbf{E}(\hat{\beta}|X)) = \mathbf{E}\beta = \beta$$

(в условном смысле несмещенность доказана в параграфе 6.5).

При небольшом уточнении формулировки сохраняется и теорема Гаусса-Маркова (вместе с доказательством). Уточнение касается класса оценок. Именно, рассматриваются (ср. с параграфом 6.5) линейные по Y оценки вида $\tilde{\beta} = CY$, где C — матрица коэффициентов, элементы которой являются функциями от регрессионной матрицы X. Условие несмещенности такой оценки, как и в параграфе 6.5, имеет вид $CX = \mathbf{1}$ (с вероятностью 1). Оценки наименьших квадратов являются эффективными в классе линейных по Y несмещенных оценок указанного вида. Доказательство, как и приведенное выше доказательство несмещенности, использует свойства условных математических ожиданий и опирается на доказательство теоремы Гаусса-Маркова, приведенное в параграфе 6.5, а также на то обстоятельство, что условное математическое ожидание $\mathbf{E}(\hat{\beta}|X)$ не зависит от условия.

Дисперсия σ^2 ошибок оценивается тем же выражением, что и в параграфе 6.6. Оператор ортогонального проектирования $P^\perp$ зависит от регрессионной матрицы X, однако доказательство равенства $\mathbf{E}(\hat{\varepsilon}'\hat{\varepsilon}|X) = \sigma^2 \mathsf{tr} P^\perp$ совпадает с доказательством аналогичного безусловного равенства в главе 6, а соотношение $\mathsf{tr} P^\perp = N - k$ справедливо с вероятностью 1. Поэтому статистика

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{N-k}$$

остается несмещенной оценкой дисперсии.

Прослеживая рассуждения параграфа 6.7, относящиеся к модели с нормально распределенными ошибками, легко обнаружить, что и они, в основном, воспроизводятся. Особый интерес представляет то обстоятельство, что условное распределение дроби

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{[(X'X)^{-1}]_{jj}}}$$

есть распределение Стьюдента при почти всех условиях. Отсюда немедленно вытекает, что и безусловное распределение этой дроби стьюдентовское, так что конструкция доверительных интервалов сохраняется и в модели со стохастическими регрессорами.

Аналогичным образом, сохраняются и результаты параграфа 6.8 о проверке линейных гипотез общего вида (матрица R коэффициентов предполагается постоянной, т.е. не зависящей от X). Мы оставляем читателю выяснение вопроса о том, в какой степени воспроизводятся в модели со стохастическими регрессорами остальные результаты главы 6.

7.2 Проблема мультиколлинеарности

Второе предположение о регрессорах — линейная независимость, является абсолютно необходимым с точки зрения абстрактной теории, однако на практике иногда сильно досаждает исследователям.

Предположим, что регрессоры вдруг оказываются линейно зависимыми. Это означает, что по меньшей мере один из них может быть линейно выражен через остальные. При этом ранг матрицы X, а вместе с ним и ранг матрицы $X'X$, оказываются строго меньше k — числа регрессоров, а тогда $X'X$ необратима. Вся цепочка рассуждений,

приведших нас в параграфе 6.3 к оценкам наименьших квадратов, а затем к их свойствам, рушится.

В содержательных терминах эта необратимость означает следующее. Проекция $\hat{Y}$ вектора Y на подпространство регрессоров $\mathcal{L}(X_1, \cdots, X_k)$ (она, разумеется, существует) может быть по-разному выражена через них. Поэтому коэффициенты этого разложения предметного (объясняющего) смысла не имеют.

Как же должен поступить исследователь, обнаруживший подобную линейную зависимость (=коллинеарность) регрессоров? Скорее всего он изменит спецификацию модели, выразив один (или даже несколько) регрессоров через остальные и исключив их тем самым. По-видимому, линейные соотношения между экзогенными величинами должны иметь какое-либо осмысленное (экономическое или управленческое) объяснение. Разумеется, могут возникнуть исключительные обстоятельства, но этой возможностью обычно пренебрегают.

К сожалению, относительно нередко регрессоры оказываются "почти линейно зависимыми"(т.е. по содержательным причинам меняют свои значения хоть и не синхронно, но очень похожим образом). С вычислительной точки зрения это выражается в том, что определитель матрицы $X'X$ близок к нулю (в некотором смысле, который нуждается в уточнении), а обращение этой матрицы приводит (по крайней мере, потенциально) к катастрофически большим погрешностям (вычисления, разумеется, призводятся на компьютере и, практически всегда, с округлением). В результате теряется доверие к оценкам коэффициентов. Может возникнуть и более "экзотическая"ситуация, когда отдельные коэффициенты регрессии незначимо отличаются от нуля, а совместно они значимы (гипотезы о параметрах обсуждались в параграфе 6.8).

Подобные явления принято называть мультиколлинеарностью. Все авторы учебников соглашаются с тезисом о важности проблемы мультиколлинеарности, но по-разному оценивают возможности исследователя в преодолении этой трудности (см., например, [19, 9, 25]). Универсального рецепта, несомненно, существовать не может, а на практике, как указывают [25], чаще всего приходится менять "правила игры".

7.3 Асимптотические свойства оценок метода наименьших квадратов

Перейдем теперь к обсуждению проблемы, которая в контексте главы 6 не затрагивалась, именно, проблемы состоятельности оценок МНК. После обсуждения в параграфе 1 стохастических регрессоров изучение состоятельности окажется более содержательным, хотя мы и начнем со специального случая "управляемых" **неслучайных** объясняющих факторов.

В данном контексте "управляемость"будет означать всего лишь, что регрессионная матрица X меняется с ростом числа наблюдений некоторым предписанным образом. Собственно говоря, в нормальных уравнениях метода наименьших квадратов присутствует не сама матрица X с растущим числом строк, а произведение $X'X$ — матрица **фиксированного** размера $k \times k$. Меняются ее элементы, представляющие собой суммы растущего числа N слагаемых. Простейшее разумное предписание поведения этих сумм — асимптотический линейный рост по N. Это приблизительно соответствует некоторой стационарности в поведении экзогенных величин $X_1, \cdots, X_k$. Мы выразим эту асимптотическую линейность стандартным образом — предположим, что существует предел

$$\lim_{N\to\infty} \frac{1}{N} X'X = Q. \tag{7.1}$$

Во избежание возникновения проблемы мультиколлинеарности (см. параграф 2) матрицу Q мы будем считать невырожденной. Для выяснения условий состоятельности обратимся теперь к уравнению (6.9):

$$\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon.$$

Легко сообразить, что при сделанных предположениях (Q невырождена) состоятельность оценки $\hat{\beta}$ вытекает из соотношения

$$\frac{1}{N} X'\varepsilon \to 0 \tag{7.2}$$

(по вероятности при $N \to \infty$). Пока мы предполагаем неслучайность регрессоров и некоррелированность ошибок, (7.2) выполняется автоматически. Действительно, векторная случайная величина $\frac{1}{N}X'\varepsilon$ имеет нулевое математическое ожидание, а дисперсии ее компонент

$\frac{1}{N}(X'\varepsilon)_j$ стремятся к нулю:

$$\mathbf{V}\left(\frac{1}{N}\sum_{i=1}^{N} X_{ij}\varepsilon_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} X_{ij}^2\sigma^2 = \frac{1}{N}[q_{jj} + o(1)]\sigma^2 \to_{N\to\infty} 0$$

(здесь q_{ij} — соответствующий элемент матрицы Q). Соотношение (7.2) из этих свойств вытекает в силу неравенства Чебышёва:

$$\mathsf{P}\left(\left|\frac{1}{N}(X'\varepsilon)_j\right| \geq \varepsilon\right) \leq \frac{\mathbf{V}\left(\frac{1}{N}(X'\varepsilon)_j\right)}{\varepsilon^2} \to 0$$

(см. аналогичное рассуждение в параграфе 2.1 при выводе достаточных условий состоятельности).

Более общие, чем (7.1), предположения, обеспечивающие состоятельность оценок наименьших квадратов, так называемые условия Гренандера, можно найти в [19], гл.9.

Упражнение. Доказать, что при $k = 2$, $X_1 \equiv 1$, $X_{i2} = i$ ("время") оценки МНК состоятельны.

Перейдем теперь к стохастическим регрессорам. Простейший по формулировке вариант условий, гарантирующих состоятельность оценок МНК, — те же соотношения (7.1) и (7.2). Следует только уточнить, что в (7.1) предел понимается теперь по вероятности, а предельная матрица Q, помимо невырожденности, как правило, предполагается еще и неслучайной. Проверка (7.1) и (7.2) практически всегда опирается на подходящий вариант закона больших чисел для зависимых величин. В приложении C приводится утверждение такого типа, достаточное для многих применений. Проиллюстрируем его использование одним примером. Соотношение (7.1) в развернутом виде означает, что

$$\frac{1}{N}\sum_{i=1}^{N} X_{ij_1}X_{ij_2} \to q_{j_1j_2}$$

($1 \leq j_1, j_2 \leq k$). Эти соотношения похожи на законы больших чисел для последовательностей $\{X_{ij_1}X_{ij_2}\}_{i=1}^{\infty}$. Легко предложить условия, когда эти законы больших чисел будут справедливы. Вот один из вариантов таких условий:

1. существуют пределы $q_{j_1j_2} = \lim_{i\to\infty} \mathbf{E}(X_{ij_1}X_{ij_2})$;
2. четвертые моменты $\mathbf{E}(X_{ij}^4)$ ограничены в совокупности:

$$\mathbf{E}(X_{ij}^4) \leq c < \infty;$$

3. коэффициенты корреляции

$$\rho(X_{mj_1}X_{mj_2}, X_{nj_1}X_{nj_2})$$

стремятся к нулю при $|m-n| \to \infty$.

Условие 1 позволяет перейти к центрированным величинам, а условия 2 и 3 обеспечивают применимость теоремы из приложения С. Детали проверки мы оставляем читателям.

Включение в модель регрессора "время": $X_{i2} = i$, имеющего "нестационарный"характер, требует небольших дополнительных усилий. Мы на этом не останавливаемся.

Сделаем еще одно общее замечание о регрессорах $X_1, \cdots, X_k$. Исследователь находится перед выбором: либо они трактуются экзогенно, и тогда о них можно делать лишь предположения общего, формального характера (типа моментных условий 1 – 3, указанных выше), либо для них, в свою очередь, предполагаются какие-то более конкретные модели. Вторая возможность может привести к расширению исходной (основной) модели, она уже будет включать не одно, а несколько уравнений. Системы структурных регрессионных уравнений будут рассматриваться в главе 8. В качестве промежуточного варианта можно предложить следующее. Для регрессоров предполагается формальная (**не** структурная) модель, например, авторегрессионная. Такую модель можно тестировать (см. ниже параграф 5). Но используется эта модель лишь для мотивировки каких-либо общих свойств поведения регрессионной матрицы, например, для (частичного) обоснования условия 3, сформулированного выше.

Перейдем теперь к обсуждению соотношения (7.2). Для стохастических регрессоров оно приобретает самостоятельное значение. Фактически (7.2) утверждает, что регрессоры и ошибки асимптотически не коррелируют. Отсутствие этого свойства иногда означает, что модель неправильно специфицирована. В главе 8 мы увидим, что для отдельного уравнения, вырванного из структурной системы, такая корреляция объясняется связями, выраженными другими уравнениями системы. В любом случае отсутствие соотношения (7.2) почти предопределяет несостоятельность оценок наименьших квадратов и вынуждает искать другие методы оценивания коэффициентов. Мы еще будем возвращаться к обсуждению этих вопросов в различных контекстах.

Асимптотическая нормальность оценок параметров (см. главы 2 и 3) позволяет строить для этих параметров и доверительные интервалы

(также асимптотические). Эта методика применима и к оценкам метода наименьших квадратов. Вместо закона больших чисел при этом используется подходящий вариант центральной предельной теоремы.

Для неслучайных регрессоров и независимых наблюдений достаточно предположить существование и невырожденность предельной матрицы Q в (7.1). Тогда распределение нормированного отклонения $\sqrt{N}(\hat{\beta}-\beta)$ слабо сходится к нормальному распределению $\mathbf{N}(0,\sigma^2Q^{-1})$. Иначе это утверждение можно записать так: распределение величины $\frac{1}{\sqrt{N}}X'\varepsilon$ слабо сходится к $\mathbf{N}(0,\sigma^2Q)$.

Равносильность этих формулировок вытекает из (7.1) и формулы пересчета ковариационной матрицы при умножении вектора на (матричный) множитель:

$$\mathsf{cov}(Q^{-1}X'\varepsilon)=Q^{-1}\mathsf{cov}(X'\varepsilon)Q^{-1}.$$

Для доказательства второго варианта утверждения об асимптотической нормальности достаточно всего лишь сослаться на многомерную центральную предельную теорему для неодинаково распределенных слагаемых — теорему Линдеберга. Теоремы Леви, сформулированной в параграфе 1.4, здесь не хватает (она относится к iid величинам). Некоторые подробности, а также обобщения, относящиеся к стохастическим регрессорам, можно найти в книге [19], гл.9. Следует только иметь в виду, что ее автор не является специалистом по предельным теоремам, поэтому допускает иногда неточности исторического характера.

Так, он приписывает усиленный закон больших чисел для независимых неодинаково распределенных величин без дисперсии А.А.Маркову (1856 – 1922), скончавшемуся за несколько лет до того, как А.Н.Колмогоров в 1929 г. получил общую формулировку этого закона для неодинаково распределенных величин, да и то с конечными дисперсиями.

Достаточное условие сходимости к нормальному закону в предположении предельной пренебрегаемости отдельных слагаемых — так называемое условие Линдеберга (а не Линдберга, как его упорно называет Грин) было получено в 1922 г., задолго до работы В.Феллера, доказавшего (1935) его необходимость. Поэтому именовать указанное достаточное условие "теоремой Линдеберга-Феллера"попросту некорректно.

7.4 Совместное распределение ошибок и обобщенный метод наименьших квадратов

Ключевые свойства вектора ошибок ε, предполагавшиеся выполненными в главе 6, формулируются на языке моментов второго порядка — дисперсий (они считаются одинаковыми) и ковариаций (они нулевые) [2]. Коротко мы записывали это в виде $\mathsf{cov}(\varepsilon) = \sigma^2\mathbf{1}$. Для многих эконометрических моделей такая структура ковариационной матрицы оказывается неудовлетворительной. Поэтому мы будем рассматривать далее различные альтернативные специальные формы этой матрицы. Такие формы должны быть достаточно конкретными, ибо в общем случае матрица ковариаций включает $N(N+1)/2$ параметров — слишком много, чтобы их можно было содержательно оценить по N наблюдениям.

В двух наиболее распространенных случаях — временных рядов и пространственных данных — естественные предположения о форме матрицы $\mathsf{cov}(\varepsilon)$ оказываются различными.

Временной ряд, как правило, описывает эволюцию некоторой характеристики фиксированного объекта (фирмы, ценной бумаги и т.п.). В этом случае на первый план выступают связи, прежде всего, корреляционные, между последовательными значениями этой характеристики. Часто можно считать, что она (характеристика), а вместе с ней и ошибки в нашей модели, ведет себя стационарным образом. На языке моментов второго порядка эта стационарность (в теории случайных процессов используются термины "стационарность в широком (=слабом) смысле"или "стационарность второго порядка") означает инвариантность их (моментов) при сдвиге шкалы времени:

$$\mathsf{cov}(\varepsilon_{t_1}, \varepsilon_{t_2}) = \mathsf{cov}(\varepsilon_{t_1+h}, \varepsilon_{t_2+h}) \tag{7.3}$$

(целое число h интерпретируется как сдвиг времени).

Для последовательности $\{\varepsilon_t\}$, стационарной в широком смысле, ковариации представляются в виде

$$\mathsf{cov}(\varepsilon_{t_1}, \varepsilon_{t_2}) = \sigma^2\rho_{|t_1-t_2|},$$

где $\sigma^2 = \mathbf{V}(\varepsilon_t)$ (в силу (7.3) дисперсия не зависит от t), а

$$\rho_{|t_1-t_2|} = \rho(\varepsilon_{t_1}, \varepsilon_{t_2})$$

[2]Центрированность ошибок уже комментировалась в параграфе 6.2. Отказываться от этого предположения мы не собираемся.

— коэффициент корреляции между ε_{t_1} и ε_{t_2} (в силу (7.3) он действительно зависит только от расстояния $|t_1 - t_2|$ между двумя моментами времени).

В параграфе 7.5 мы будем обсуждать автокорреляционные модели ошибок, для которых коэффициенты корреляции ρ_k описываются при помощи фиксированного числа параметров.

Пространственные данные, напротив, обычно описывают характеристики различных объектов (фирм, ценных бумаг и т.п.) в один и тот же момент времени. В этом случае связями между этими объектами часто можно пренебречь и считать соответствующие ошибки некоррелированными: $\mathsf{cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = 0$ $(i_1 \neq i_2)$, однако, вообще говоря, разнораспределенными. В теории второго порядка эта разная распределенность будет проявляться через зависимость дисперсии $\mathbf{V}(\varepsilon_i)$ от номера наблюдения. Соответствующие модели ошибок мы будем рассматривать в параграфе 7.6.

Для панельных данных обычно используется некоторая комбинация идей, относящихся к временным рядам и пространственным данным — см. также параграф 7.7.

Во всех подобных ситуациях имеется общее ядро — матрица ковариаций $\mathsf{cov}(\varepsilon) = V$, зависящая от некоторого относительно небольшого набора параметров. Ее параметры следует оценивать наряду с коэффициентами β_j линейной регрессии.

Как и в главе 6, мы начнем с обсуждения процедуры оценивания коэффициентов линейной регрессии. Заметим сначала, что оценки наименьших квадратов $\hat{\beta} = (X'X)^{-1}X'Y$ являются несмещенными при любой матрице V, однако доказательство их эффективности (см. параграф 6.5) существенным образом зависело от предположения $V = \sigma^2\mathbf{1}$. Довольно легко привести примеры, когда оценки наименьших квадратов перестают быть эффективными — см. параграф 7.6. Подчеркнем однако, что они остаются интуитивно приемлемыми.

Что же касается оценки дисперсии ошибок, полученной в параграфе 6.6, то она, вообще говоря, может потерять всякий смысл (если отсутствует соответствующий параметр). Как следствие, эту оценку нет основания использовать и для других целей, например, для оценивания матрицы ковариаций $\mathsf{cov}(\hat{\beta})$. Для стационарных временных рядов, имеющих постоянную дисперсию, свойства этой оценки будут обсуждаться в параграфе 7.5.

Таким образом, важной задачей оказывается статистическая проверка классических предположений об ошибках. Если эти предположения нарушены, целесообразно использовать процедуры, отличающиеся от тех, которые изучались в главе 6. Одной из таких процедур является так называемый обобщенный метод наименьших квадратов (английская аббревиатура GLS — generalized least squares). Обсудим этот метод сначала в чисто учебной ситуации, когда предполагается, что матрица V известна (в реальных задачах такого, разумеется, не бывает) и невырождена.

Докажем, что найдется невырожденная матрица L, удовлетворяющая соотношению $V^{-1} = L'L$. Такая матрица не единственная, и мы приведем лишь один из способов ее нахождения.

В курсах линейной алгебры доказывается, что симметричную матрицу (а V, как и любая матрица ковариаций, симметрична) можно ортогональным преобразованием привести к диагональному виду. Это означает, что найдется такая ортогональная матрица U, что $U'VU = \Lambda$ диагональна. Поскольку V еще и положительно определена, диагональные элементы λ_{ii} матрицы Λ положительны. Определим положительный квадратный корень $\Lambda^{1/2} = diag(\lambda_{11}^{1/2}, \cdots, \lambda_{NN}^{1/2})$ и положим $L' = U\Lambda^{-1/2}$. Тогда $LVL' = \mathbf{1}$, $V = L^{-1}L'^{-1} = (L'L)^{-1}$ и $V^{-1} = L'L$.

Умножим основное соотношение нашей регрессионной модели $Y = X\beta + \varepsilon$ на матрицу L слева и обозначим $Y^* = LY$, $X^* = LX$, $\varepsilon^* = L\varepsilon$. Мы получаем новую модель $Y^* = X^*\beta + \varepsilon^*$ с теми же коэффициентами регрессии и ошибками, удовлетворяющими классическим предположениям. Действительно,

$$\mathsf{cov}(\varepsilon^*) = \mathbf{E}(\varepsilon^*\varepsilon^{*'}) = \mathbf{E}(L\varepsilon\varepsilon'L') = L\mathsf{cov}(\varepsilon)L' = LVL' = \mathbf{1}).$$

Эффективной линейной несмещенной оценкой вектора коэффициентов β по теореме Гаусса-Маркова является оценка

$$\hat{\beta}_{GLS} = (X^{*'}X^*)^{-1}X^{*'}Y^* = (X'L'LX)^{-1}X'L'LY = \\ = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

(она называется оценкой обобщенного метода наименьших квадратов).

Важно подчеркнуть, что запас линейных несмещенных оценок в исходной и преобразованной моделях одинаков. Поэтому и понятие эффективной линейной оценки одно и то же в обеих моделях.

Если дополнительно предполагать, что вектор ошибок ε распределен нормально, то и преобразованный вектор ε^* будет иметь нормальное распределение. В этом случае оценка $\hat{\beta}_{GLS}$ будет эффективна в классе всех (не обязательно линейных) несмещенных оценок (ср. с аналогичным результатом, упоминавшимся в параграфе 6.7).

Поскольку в реальных задачах матрица V неизвестна, процедура построения оценки вектора коэффициентов β усложняется. Обычно матрица V тем или иным способом оценивается, а затем в качестве оценки вектора β берется выражение

$$\hat{\beta}_{GLS} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y,$$

где $\hat{V}$ — оценка матрицы V. При этом свойство несмещенности (не говоря уже об эффективности), вообще говоря, пропадает, однако сама процедура оценивания остается вполне осмысленной. Конечно, свойства $\hat{\beta}_{GLS}$ во многом зависят от способа оценивания матрицы V. Мы еще будем возвращаться к обсуждению этих вопросов в следующих параграфах этой главы.

В учебниках по эконометрике изложенный вариант обобщенного метода наименьших квадратов иногда снабжается эпитетом "feasible"(русским переводом может быть слово "осуществимый"или "реализуемый"; в [9] используется не слишком удачный, на наш взгляд, термин "доступный").

Имеется один важный случай, когда при построении оценок $\hat{\beta}_{GLS}$ можно обойтись без предварительного оценивания матрицы ковариаций V. Это — случай, когда V известна с точностью до скалярного множителя: $V = \sigma^2 C$, где C — известная матрица. Действительно, выражение

$$(X'V^{-1}X)^{-1}X'V^{-1}Y$$

для оценок обобщенного метода наименьших квадратов в этих предположениях сводится к выражению

$$(X'C^{-1}X)^{-1}X'C^{-1}Y,$$

уже не содержащему неизвестный параметр σ^2. Тем самым, этот метод автоматически осуществим (feasible), и оценки $\hat{\beta}_{GLS}$ эффективны! Мы воспользуемся этим замечанием в параграфе 6.

Кроме процедуры обобщенного метода наименьших квадратов существуют и другие способы оценивания, основанные на общих

статистических принципах, например, на принципе максимального правдоподобия. Эти способы целесообразно обсуждать в более конкретных модельных предположениях об ошибках.

7.5 Авторегрессионные стационарные последовательности и корреляция ошибок

Последовательность случайных величин $\{\varepsilon_t\}$ называется авторегрессионной, если она удовлетворяет линейному рекуррентному уравнению с постоянными коэффициентами:

$$\varepsilon_t = \delta + \phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + u_t, \tag{7.4}$$

где $\{u_t\}$ — слабый белый шум. Как правило, предполагается (или неявно подразумевается), что вспомогательный белый шум $\{u_t\}$ "не коррелирует с прошлым", т.е. ковариации $\mathsf{cov}(u_t, \varepsilon_{t-1})$, $\mathsf{cov}(u_t, \varepsilon_{t-2})$,... равны нулю. И мы также будем придерживаться этого соглашения. В качестве моделей ошибок используются центрированные последовательности, поэтому в рамках настоящего параграфа мы предположим, что $\delta = 0$ и что все математические ожидания $\mathbf{E}\varepsilon_t$ также нулевые.

Можно дать естественную неформальную трактовку ошибок, подчиняющихся авторегрессионному соотношению. В каждый момент времени t ошибка включает составляющие, связанные с тем, что ранее возникшие источники ошибки продолжают действовать (в некотором измененном, часто можно считать — ослабленном, виде), и составляющую, описывающую дополнительные, только что возникшие, "сиюминутные", источники ошибки (имеется в виду белый шум u_t, не коррелирующий с прошлым).

Авторегрессионная модель ошибок включает в качестве параметров коэффициенты авторегрессии $\phi_1, \cdots, \phi_p$ и дисперсию σ_u^2 белого шума $\{u_t\}$. Порядок p авторегрессии также может варьироваться, хотя в моделях ошибок редко бывает большим.

Традиционный способ задания авторегрессионной последовательности — зафиксировать p подряд идущих ее членов и выразить через них все остальные при помощи рекуррентного соотношения. Например, можно зафиксировать $\varepsilon_0, \varepsilon_{-1}, \cdots, \varepsilon_{-p+1}$ и

написать

$$\varepsilon_1 = \phi_1\varepsilon_0 + \phi_2\varepsilon_{-1} + \cdots + \phi_p\varepsilon_{-p+1} + u_1,$$
$$\varepsilon_2 = \phi_1\varepsilon_1 + \phi_2\varepsilon_0 + \cdots + \phi_p\varepsilon_{-p+2} + u_2 =$$
$$= (\phi_1^2+\phi_2)\varepsilon_0+(\phi_1\phi_2+\phi_3)\varepsilon_{-1}+\cdots+(\phi_1\phi_{p-1}+\phi_p)\varepsilon_{-p+2}+\phi_1\phi_p\varepsilon_{-p+1}+\phi_1 u_1+u_2$$

и т.д. Аналогично можно найти и предыдущие члены последовательности. Для этого всего лишь надо переписать рекуррентное соотношение в виде

$$\varepsilon_{t-p} = \frac{1}{\phi_p}\varepsilon_t - \frac{\phi_1}{\phi_p}\varepsilon_{t-1} - \cdots - \frac{\phi_{p-1}}{\phi_p}\varepsilon_{t-p+1} - \frac{1}{\phi_p}u_t.$$

В общем случае авторегрессионная последовательность $\{\varepsilon_t\}$ не обладает свойством слабой стационарности. Более того, стационарные решения уравнения (7.4), отличные от нулевого, существуют не для всех наборов коэффициентов. Для выяснения этого вопроса выпишем уравнения, которым должны подчиняться дисперсия $\gamma_0 = \mathbf{E}(\varepsilon_t^2)$ и ковариации $\gamma_k = \mathbf{E}(\varepsilon_t\varepsilon_{t-k})$, $k \geq 1$, центрированной стационарной последовательности $\{\varepsilon_t\}$, удовлетворяющей авторегрессионному уравнению

$$\varepsilon_t = \phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + u_t \tag{7.5}$$

(общий случай уравнения (7.4) с $\delta \neq 0$ рассматривается почти так же).

Итак,

$$\gamma_0 = \mathbf{E}(\varepsilon_t^2) = \mathbf{E}[\varepsilon_t(\phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + u_t] =$$
$$= \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \mathbf{E}[(\phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + u_t)u_t]$$
$$= \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_u^2,$$
$$\gamma_1 = \mathbf{E}(\varepsilon_t\varepsilon_{t-1}) = \mathbf{E}[(\phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p}) + u_t)\varepsilon_{t-1}] =$$
$$= \phi_1\gamma_0 + \phi_2\gamma_1 + \cdots + \phi_p\gamma_{p-1},$$
$$\gamma_2 = \cdots = \phi_1\gamma_1 + \phi_2\gamma_0 + \phi_3\gamma_1 + \cdots + \phi_p\gamma_{p-2},$$
$$\gamma_p = \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + \cdots + \phi_p\gamma_0.$$

Выписанные уравнения образуют замкнутую систему из $p+1$ уравнений с таким же числом неизвестных. Они называются уравнениями Юла-Уолкера (Yule-Walker equations). Остальные ковариации рекуррентно находятся через $\gamma_0, \gamma_1, \cdots, \gamma_p$ при помощи аналогичных соотношений

$$\gamma_{p+k} = \mathbf{E}(\varepsilon_t\varepsilon_{t-p-k}) = \phi_1\gamma_{p+k-1} + \cdots + \phi_p\gamma_k \tag{7.6}$$

Легко установить, что даже в простейшем случае $p = 1$ уравнения Юла-Уолкера могут не иметь подходящего решения. Действительно, при $p = 1$ имеем

$$\gamma_0 = \phi_1\gamma_1 + \sigma_u^2, \gamma_1 = \phi_1\gamma_0, \tag{7.7}$$

откуда

$$\gamma_0 = \frac{\sigma_u^2}{1 - \phi_1^2}. \tag{7.8}$$

Если $|\phi_1| > 1$, то решение системы (7.7) не имеет вероятностного смысла (дисперсия γ_0 должна быть положительной), а при $|\phi_1| = 1$ решение вообще не существует. Если $|\phi_1| < 1$, мы получаем осмысленные выражения для γ_0 и γ_1. Более того, из (7.6) легко найти $\gamma_s = \phi_1^s\gamma_0$ $(s = 1, 2, \cdots)$. Последовательности $\{\varepsilon_t\}$ с такими ковариационными характеристиками действительно существуют. Построить $\{\varepsilon_t\}$ можно, опираясь на следующие наводящие соображения. Из формулы (7.5) при $p = 1$ следует, что

$$\varepsilon_t = \phi_1(\phi_1\varepsilon_{t-2} + u_{t-1}) + u_t = \cdots = u_t + \phi_1 u_{t-1} + \cdots + \phi_1^s u_{t-s} + \phi_1^{s+1}\varepsilon_{t-s-1}.$$

Формально устремляя $s \to \infty$, можно предположить, что

$$\varepsilon_t = \sum_{s=0}^{\infty} \phi_1^s u_{t-s}.$$

Нетрудно проверить, что последний ряд сходится (в среднем квадратичном) и что его сумма стационарна и удовлетворяет авторегрессионному соотношению, а потому имеет требуемые ковариации[3].

Для авторегрессии произвольного порядка p можно получить аналогичные условия существования стационарной последовательности, удовлетворяющей (7.5). Рассмотрим так называемое характеристическое уравнение

$$\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_{p-1}\lambda - \phi_p = 0.$$

Для того чтобы уравнение (7.5) имело стационарное (в слабом смысле) решение, необходимо и достаточно, чтобы все корни характеристического уравнения лежали в открытом единичном круге $\{\lambda \in \mathbb{C} : |\lambda| < 1\}$ плоскости комплексных чисел[4]. Мы не будем

[3]Мы не останавливаемся на этом подробно, поскольку обсуждение увело бы нас слишком в сторону от основной темы

[4]В этой формулировке предполагается, что белый шум $\{u_t\}$ не вырожден, т.е. $\sigma_u \neq 0$.

доказывать это утверждение. Отметим однако, что один из подходов к доказательству — обобщить рассуждения, изложенные выше для $p = 1$.

Совокупность стационарных последовательностей, удовлетворяющих (7.5) (или, более общим образом, (7.4)) часто обозначается $\mathsf{AR}(p)$.

Перейдем теперь к обсуждению свойств линейной регрессионной модели с ошибками класса $\mathsf{AR}(p)$, или, как еще говорят, с автокорреляцией ошибок порядка p. Помимо собственно свойств этой модели, следует обсудить вопросы о том, как выбрать p, и о том, есть ли вообще необходимость в допущении автокорреляции ошибок.

Мы уже указывали в предыдущем параграфе, что обычные оценки наименьших квадратов для коэффициентов β остаются несмещенными, хотя и перестают, вообще говоря, быть эффективными. Сейчас у нас есть возможность дополнить это обсуждение.

В предположениях настоящего параграфа главной характеристикой ошибок по-прежнему является дисперсия $\gamma_0 = \sigma_\varepsilon^2$, не зависящая от номера наблюдения. К сожалению, оценка этой дисперсии через сумму квадратов остатков, изучавшаяся в главе 6, перестает быть несмещенной. Более того, во многих типичных ситуациях смещение оказывается отрицательным (об этом можно прочитать в [4], гл.8, или в более позднем издании [23]; в [9], гл.6, излишне категорично утверждается, что смещение всегда отрицательно). Недооценка дисперсии ошибок может привести к разнообразным заблуждениям при реализации последующих статистических процедур, например, при определении статистической значимости коэффициентов регрессии.

Обратимся теперь к обобщенному методу наименьших квадратов (см. параграф 4). Если коэффициенты авторегрессии $\phi_1, \cdots, \phi_p$ считать известными (мы используем все тот же учебный прием — начать с более простой, хотя и нереалистичной, ситуации), оценки GLS можно получить следующим простым приемом. Введем новые величины ($t \geq p + 1$)

$$Y_t^* = Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p},$$
$$X_{tj}^* = X_{tj} - \phi_1 X_{t-1,j} - \cdots - \phi_p X_{t-p,j},$$
$$\varepsilon_t^* = \varepsilon_t - \phi_1 \varepsilon_{t-1} - \cdots - \phi_p \varepsilon_{t-p} = u_t.$$

Для них выполняется соотношение

$$Y_t^* = \beta_1 X_{t1}^* + \cdots + \beta_k X_{tk}^* + u_t$$

($t \geq p+1$), так что в модифицированной модели ошибки удовлетворяют классическим предположениям. К сожалению мы теряем при этом p

первых наблюдений, что во многих практических задачах нежелательно. Восполнить понесенные потери можно следующим образом. Положим

$$Y_1^* = C_{11}Y_1,$$
$$Y_2^* = C_{21}Y_1 + C_{22}Y_2,$$
$$\cdots$$
$$Y_p^* = C_{p1}Y_1 + C_{p2}Y_2 + \cdots + C_{pp}Y_p \quad (7.9)$$

и определим коэффициенты $C_{\cdot\cdot}$ так, чтобы дисперсии и ковариации этих модифицированных величин приняли требуемые значения:

$$\mathbf{E}(Y_i^{*2}) = \sigma_u^2$$

$(i = 1, \cdots, p)$,

$$\mathbf{E}(Y_{i_1}^* Y_{i_2}^*) = 0$$

$(1 \le i_1 < i_2 \le p)$. Треугольный характер соотношений (7.9) позволяет легко сделать это при малых p.

Пусть $p = 1$. Следует искать единственный коэффициент C_{11} из уравнения

$$C_{11}^2\gamma_0 = \sigma_u^2.$$

Знак C_{11} не имеет какого-либо значения, поэтому можно взять

$$C_{11} = \frac{\sigma_u}{\sqrt{\gamma_0}} = \frac{\sigma_u}{\sigma_\varepsilon} = \sqrt{1 - \phi_1^2}.$$

Аналогично, при $p = 2$ получаем

$$C_{11} = \frac{\sigma_u}{\sigma_\varepsilon},$$

а для C_{21} и C_{22} имеем уравнения

$$C_{21}^2\gamma_0 + 2C_{21}C_{22}\gamma_1 + C_{22}^2\gamma_0 = \sigma_u^2,$$
$$C_{11}C_{21}\gamma_0 + C_{11}C_{22}\gamma_1 = 0.$$

Из второго уравнения получаем

$$C_{21} = -\frac{\gamma_1}{\gamma_0}C_{22},$$

так что

$$C_{22}^2\frac{\gamma_0^2 - \gamma_1^2}{\gamma_0} = \sigma_u^2.$$

Отсюда

$$C_{22} = \frac{\sigma_u}{\sigma_\varepsilon}\frac{1}{\sqrt{1-\rho_1^2}}, \qquad C_{21} = -\frac{\sigma_u}{\sigma_\varepsilon}\frac{\rho_1}{\sqrt{1-\rho_1^2}},$$

где $\rho_1 = \gamma_1/\gamma_0$ — соответствующий коэффициент корреляции. Выразить эти коэффициенты $C_{\cdot\cdot}$ через ϕ_1 и ϕ_2 несколько сложнее. Легко проверить, что $\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\phi_1}{1-\phi_2}$. Выражение для $\frac{\sigma_u}{\sigma_\varepsilon}$ достаточно громоздко; мы его не приводим.

В учебной литературе обычно обсуждается простейшая регрессионная модель $\mathsf{AR}(1)$ для ошибок, в которой $\phi_1 = \rho_1$. Изложим популярную итеративную процедуру, позволяющую оценить в этом случае коэффициент ρ_1 (удобно трактовать его именно как коэффициент корреляции). Она называется процедурой Кохрейна-Оркатта[5] (Cochrane-Orcutt procedure).

На первом шаге коэффициенты β_j основной регрессии оцениваются обычным методом наименьших квадратов. Остатки $\hat\varepsilon_t$ этой регрессии используются на следующем шаге для получения оценки коэффициента ρ_1 из вспомогательного авторегрессионного уравнения (7.5) вида $\varepsilon_t = \rho_1\varepsilon_{t-1} + u_t$:

$$\hat\rho_1 = \frac{\sum_2^T \hat\varepsilon_{t-1}\hat\varepsilon_t}{\sum_2^T \hat\varepsilon_{t-1}^2} \tag{7.10}$$

(ср. с параграфом 6.4). На третьем шаге с помощью $\hat\rho_1$ делается преобразование модели, имитирующее описанный выше переход к некоррелированным ошибкам, и строятся оценки обобщенного метода наименьших квадратов для коэффициентов основной регрессии. На четвертом шаге остатки $\hat{\hat\varepsilon}_t$, полученные с помощью этих GLS-оценок, используются для нахождения следующего приближения $\hat{\hat\rho}_1$ для коэффициента автокорреляции и т.д. Принято считать, что этот итеративный процесс быстро сходится (в практическом смысле, т.е. с наперед заданной точностью) и что оценки последнего шага эффективнее первоначальных GLS-оценок. Корректную теоретическую постановку соответствующего вопроса не так легко дать, однако обсуждение этой проблемы выходит за рамки наших лекций.

Процедура Кохрейна–Оркатта почти непосредственно обобщается на $\mathsf{AR}(p)$-модель ошибок с произвольным p.

Известны (см., например, [24, 9]) и другие процедуры оценивания коэффициента автокорреляции, используемые в практических расчетах.

[5] D.Cochrane, не путать с известным статистиком Кокреном (W.G.Cochran)

Вернемся к обсуждению вопросов оценивания в модели с AR(1)-ошибками. Располагая оценками коэффициентов регрессии β_j и оценкой коэффициента автокорреляции ρ_1, можно оценить дисперсии σ_u^2 и σ_ε^2. Вспомогательная дисперсия σ_u^2 оценивается обычным образом через остатки, а дисперсия σ_ε^2 после этого с использованием соотношения (7.8). Если уж мы соглашаемся с оценками коэффициентов регрессии, мы, видимо, вынуждены согласиться и с оценкой дисперсии σ_ε^2.

Далее можно использовать эти оценки и для решения последующих задач, обсуждавшихся в гл.6, т.е. для построения доверительных интервалов и проверки гипотез о коэффициентах регрессии. На практическом уровне никаких изменений при этом не происходит, а теоретическое обоснование, как уже отмечалось выше, не входит в наши планы.

Следует выделить, однако, новую задачу — задачу выбора между двумя моделями. Одна из них — классическая модель, изучавшаяся в гл.6. Другая — модель с автокорреляцией ошибок, требующая использования других статистических приемов. Естественно взять в качестве основной гипотезу об отсутствии автокорреляции ошибок $\rho_1 = 0$ (мы продолжаем обсуждать простейшую схему автокорреляции первого порядка), а в качестве альтернативной — гипотезу $\rho_1 > 0$ (альтернатива $\rho_1 < 0$ рассматривается точно так же).

Разумной характеристикой корреляции ошибок является эмпирический коэффициент корреляции

$$r = \frac{\sum_2^T \hat{\varepsilon}_{t-1}\hat{\varepsilon}_t}{\sqrt{\sum_2^T \hat{\varepsilon}_{t-1}^2 \sum_2^T \hat{\varepsilon}_t^2}}$$

(это выражение отличается, хотя и незначительно, от (7.10)), однако чаще всего используется статистика DW, предложенная Дёрбином (Durbin) и Ватсоном (Watson) в 1950 г. ([16, 17, 18])[6]:

$$DW = \frac{\sum_2^T (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_1^T \hat{\varepsilon}_t^2}.$$

Пользуясь рассуждениями, аналогичными приведенным в параграфе 3, можно доказать, что при некоторых естественных предположениях $\hat{\rho}_1$ и r

[6]Дёрбин=Дурбин=Дарбин, Ватсон=Уотсон. Мы придерживаемся варианта, принятого в русском переводе книги Себера [10]. Как сообщил Я.Ю.Никитин (private communication), лично встречавшийся с Дёрбином, именно такое произношение его фамилии является правильным. Написание "Ватсон"соответствует традициям, преобладающим в математической литературе на русском языке.

— состоятельные оценки теоретического коэффициента корреляции ρ_1, а DW — состоятельная оценка величины $2(1-\rho_1)$. (Заинтересовавшийся читатель в качестве упражнения мог бы, предполагая состоятельность $\hat{\rho}_1$, найти расхождение между DW и $2(1-\hat{\rho}_1)$ и проверить, что оно стремится к нулю при $T \to \infty$.)

Дёрбин и Ватсон нашли определенные преимущества статистики DW, оказавшиеся весьма удобными для практических расчетов. Опишем схематично их результаты для задачи проверки основной гипотезы $H_0 : \rho_1 = 0$ против односторонней альтернативы $H_1 : \rho_1 > 0$ (альтернатива $\rho_1 < 0$ рассматривается совершенно аналогично, "зеркальным"образом). Прежде всего, они установили, что, несмотря на то, что распределение случайной величины DW при основной гипотезе H_0 зависит от регрессионной матрицы X, существуют случайные величины D^- и D^+, имеющие распределения, уже не зависящие от X, ограничивающие DW с двух сторон: $D^- \leq DW \leq D^+$. Эти распределения затабулированы, а процентные точки их традиционно обозначаются d_L и d_U (L — lower, U — upper). В терминах исходной статистики DW, предложенной Дёрбином и Ватсоном, критерий можно описать следующим образом. По уровню значимости ε определяются критические значения d_L и d_U, $0 < d_L < d_U < 2$, такие, что H_0 отвергается, если $DW < d_L$, и принимается, если $DW > d_U$. Промежуток $\langle d_L, d_U \rangle$ иногда называют зоной неопределенности. В этом случае Дёрбин и Ватсон предложили приближенные процедуры, которые "как будто весьма хорошо работают на практике"([10], с.165). Одна из этих процедур основана на наблюдении, что статистика $DW/4$ хорошо аппроксимируется бета-распределением с теми же математическим ожиданием и дисперсией (более подробно см. [10, 18]) . При двусторонней альтернативе $H_1 : \rho_1 \neq 0$ можно использовать "симметризованную"процедуру, выбрав критические значения d_L, d_U (и симметричные им $4-d_U$, $4-d_L$) по уровню значимости $\varepsilon/2$.

7.6 Неоднородные пространственные данные

Как уже отмечалось в параграфе 4, пространственные данные (мы возвращаемся к обозначению i для номера наблюдения) чаще всего можно считать некоррелированными. Неоднородность их при этом в теории второго порядка сводится к зависимости дисперсии ошибки от номера наблюдения: $\mathbf{E}(\varepsilon_i^2) = \sigma_i^2$. Такая неоднородность в учебниках

по эконометрике часто называется трудновыговариваемым словом "гетероскедастичность"(heteroscedasticity), в противоположность однородным, "гомоскедастичным"данным. Термин этот восходит к XIX веку, когда "скедастической линией"называли график условной дисперсии как функции условия[7]. В определенных отношениях эта терминология является анахронизмом, однако широко распространенным.

В общем случае дисперсий σ_i^2 слишком много, чтобы их можно было содержательно оценивать. Поэтому используются модельные представления с малым числом параметров. Такие модельные представления должны удовлетворять двум естественным требованиям — чтобы они имели содержательное (экономическое) объяснение и чтобы соответствующие параметры можно было удобным образом оценивать.

Мы рассмотрим сначала наиболее простую и наиболее известную схему такого рода, позволяющую без больших усилий пользоваться техникой наименьших квадратов. Именно, предположим, что изменение дисперсии σ_i^2 от наблюдения к наблюдению объясняется влиянием на нее регрессоров. Естественная форма такого влияния

$$\sigma_i^2 = \sigma^2 g(X_{i\cdot}), \tag{7.11}$$

где σ^2 — единственный параметр этого модельного представления, g — строго положительная функция, не содержащая каких-либо дополнительных свободных параметров, а $X_{i\cdot} = (X_{i1}, \cdots, X_{ik})$ — i-я строка регрессионной матрицы X (набор $(X_{i1}, \cdots, X_{ik})$ значений регрессоров в i-м наблюдении). В стандартных учебниках (см., например, [19, 9]) рассматривается частный случай (7.11), отвечающий квадратичной функции g (точнее, $g(x) = x^2$, в качестве аргумента g подставляется один из регрессоров, например, X_{i2}), однако общее представление (7.11) исследовать ничем не сложнее. Более того, можно даже допустить зависимость g от каких-нибудь дополнительных объясняющих величин $Z_{i\cdot}$, не выражающихся через X_i (впрочем, во многих случаях проще, видимо, включить эти дополнительные факторы в список регрессоров).

Поскольку матрица V ковариаций ошибок предположена известной с точностью до скалярного коэффициента σ^2 ($V = \sigma^2 C$, C — диагональная матрица), мы можем воспользоваться замечанием, сделанным в конце

[7]Это обстоятельство настолько забылось, что даже появился термин "условная гетероскедастичность".

параграфа 4, и сразу написать (эффективные и несмещенные) оценки обобщенного метода наименьших квадратов

$$\hat{\beta}_{GLS} = (X'C^{-1}X)^{-1}X'C^{-1}Y.$$

В нашем контексте (корреляция ошибок отсутствует) соответствующая процедура из параграфа 4 допускает очень простое толкование. Представление данных

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

мы преобразуем к виду

$$\frac{Y_i}{\sqrt{g(X_{i\cdot})}} = \beta_1 \frac{X_{i1}}{\sqrt{g(X_{i\cdot})}} + \cdots + \beta_k \frac{X_{ik}}{\sqrt{g(X_{i\cdot})}} + \frac{\varepsilon_i}{\sqrt{g(X_{i\cdot})}}.$$

Новая ошибка

$$\varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{g(X_{i\cdot})}}$$

имеет теперь постоянную дисперсию σ^2,

$$X_{ij}^* = \frac{X_{ij}}{\sqrt{g(X_{i\cdot})}}$$

рассматриваются как значения новых регрессоров, а

$$Y_i^* = \frac{Y_i}{\sqrt{g(X_{i\cdot})}}$$

— как значения новой объясняемой величины.

Чтобы оценить оставшийся параметр σ^2 — дисперсию ошибки преобразованного регрессионного уравнения, можно использовать обычную формулу

$$s^2 = \frac{\hat{\varepsilon}^{*'}\hat{\varepsilon}^*}{N-k}.$$

Как и в гл.6, эта оценка — несмещенная.

Рассмотрим теперь одну из реализаций более сложной схемы. Предположим, что дисперсии ошибок линейно выражаются через некоторые функции от регрессоров (а также, возможно, и еще некоторых наблюдаемых величин $Z_{i\cdot}$):

$$\sigma_i^2 = \theta_1 g_1(X_{i\cdot}) + \cdots + \theta_r g_r(X_{i\cdot}). \tag{7.12}$$

Можно предложить следующую последовательность действий. На первом этапе основное регрессионное уравнение оценивается обычным методом наименьших квадратов (напомним, что OLS-оценки остаются интуитивно приемлемыми, хотя и не обязательно эффективными, и в теперешней "гетероскедастичной"ситуации). Остатки $\hat{\varepsilon}_i$ этой регрессии используются на втором этапе для оценивания коэффициентов $\theta_1, \cdots, \theta_r$. Для этого формируется вспомогательная регрессия вида

$$\hat{\varepsilon}_i^2 = \theta_1 g_1(X_{i\cdot}) + \cdots + \theta_r g_r(X_{i\cdot}) + \nu_i. \tag{7.13}$$

Мы при этом исходим из ощущения сходства между интересующей нас дисперсией σ_i^2 и квадратом остатка — обе эти величины отражают, хотя и по-разному, степень разброса или вариативности в рамках нашей основной регрессионной модели.

Во вспомогательной регрессии $g_1(X_{i\cdot}), \cdots, g_r(X_{i\cdot})$ выступают в качестве объясняющих величин (вспомогательных регрессоров), а $\hat{\varepsilon}_i^2$ — в качестве вспомогательной объясняемой величины.

Оценки $\hat{\theta}_1, \cdots, \hat{\theta}_r$ обычного метода наименьших квадратов дают возможность предложить и оценки дисперсий (прогнозные значения, fitted values, для вспомогательной регрессии):

$$\hat{\sigma}_i^2 = \hat{\theta}_1 g_1(X_{i\cdot}) + \cdots + \hat{\theta}_r g_r(X_{i\cdot}).$$

На третьем этапе мы используем эти оценки для нахождения оценок $\hat{\beta}_{GLS}$ обобщенного метода наименьших квадратов. Можно надеяться, что эти оценки будут более эффективными, чем OLS-оценки.

При желании мы можем наш процесс продолжить — образовать новые остатки, с их помощью заново оценить коэффициенты $\theta_1, \cdots, \theta_r$ и т.д.

В некоторых частных случаях (один из них разбирается ниже) изложенная процедура дает состоятельные, хотя и смещенные оценки дисперсий.

Иллюстрацией данной процедуры является случай, когда дисперсия ошибки принимает только два значения (оба они, разумеется, считаются неизвестными).

Итак, предположим, что $\sigma_i^2 = A$ при $i = 1, \cdots, N_1$, $\sigma_i^2 = B$ при $i = N_1 + 1, \cdots, N_1 + N_2 = N$. Введем две индикаторные величины, I_1 и I_2, выделяющие эти значения:

$$I_{1i} = 1, \quad i \leq N_1, \quad I_{1i} = 0, \quad i > N_1, \quad I_2 = 1 - I_1.$$

С их помощью дисперсии σ_i^2 представляются в виде

$$\sigma_i^2 = AI_{1i} + BI_{2i}.$$

Отметим, что целесообразно ввести эти индикаторы в список регрессоров основной модели (вместо константы, если она там первоначально присутствовала). Из формул параграфа 6.4 легко получаем

$$\hat{A} = \frac{1}{N_1}\sum_{i=1}^{N_1}\hat{\varepsilon}_i^2, \quad \hat{B} = \frac{1}{N_2}\sum_{i=N_1+1}^{N}\hat{\varepsilon}_i^2.$$

Мы не будем обсуждать дальнейшие свойства этих оценок.

Замечание. Небольшие размышления подсказывают, что и представление (7.12) можно дальше обобщать, не меняя, по существу, рецептуру оценивания. Предположим, что

$$\sigma_i^2 = h(\theta_1 g_1(X_{i\cdot}, Z_{i\cdot}) + \cdots + \theta_r g_r(X_{i\cdot}, Z_{i\cdot}), X_{i\cdot}, Z_{i\cdot}), \tag{7.14}$$

где h — строго положительная функция, обратимая по первому аргументу. Пусть h^* — обратная (по первому аргументу) к h, так что

$$h^*(\sigma_i^2, X_{i\cdot}, Z_{i\cdot}) = \theta_1 g_1(X_{i\cdot}, Z_{i\cdot}) + \cdots + \theta_r g_r(X_{i\cdot}, Z_{i\cdot}).$$

Тогда, аналогично вспомогательной регрессии (7.13), можно рассмотреть регрессию $h^*(\hat{\varepsilon}_i^2, X_{i\cdot}, Z_{i\cdot})$ на набор регрессоров $g_1(X_{i\cdot}, Z_{i\cdot}), \cdots, g_r(X_{i\cdot}, Z_{i\cdot})$ и получить оценки $\hat{\theta}_1, \cdots, \hat{\theta}_r$ коэффициентов $\theta_1, \cdots, \theta_r$. После этого дисперсии σ_i^2 оцениваются естественным образом

$$\hat{\sigma}_i^2 = h(\hat{\theta}_1 g_1(X_{i\cdot}, Z_{i\cdot}) + \cdots + \hat{\theta}_r g_r(X_{i\cdot}, Z_{i\cdot}), X_{i\cdot}, Z_{i\cdot})$$

и т.д. В литературе (см., например, [25]) обсуждается, в частности, так называемая "мультипликативная форма"неоднородности, укладывающаяся в эту схему:

$$\sigma_i^2 = \exp(\theta_1 g_1 + \cdots + \theta_r g_r).$$

Обсудим теперь проблему выбора между двумя регрессионными моделями — однородной и неоднородной[8]. Большинство тестов, используемых при этом, проверяют основную гипотезу однородности против альтернативы, предполагающей ту или иную конкретную форму неоднородности.

[8]Право же, выражение "модель с гетероскедастичностью", которое можно встретить в учебниках, выглядит менее привлекательным.

Один из наиболее известных приемов, тест Голдфельда-Квандта (Goldfeld-Quandt test), используется в случае неоднородности вида (7.11):

$$\sigma_i^2 = \sigma^2 g(X_{i\cdot}, Z_{i\cdot}).$$

Наблюдения разбиваются на три группы — с "малыми", "средними"и "большими"значениями $g(X_{i\cdot}, Z_{i\cdot})$. Формально средняя группа не обязательна — она служит только для того, чтобы более резко отделить "большие"значения от "малых". Наблюдения средней группы просто отбрасываются. Обычно в учебниках приводятся "практические рекомендации", согласно которым в среднюю группу включаются от 15% до 20% из общего числа наблюдений. При этом крайние группы предполагаются примерно одинаковыми по размеру. Предположим для определенности, что $N = n_1 + n_2 + n_3$, где n_1, n_2, n_3 — численности групп, начиная с "малых"значений $g(X_{i\cdot}, Z_{i\cdot})$. Таким образом, первую группу составляют n_1 наблюдений с наименьшими значениями g, а третью — n_3 наблюдений с наибольшими значениями g.

Далее, отдельно в первой и третьей группах, оцениваются коэффициенты регрессии β обычным методом наименьших квадратов, а затем, также по обычной формуле (через остатки), дисперсия наблюдений отдельно взятой группы (т.е. так, как будто в пределах группы дисперсии одинаковы). Пусть σ_*^2 и σ_{***}^2 — полученные оценки дисперсий. В предположении справедливости основной гипотезы однородности отношение $\sigma_{***}^2/\sigma_*^2$ имеет (по крайней мере асимптотически) распределение Фишера $\mathbf{F}_{n_3-k,n_1-k}$. В предположении альтернативной гипотезы можно думать, что это отношение будет смещено вверх (вправо). Поэтому, выбрав по уровню значимости ε верхнюю критическую точку $\mathbf{F}$-распределения, мы получаем естественный рецепт — отвергать H_0, если отношение оцененных дисперсий превышает это критическое значение.

В более общей модели неоднородности (7.14) учебники рекомендуют BP-тест (Breusch-Pagan test). Опишем его схематически, следуя [25].

На первом шаге к исходной модели применяется обычный метод наименьших квадратов и строится величина

$$\hat{\sigma}^2 = \frac{1}{N}\sum \hat{\varepsilon}_i^2$$

(оценка максимального правдоподобия дисперсии в предположении однородности).

Затем образуются "нормированные"квадраты остатков $\hat{\varepsilon}_i^2/\hat{\sigma}^2$ и строится регрессия этих нормированных квадратов на набор вспомогательных регрессоров $g_1 \equiv 1, \cdots, g_r$ (см. (7.14); заметим, что наличие в (7.14) функции h никак не учитывается). Согласно [14] в случае однородных нормально распределенных ошибок регрессионная сумма квадратов RSS вспомогательной регрессии, деленная пополам, имеет асимптотически распределение χ^2_{r-1}. Большие значения величины $RSS/2$, по-видимому, указывают на нарушение основной гипотезы (возможно, в пользу (7.14); как указывает Грин [19], имеются основания считать, что BP-тест чувствителен к нарушениям предположения нормальности).

7.7 Панельные данные

Регрессионные модели, используемые для описания панельных данных, довольно разнообразны (см. [19]). Мы обозначим только некоторые идеи из этой области. Прежде всего следует отметить, что специфику подобных данных ("двумерный", в противоположность линейному, характер множества наблюдений) можно попробовать вообще не учитывать и пользоваться OLS. Однако при этом во многих случаях будут получаться неэффективные оценки. Поэтому разработка методов, специально ориентированных на панели, это, прежде всего, борьба за эффективность. Разумеется нельзя забывать и о том, что выбор спецификации модели — дело довольно тонкое, и это еще один повод к изучению подобных подходов.

Мы начнем с простого замечания об индикаторных величинах. Такие индикаторы имеют вполне отчетливый смысл. Одна категория индикаторов может описывать аддитивным образом отличия фирм или других подобных образований и не иметь отношения к временно́й динамике. Вторая категория индикаторов может описывать как раз изменения во времени, единые для всех фирм. Тогда мы получим следующую спецификацию

$$Y_{it} = \sum_{j=1}^{k} \beta_j X_{it,j} + \sum_{i'=2}^{N} \gamma_{i'} I_{i'}(i) + \sum_{t'=2}^{T} \delta_{t'} I_{t'}(t) + \varepsilon_{it}.$$

Здесь $I_{i'}$ — индикатор фирмы с номером i', а $I_{t'}$ — индикатор момента времени t'.

Кроме того, предполагается, что один из основных регрессоров (как обычно, первый) — константа. В противном случае следует суммировать по i' и t', начиная с единицы. Ошибки ε_{it} в простейшем случае предполагаются удовлетворяющими стандартным классическим условиям — образующими (слабый) белый шум.

Подобная спецификация, скорее всего, может возникнуть как альтернатива спецификации без индикаторов. Выбор между этими двумя вариантами можно сделать, проверяя гипотезу равенства нулю всех коэффициентов γ и δ. Если панель вытянута в одном из направлений (скажем, довольно часто встречаются задачи, в которых N много больше T), введение индикаторов приводит к значительным потерям в числе степеней свободы, а потому и в эффективности. Количество коэффициентов регрессии, очевидно, равно $k+(N-1)+(T-1)$, так что остается

$$NT - (k + N + T - 2) = (N - 1)(T - 1) - k + 1$$

степеней свободы (еще одна степень свободы позже расходуется на дисперсию ошибок). Разумеется, в каких-то задачах часть этих индикаторов не потребуется (возможно, придется проверять гипотезу о равенстве нулю группы коэффициентов), а коэффициенты при других индикаторах может оказаться целесообразным считать равными (опять же проверка линейной гипотезы, только чуть более общего вида).

Другая модель, которую мы рассмотрим, трактует влияние номера фирмы и номера момента времени стохастически, через их вклад в ошибку. Более точно, можно рассмотреть следующую спецификацию (error-components model):

$$Y_{it} = \sum_{j=1}^{k} \beta_j X_{it,j} + \varepsilon_{it},$$

где

$$\varepsilon_{it} = u_i + v_t + w_{it}.$$

Предполагается, что компоненты u, v и w ошибки ε являются белыми шумами, не коррелирующими между собой.

В этой модели число коэффициентов регрессии остается равным k, число дополнительно возникающих параметров — два (вместо одной дисперсии σ^2 появляются три — σ_u^2, σ_v^2 и σ_w^2). Ошибки ε_{it} становятся

в известном смысле коррелированными:

$$\mathbf{E}(\varepsilon_{i_1 t}\varepsilon_{i_2 t}) = \sigma_v^2 \quad (i_1 \neq i_2),$$

$$\mathbf{E}(\varepsilon_{i t_1}\varepsilon_{i t_2}) = \sigma_u^2 \quad (t_1 \neq t_2).$$

Оценив дисперсии компонент ошибки (читатель может сам поизобретать такие методы — здесь широкое поле для фантазии), мы сможем применить обобщенный метод наименьших квадратов и получить для коэффициентов регрессии оценки, которые, можно надеяться, асимптотически окажутся эффективнее оценок обычного метода наименьших квадратов.

Более сложные (и, может быть, более реалистичные) модели, которые мы лишь упомянем, включают (ср. с параграфами 7.5 и 7.6) автокорреляцию ошибок (в направлении t) и/или неодинаковость дисперсий (в направлении i). Для их оценивания может использоваться обобщенный метод наименьших квадратов.

7.8 Корреляция между регрессорами и ошибками

При обсуждении стохастических регрессоров в параграфе 1 мы предполагали, что $\mathbf{E}(\varepsilon_i|X) = 0$, $\mathbf{E}(\varepsilon_i^2|X) = \sigma^2$ и $\mathbf{E}(\varepsilon_{i_1}\varepsilon_{i_2}|X) = 0$ $(i_1 \neq i_2)$. Эти соотношения, вообще говоря, нарушаются, если допустить корреляцию (или более сложную зависимость) между ошибками и (не обязательно всеми) регрессорами, а оценки наименьших квадратов оказываются тогда смещенными.

Если допустить, что указанная корреляция сохраняется и асимптотически, то оценки эти окажутся и несостоятельными (во всяком случае нет особых причин считать их состоятельными). Мы сейчас рассмотрим наиболее распространенную "двухшаговую"процедуру, дающую состоятельные (и, в некотором смысле, оптимальные) оценки.

Начинается построение таких оценок с нахождения специфических вспомогательных величин, которые мы будем называть первичными инструментами и обозначать $Z_1, \cdots, Z_l$. Как правило, в число первичных инструментов включаются все регрессоры, не коррелирующие с ошибками (в частности, константа). Общее число инструментов должно быть не меньше числа основных регрессоров (более точно это описано ниже). Где искать недостающие инструменты — иногда непростой вопрос. Наиболее важный пример их связан с системами регрессионных уравнений и будет рассматриваться в следующей главе.

Главные свойства первичных инструментов (они по мере обсуждения будут уточняться) — отсутствие корреляции с ошибками и, напротив, наличие корреляции с основными регрессорами. Ясно, что основные регрессоры, не коррелирующие с ошибками, удовлетворяют и второму свойству — коррелируют сами с собой (этим и объясняется то обстоятельство, что их включают в список первичных инструментов).

На первом шаге описываемой процедуры строятся регрессии всех основных регрессоров $X_{\cdot j}$ на полный набор первичных инструментов $Z_1, \cdots, Z_l$. При этом используется обычный метод наименьших квадратов. Соответствующие прогнозные (fitted) значения $\hat{X}_{\cdot j}$ называются целевыми инструментами, отвечающими основным регрессорам. Очевидно, что для регрессоров, не коррелирующих с ошибками (т.е. входящих в список первичных инструментов) $\hat{X}_{\cdot j} = X_{\cdot j}$ — прогнозировать какую-либо величину по информации, содержащей ее саму, занятие банальное. Для остальных регрессоров целевые инструменты можно представлять себе как карикатуры на них, главное достоинство которых — отсутствие корреляций с ошибками (это утверждение на самом деле справедливо только асимптотически; оно вытекает из уточненных предположений об инструментах, которые сделаны ниже).

На втором шаге строится регрессия объясняемой величины Y на набор целевых инструментов (т.е. регрессия, которую можно рассматривать как карикатуру на исходную основную модель).

Как мы сейчас увидим, при достаточно разумном уточнении предположений об инструментах оценки наименьших квадратов для этой регрессии оказываются не карикатурными, а состоятельными. Для того чтобы выяснить это, введем сначала необходимые обозначения.

Матрицу наблюдений первичных инструментов мы обозначим (а как иначе?) Z. У нее N строк (по числу наблюдений) и l столбцов (по числу инструментов). Матрицу прогнозных значений, состоящую из столбцов $\hat{X}_{\cdot j}$, $j = 1, \cdots, k$, мы обозначим $\hat{X}$. Легко сообразить, что всю совокупность регрессий первого шага можно записать единым образом в матричной форме:

$$\hat{X} = Z(Z'Z)^{-1}Z'X.$$

Отдельной регрессии при этом соответствует аналогичное соотношение

$$\hat{X}_j = Z(Z'Z)^{-1}Z'X_j.$$

На втором шаге матрица $\hat{X}$ используется в качестве регрессионной, так

что оценки коэффициентов основной регрессии, предлагаемые описанной процедурой, имеют вид

$$\tilde{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y.$$

Если $k = l$, то размеры матриц X и Z совпадают, и формула значительно упрощается:

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

Разумеется, проводя формальные преобразования, мы всюду предполагали, что возникающие обратные матрицы существуют. Теперь пришло время сформулировать условия, которые эту обратимость обеспечивают.

Это делается почти аналогично тому, как вводились в параграфе 3 условия, дававшие в тех предположениях состоятельность оценок наименьших квадратов. Все пределы в написанных ниже соотношениях понимаются как пределы по вероятности.

Первое условие — условие невырожденности совокупности первичных инструментов:

$$\lim_{N\to\infty} \frac{1}{N} Z'Z = Q_{ZZ},$$

где Q_{ZZ} — невырожденная матрица. Как и в параграфе 3 предельная матрица Q_{ZZ} предполагается неслучайной.

Второе условие относится к взаимоотношениям первичных инструментов и ошибок — надлежит обеспечить отсутствие корреляций (хотя бы в асимптотическом смысле):

$$\frac{1}{N} Z'\varepsilon \to_{N\to\infty} 0.$$

Наконец, третье условие обеспечивает асимптотическую невырожденность матрицы целевых инструментов и, тем самым, корректность второго шага описанной процедуры:

$$\frac{1}{N} Z'X \to_{N\to\infty} Q_{ZX},$$

где Q_{ZX} — матрица полного ранга (т.е. ранга k — вспомним, что $l \geq k$). И здесь матрица Q_{ZX} предполагается неслучайной.

Для доказательства состоятельности вектора $\tilde{\beta}$ оценок заметим, что

подстановка $Y = X\beta + \varepsilon$ дает

$$\tilde{\beta} = \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\varepsilon =$$
$$\beta + \left(\frac{1}{N}X'Z\left(\frac{1}{N}Z'Z\right)^{-1}\frac{1}{N}Z'X\right)^{-1}\frac{1}{N}X'Z\left(\frac{1}{N}Z'Z\right)^{-1}\frac{1}{N}Z'\varepsilon.$$

При $N \to \infty$

$$\frac{1}{N}X'Z\left(\frac{1}{N}Z'Z\right)^{-1}\frac{1}{N}Z'X \to Q'_{ZX}Q_{ZZ}^{-1}Q_{ZX}$$

и предельная матрица имеет порядок и ранг k. Поэтому обратная к допредельной матрице существует при достаточно больших N и стремится к $(Q'_{ZX}Q_{ZZ}^{-1}Q_{ZX})^{-1}$, т.е. к конечному пределу. Аналогично существует конечный предел

$$\lim \frac{1}{N}X'Z\left(\frac{1}{N}Z'Z\right)^{-1} = Q'_{ZX}Q_{ZZ}^{-1}.$$

В то же время

$$\lim \frac{1}{N}Z'\varepsilon = 0.$$

Вычисляя предел произведения, получаем

$$\tilde{\beta} - \beta \to_{N\to\infty} 0$$

по вероятности.

Еще раз подчеркнем, что без конкретных примеров изложенные идеи повисают в воздухе. Наиболее важные примеры появятся в следующей главе. Там же мы рассмотрим и вопрос обнаружения корреляции между регрессорами и ошибками.

Глава 8

Системы регрессионных уравнений

До сих пор мы предполагали, что содержательные экономические теории позволяют выделить такой относительно замкнутый фрагмент большого экономического мира, который можно описать одним уравнением. Это предположение далеко не всегда выполняется. Дело в том, что "относительная замкнутость", упомянутая в предыдущей фразе, означает не просто возможность написания такого уравнения, но и возможность удовлетворить предположениям, которые делались для обеспечения осмысленности тех или иных статистических процедур.

Более широкие возможности открывают эконометрические модели, включающие несколько уравнений (см. примеры подобных моделей в главе 5). Некоторые приемы, реализующие эти возможности, будут описаны ниже.

8.1 Системы уравнений как источник первичных инструментов

В этом параграфе мы описываем некоторые общие идеи, связанные с оцениванием коэффициентов (и вообще параметров) в системах регрессионных уравнений. Главная трудность состоит в том, что отдельно взятое уравнение, как правило, не удовлетворяет стандартным предположениям (см. гл.7). Обычная запись уравнений, в которой слева стоит объясняемая (эндогенная, внутренняя) величина, а справа — объясняющие, также во многом должна быть уточнена.

Действительно, эндогенных величин в системе уравнений столько же, сколько уравнений, а потому в отдельно взятом уравнении их вполне может оказаться (и оказывается) несколько. Те из них, которые находятся в правой части (мы дальше обсудим на примере вопрос о

том, какие из них следует помещать налево, а какие — направо), в пределах нашего отдельно взятого уравнения похожи на регрессоры, но, вообще говоря, коррелируют с ошибками. Тем самым, стандартные предположения не выполнены. В параграфе 7.8 отмеченная выше корреляция уже обсуждалась, и была описана схема, позволяющая с этой трудностью справляться. Теперь мы можем уточнить эту процедуру, сказав, что первичные инструменты в теперешнем контексте систем регрессионных уравнений возникают естественным путем — в качестве них берутся объясняющие величины, которые фигурируют в остальных уравнениях системы. Формальная процедура оценивания, реализующая эту идею, — двухшаговый метод наименьших квадратов — будет изложена в следующем параграфе.

Разумеется, первичных инструментов, требующихся для этого метода, должно найтись достаточное количество. В противном случае уравнение называется неидентифицируемым. Его коэффициенты (по крайней мере, некоторые) оценить указанной процедурой не удается. Неидентифицируемость уравнения чаще всего свидетельствует о каких-то трудностях содержательного (экономического) характера.

При использовании двухшагового метода наименьших квадратов выделяется одно из уравнений системы, а остальные уравнения учитываются лишь формально — для нахождения инструментов. Известны более сложные процедуры, в которых вся система исследуется как единое целое. Эти методы значительно более чувствительны к ошибкам спецификации модели.

8.2 Двухшаговый метод наименьших квадратов

Рассмотрим одно из уравнений системы, записанное в виде

$$Y = \beta_1 X_1 + \cdots + \beta_k X_k + \gamma_1 Y_1^* + \cdots + \gamma_m Y_m^* + \varepsilon. \tag{8.1}$$

Здесь $X_1, \cdots, X_k$ — предопределенные величины, которые не коррелируют с ошибкой ε, а $Y_1^*, \cdots, Y_m^*$, так же как и Y, эндогенные величины, объясняемые моделью (т.е. всей системой). Поскольку мы будем обсуждать только оценивание коэффициентов отдельно взятого уравнения, нумерацию величин можно можно приспособить к этой локальной цели и избежать громоздких обозначений, возникающих при обсуждении всей системы (они нам не понадобятся).

Полную совокупность предопределенных величин системы мы обозначим $Z_1, \cdots, Z_l$. Без ограничения общности можно считать, что $Z_1 = X_1, \cdots, Z_k = X_k$.

Величины $Z_1, \cdots, Z_l$ будем рассматривать как первичные инструменты. Действуя по схеме параграфа 7.8, на первом шаге построим регрессии величин $X_1, \cdots, X_k,\ Y_1^*, \cdots, Y_m^*$, входящих в правую часть уравнения (8.1), на полный набор первичных инструментов $Z_1, \cdots, Z_l$ и получим, тем самым, целевые инструменты $\hat{X}_1, \cdots, \hat{X}_k$, $\hat{Y}_1^*, \cdots, \hat{Y}_m^*$. При этом, по очевидным причинам,

$$\hat{X}_1 = X_1(= Z_1), \cdots, \hat{X}_k = X_k(= Z_k),$$

так что, собственно говоря, эти регрессии и строить не нужно. На втором шаге процедуры построим регрессию величины Y на набор построенных целевых инструментов. Коэффициенты этой регрессии и будут оценками двухшагового метода наименьших квадратов.

При выполнении предположений, обсуждавшихся в параграфе 7.8, они состоятельны.

8.3 Структурные и приведенные системы. Косвенный метод наименьших квадратов

Для более ясного представления о месте двухшагового метода наименьших квадратов в теории систем регрессионных уравнений рассмотрим некоторые возможные альтернативы и выявим их минусы и плюсы. Мы уже отмечали, что основным аргументом, вызвавшим появление двухшаговой процедуры предыдущего параграфа, является вхождение **нескольких** эндогенных величин в рассматриваемое уравнение. Каждая из них, вообще говоря, коррелирует с ошибкой. Вырывая отдельное уравнение из системы, мы только одну из них можем поместить в левую часть уравнения (т.е. трактовать как объясняемую). Остальные эндогенные величины, входящие в уравнение, при этом трактуются как регрессоры, коррелирующие с ошибкой.

Можно попробовать исключить из нашего уравнения остальные эндогенные величины с помощью остальных уравнений системы. Посмотрим, к чему приведет эта идея.

Итак, в нашем распоряжении имеется первоначальная система линейных уравнений (нелинейные системы мы не рассматриваем),

написанная из тех или иных содержательных экономических соображений, т.е. выражающая определенный фрагмент экономической теории. Системы, возникающие подобным образом, принято называть структурными. Будем считать, что количество уравнений совпадает с количеством эндогенных величин, входящих в систему (это предположение в основном согласуется со здравым смыслом, другие возможности читатель может продумать самостоятельно). Хорошо известно, что система линейных уравнений, в которой число уравнений совпадает с числом неизвестных, обычно имеет единственное решение. Этот случай мы и рассмотрим (обдумывание и интерпретация других возможностей снова предоставляется читателю). "Решая"структурную систему относительно эндогенных величин, мы получаем выражения для них через остальные (т.е. предопределенные) величины. Слово "решая"мы намеренно заключили в кавычки. Дело в том, что все (или, по крайней мере, большинство) коэффициентов первоначальной структурной системы — неизвестные параметры. Поэтому коэффициенты приведенной системы также неизвестны, хотя можно написать формулы, выражающие их через структурные коэффициенты.

Уравнения приведенной системы можно оценивать по отдельности, ибо предопределенные величины с ошибками не коррелируют. Заметим однако, что при переходе от структурной системы к приведенной ошибки первоначальной системы "смешиваются". В то же время предположения об ошибках обычно формулируются и обосновываются в структурных терминах. Какими окажутся при этом свойства ошибок приведенных уравнений, определяется тем процессом "решения", который дает приведенные уравнения. Тем не менее, во многих случаях можно считать, что ошибки приведенных уравнений удовлетворяют классическим предположениям главы 6 (для нас сейчас это не главное).

Предположим, что мы смогли обычным методом наименьших квадратов состоятельным образом оценить коэффициенты приведенных уравнений. Что дальше? Это зависит от целей эконометрического исследования. Если нам нужно лишь определить прогнозные значения эндогенных величин, цель фактически достигнута — остается лишь воспользоваться оценками приведенных коэффициентов. Возвращаться к исходной структурной системе и ее коэффициентам уже не нужно. Если же нас действительно интересуют структурные коэффициенты, то их придется восстанавливать по приведенным (точнее, по оценкам

наименьших квадратов для них). Для этого соотношения между структурными и приведенными коэффициентами нужно решить относительно структурных коэффициентов. Этот прием называется непрямым или косвенным (indirect) методом наименьших квадратов (для оценивания структурных коэффициентов).

К сожалению на деле все оказывается не так просто. Во-первых, количество приведенных коэффициентов может отличаться от количества структурных. Во-вторых, соотношения, их связывающие, отнюдь не являются линейными. Как следствие, все потенциальные трудности, связанные с решением систем уравнений, могут возникнуть.

Опишем различные возможности:

1. Существует единственный набор структурных коэффициентов, соответствующий (оцененным) значениям приведенных коэффициентов; тогда структурная система называется точно идентифицируемой (exactly identifiable) или даже (в зависимости от контекста) точно идентифицированной (exactly identified).

2. Существует более одного набора структурных коэффициентов, соответствующих данным значениям приведенных коэффициентов; тогда структурная система называется неидентифицируемой (unidentifiable). Для некоторых структурных коэффициентов, тем не менее, все эти наборы могут дать одно и то же значение. Такие коэффициенты следует называть идентифицируемыми. Остальные коэффициенты — неидентифицируемыми. Точно так же могут оказаться идентифицируемыми отдельные уравнения структурной системы.

3. Не существует ни одного набора структурных коэффициентов, соответствующего данным значениям приведенных коэффициентов. В этом случае структурная система называется сверхидентифицируемой (overidentifiable).

Последняя возможность наиболее интересна. Она возникает в том случае, когда, грубо говоря, структурных коэффициентов меньше, чем приведенных. Более точно, можно сказать так. Уравнений для нахождения структурных коэффициентов через приведенные слишком много, и система их противоречива. Однако это обстоятельство не лишает смысла исходную регрессионную модель. При использовании метода наименьших квадратов стохастические

ошибки (они ненаблюдаемы) в некотором смысле игнорируются. Поэтому статистические процедуры позволяют найти лишь оценки коэффициентов, а не сами коэффициенты. При этом регрессионные уравнения выполняются (как и положено по исходным предположениям) только приблизительно, с точностью до остатка (оцененной ошибки). С практической точки зрения можно поступать так: некоторые из уравнений, связывающих приведенные коэффициенты со структурными, отбросить и искать оценки структурных коэффициентов из остальных уравнений. Это отбрасывание можно делать по-разному и получать разные оценки одних и тех же структурных коэффициентов. Все такие оценки вполне осмыслены и их можно объявить оценками непрямого (косвенного) метода наименьших квадратов. Можно брать и подходящие линейные комбинации их. Таким образом, косвенный метод неоднозначен. Напротив, двухшаговый метод наименьших квадратов дает однозначный рецепт, который, как можно увидеть на примерах, в некотором смысле оптимален. Мы не останавливаемся на этом более подробно, однако в следующем параграфе детально разберем один важный пример.

8.4 Простейшие модели спроса и предложения

Мы несколько изменим модель примера 1 из параграфа 5.2. Во-первых, воспользуемся условием равновесия для уменьшения числа уравнений (подобное действие всегда предшествует процедурам оценивания). Во-вторых, добавим в уравнение предложения еще одну экзогенную величину T, имеющую смысл температуры воздуха (некоторое усредненное значение для данного цикла). В-третьих, для удобства заменим обозначения p, q и r соответствующими заглавными буквами, сохраняя малые буквы для принятого обозначения отклонений от средних. Тем самым, будем рассматривать систему:

$$Q = \beta_1 + \beta_2 P + \gamma_1 I + \varepsilon^D,$$

$$Q = \beta_3 + \beta_4 P + \gamma_2 R + \gamma_3 T + \varepsilon^S.$$

Предположим также, что между экзогенными величинами нет коллинеарности или мультиколлинеарности.

При выбранной записи оба уравнения содержат Q в левой части. Как будет видно, это не слишком принципиальное обстоятельство, хотя в

теоретических исследованиях обычно считают, что в каждом уравнении в левой части стоит своя эндогенная величина, т.е. устанавливают некое однозначное в обе стороны соответствие между уравнениями и объясняемыми величинами.

Мы сделаем еще одно стандартное действие — перейдем к отклонениям от средних, и запишем наши уравнения в виде

$$q = \beta_2 p + \gamma_1 i + \varepsilon^D,$$

$$q = \beta_4 p + \gamma_2 r + \gamma_3 t + \varepsilon^S$$

(заметим, что ошибки при этом переходе изменяются, хотя мы и сохранили для них старые обозначения; меняются и свойства ошибок вдоль серии наблюдений, впрочем, асимптотически это обстоятельство несущественно). Вопрос об оценивании свободных членов β_1 и β_3 мы для краткости опустим.

Двухшаговый метод наименьших квадратов действует следующим образом. На первом шаге первичные инструменты i, r, t используются для получения целевого инструмента

$$\hat{p} = \hat{\pi}_1 i + \hat{\pi}_2 r + \hat{\pi}_3 t,$$

заменяющего p на втором шаге. На этом втором шаге для оценивания уравнения спроса строится регрессия q на $\hat{p}$ и i, а для оценивания уравнения предложения — регрессия q на $\hat{p}$, r и t. Обе эти регрессии действительно можно построить, т.к. в первом уравнении отсутствуют r и t, а между $\hat{p}$ и i коллинеарности нет, и, аналогично, во втором уравнении отсутствует i, а между $\hat{p}$, r и t коллинеарности нет.

Мы видим, что отсутствие некоторых экзогенных величин в отдельно взятом уравнении — важное обстоятельство; если бы в качестве регрессоров использовались все четыре величины $\hat{p}$, i, r и t, возникла бы очевидная коллинеарность. Эти соображения, действующие и в общем случае, обычно формулируют в виде так называемого порядкового условия идентифицируемости: число отсутствующих в уравнении предопределенных величин (в нашем примере все они экзогенны) не меньше числа эндогенных величин, присутствующих в правой части (мы считаем при этом, что еще одна эндогенная величина стоит слева). Порядковое условие по своему смыслу аналогично сходному условию в теории систем линейных уравнений и, так же как и последнее, не учитывает некоторых тонкостей. Необходимое и достаточное условие, имеющее ранговый характер, мы не приводим (см., например, [24]).

Посмотрим теперь, как выглядит в нашем контексте косвенный метод наименьших квадратов (опять ограничимся уравнениями в отклонениях).

Выпишем приведенные уравнения:

$$p = \frac{1}{\beta_4 - \beta_2}\left[\gamma_1 i - \gamma_2 r - \gamma_3 t + \varepsilon^D - \varepsilon^S\right],$$

$$q = \frac{1}{\beta_4 - \beta_2}\left[\beta_4\gamma_1 i - \beta_2\gamma_2 r - \beta_2\gamma_3 t + \beta_4\varepsilon^D - \beta_2\varepsilon^S\right]$$

или

$$p = \pi_{11} i + \pi_{12} r + \pi_{13} t + u,$$

$$q = \pi_{21} i + \pi_{22} r + \pi_{23} t + v,$$

где $\pi_{..}$ — приведенные коэффициенты, а u и v — новые ошибки. Особый случай $\beta_2 = \beta_4$ пока не будем обсуждать. Поскольку приведенных коэффициентов 6, а структурных — только 5, возникает сверхидентифицируемость. Легко предположить, что она связана с уравнением спроса, в котором отсутствует "слишком много"экзогенных величин — порядковое условие выполнено в виде строгого неравенства.

Выпишем соотношения, восстанавливающие структурные коэффициенты через приведенные:

$$\beta_4 = \frac{\pi_{21}}{\pi_{11}}, \quad \beta_2 = \frac{\pi_{22}}{\pi_{12}}, \quad \beta_2 = \frac{\pi_{23}}{\pi_{13}},$$

$$\gamma_1 = \pi_{11}(\beta_4 - \beta_2), \quad \gamma_2 = -\pi_{12}(\beta_4 - \beta_2), \quad \gamma_3 = -\pi_{13}(\beta_4 - \beta_2).$$

Единственный точно идентифицируемый коэффициент — коэффициент β_4, а для β_2 и, как следствие, для γ_1, γ_2, γ_3 возможны разные представления (заметим, что мы выписали не все представления для γ_2, γ_3).

Интересно отметить, что только один коэффициент уравнения предложения точно идентифицируем, так что сделанное ранее предположение о том, что сверхидентифицируемость связана с уравнением спроса, не вполне точно.

В более "короткой"системе, где температура T отсутствует, все коэффициенты точно идентифицируемы. Довольно простая выкладка показывает, что оценки двухшагового и косвенного методов при этом совпадают (мы ее не приводим, т.к. ниже разбирается более интересный, хотя и более сложный, результат).

Вернемся к нашей основной системе и докажем, что оценки двухшагового и косвенного методов для точно идентифицируемого коэффициента β_4 совпадают. Прежде всего заметим, что $\hat{\pi}_1 = \hat{\pi}_{11}$, $\hat{\pi}_2 = \hat{\pi}_{12}$, $\hat{\pi}_3 = \hat{\pi}_{13}$ (в обоих случаях строится регрессия p на набор регрессоров i, r, t). Далее, выражения для $\hat{\pi}_{21}$, $\hat{\pi}_{22}$, $\hat{\pi}_{23}$ получаются из выражений для $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\pi}_3$ заменой p на q.

Выпишем оценки двухшагового метода. Для этого нам потребуется специальное обозначение. Пусть X и Z — две (прямоугольные) матрицы одинакового размера, $X_1, \cdots, X_k$, $Z_1, \cdots, Z_k$ — их столбцы. Тогда матрица $Z'X$ имеет вид

$$Z'X = \begin{pmatrix} Z_1'X_1 & \dots & Z_1'X_k \\ \vdots & \ddots & \vdots \\ Z_k'X_1 & \dots & Z_k'X_k \end{pmatrix},$$

т.е. является функцией от векторов $X_1, \cdots, X_k$, $Z_1, \cdots, Z_k$. Наше обозначение — это обозначение для определителя матрицы $Z'X$:

$$\det(Z'X) = D(Z_1, \cdots, Z_k; X_1, \cdots, X_k).$$

Заметим, что функция D линейна по каждому из своих $2k$ аргументов, обращается в 0, если два аргумента $Z_\cdot$ или два аргумента $X_\cdot$ совпадают, а также меняет знак при перестановке двух аргументов из одной группы.

При помощи введенного обозначения оценки наименьших квадратов (X и Y — стандартные обозначения регрессионной матрицы и объясняемой величины) можно записать в виде:

$$\hat{\beta}_j = \frac{D(X_1, \cdots, X_k; X_1, \cdots, X_{j-1}, Y, X_{j+1}, \cdots, X_k)}{D(X_1, \cdots, X_k; X_1, \cdots, X_k)}.$$

В обсуждаемом частном случае (естественно, обозначения отличаются от только что использованных общих) регрессионная матрица первого шага имеет вид

$$X = (irt),$$

так что, используя сокращение

$$D = D(i, r, t; i, r, t),$$

мы можем написать

$$\begin{aligned}\hat{\pi}_{11} = \hat{\pi}_1 = D^{-1}D(i,r,t;p,r,t),\\ \hat{\pi}_{21} = D^{-1}D(i,r,t;q,r,t),\\ \hat{\pi}_{12} = \hat{\pi}_2 = D^{-1}D(i,r,t;i,p,t),\\ \hat{\pi}_{22} = D^{-1}D(i,r,t;i,q,t),\\ \hat{\pi}_{13} = \hat{\pi}_3 = D^{-1}D(i,r,t;i,r,p),\\ \hat{\pi}_{23} = D^{-1}D(i,r,t;i,r,q).\end{aligned}$$

На втором шаге в роли регрессионной матрицы X выступают матрицы

- $(\hat{p}i)$(для уравнения спроса),
- $(\hat{p}rt)$ (для уравнения предложения).

Поэтому для коэффициента β_4 оценкой двухшагового метода является

$$\hat{\beta}_{4,2SLS} = \frac{D(\hat{p},r,t;q,r,t)}{D(\hat{p},r,t;\hat{p},r,t)}.$$

Подставляя $\hat{p} = \hat{\pi}_1 i + \hat{\pi}_2 r + \hat{\pi}_3 t$ и пользуясь свойствами функции D, легко получаем

$$\begin{aligned}D(\hat{p},r,t;q,r,t) = \hat{\pi}_1 D(i,r,t;q,r,t),\\ D(\hat{p},r,t;\hat{p},r,t) = \hat{\pi}_1^2 D(i,r,t;i,r,t) = \hat{\pi}_1 D(i,r,t;p,r,t).\end{aligned}$$

Отсюда

$$\hat{\beta}_{4,2SLS} = \frac{D(i,r,t;q,r,t)}{D(i,r,t;p,r,t)},$$

что, очевидно, совпадает с оценкой косвенного метода

$$\hat{\beta}_{4,indirect} = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}}.$$

Для коэффициента β_2 оценкой двухшагового метода является

$$\hat{\beta}_{2,2SLS} = \frac{D(\hat{p},i;q,i)}{D(\hat{p},i;\hat{p},i)}.$$

Подставляя выражение для $\hat{p}$ и пользуясь свойствами функции D, можно получить более явное выражение для этой оценки.

8.5 Специальные варианты систем регрессионных уравнений

Мы рассмотрим две практически важные ситуации, когда может оказаться полезным изменить статистическую технику.

Первая ситуация называется *"рекурсивные (recursive) системы уравнений"*. Мы увидим, что эти уравнения можно рассматривать по отдельности. Вторая ситуация называется *"уравнения, кажущиеся несвязанными"(seemingly unrelated equations)*. Как окажется, объединение отдельных подобных уравнений в систему может увеличить эффективность статистических процедур (впрочем, фактически эта система в дальнейшем трактуется как одно уравнение с корреляцией в векторе ошибок).

Перейдем к обсуждению рекурсивных систем. Основная идея очень проста и заключается в том, что правильное упорядочивание уравнений, т.е. правильный порядок принятия их во внимание, может позволить на каждом этапе рассматривать одно единственное уравнение и не обращать внимания на остальные. Поскольку возможность такого упорядочивания определяется визуально, а никакой специальной теории не требуется, мы ограничимся простейшим примером, представляющим собой небольшую модификацию моделей спроса и предложения, рассмотренных в предыдущем параграфе. Подобные примеры обсуждаются во всех учебниках.

Итак, рассмотрим структурную систему из двух уравнений:

$$Q_t = \beta_1 + \beta_2 P_{t-1} + \varepsilon_t^Q,$$

$$P_t = \beta_3 + \beta_4 Q_t + \beta_5 R_t + \varepsilon_t^P$$

и предположим (для структурных систем это предположение вполне естественно), что ошибки ε_t^Q и ε_t^P не коррелируют. Главное отличие (она и создает рекурсивность) — отсутствие в первом уравнении (уравнении предложения) слагаемого P_t, т.е. текущей цены. Тем самым, уравнение предложения можно рассматривать отдельно. Лаговое значение цены P_{t-1} относится к предопределенным величинам и не коррелирует с ошибкой ε_t^Q. На втором этапе мы можем трактовать уже Q_t как предопределенную величину — она не коррелирует с ε_t^P! Затем мы как бы возвращаемся к первому уравнению в следующий момент времени ("как бы", поскольку с точки зрения вычислений возвращаться не нужно — уже все коэффициенты оценены).

Перейдем теперь к уравнениям, кажущимся несвязанными[1]. В этих уравнениях нет эндогенных величин в правых частях, т.е. формально они не сцеплены, и каждое из них можно оценивать отдельно. Однако, если предположить, что ошибки в этих уравнениях коррелируют между собой, то объединение их в систему может дать выигрыш в эффективности. Классические примеры — уравнения спроса на взаимосвязанные (или однотипные) товары, либо же уравнения для инвестиций, осуществляемых компаниями в одной отрасли.

Для понимания тех преимуществ, которые создает объединение уравнений в систему, достаточно рассмотреть случай двух уравнений:

$$Y_{(1)} = X_{(1)}\beta_{(1)} + \varepsilon_{(1)},$$
$$Y_{(2)} = X_{(2)}\beta_{(2)} + \varepsilon_{(2)}.$$

Здесь $X_{(1)}$ и $X_{(2)}$ — регрессионные матрицы, они могут включать как одни и те же, так и различные регрессоры, а $\varepsilon_{(1)}$ и $\varepsilon_{(2)}$ — ошибки, которые предполагаются коррелированными:

$$\mathsf{cov}(\varepsilon_{(1)i}, \varepsilon_{(2)i}) = \sigma_{12},$$
$$\mathbf{V}(\varepsilon_{(1)i}) = \sigma_1^2,$$
$$\mathbf{V}(\varepsilon_{(2)i}) = \sigma_2^2.$$

Остальные ковариации (при $i_1 \neq i_2$) будем считать нулевыми:

$$\mathsf{cov}(\varepsilon_{(1)i_1}, \varepsilon_{(2)i_2}) = 0,$$
$$\mathsf{cov}(\varepsilon_{(1)i_1}, \varepsilon_{(1)i_2}) = 0,$$
$$\mathsf{cov}(\varepsilon_{(2)i_1}, \varepsilon_{(2)i_2}) = 0.$$

Количество наблюдений N одно и то же.

Составим вектор Y, расположив обе серии наблюдений в одну последовательность

$$Y^T = (Y_{(1)1}, \cdots, Y_{(1)N}, Y_{(2)1}, \cdots, Y_{(2)N}),$$

регрессионную матрицу

$$X = \begin{pmatrix} X_{(1)} & 0 \\ 0 & X_{(2)} \end{pmatrix},$$

[1] В [9] используется неудачный термин "внешне не связанные уравнения". Слово "внешне"в русском языке слишком многозначно, а уравнения эти связаны именно "извне", через общую окружающую экономическую среду.

вектор коэффициентов

$$\beta^T = (\beta_{(1)}^T, \beta_{(2)}^T)$$

и вектор ошибок

$$\varepsilon^T = (\varepsilon_{(1)}^T, \varepsilon_{(2)}^T).$$

Матрица ковариаций V вектора ε имеет, очевидно, вид

$$\begin{pmatrix} \sigma_1^2\mathbf{1} & \sigma_{12}\mathbf{1} \\ \sigma_{12}\mathbf{1} & \sigma_2^2\mathbf{1} \end{pmatrix},$$

где $\mathbf{1}$ — единичная матрица порядка N.

Преимущество модели

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

перед исходной системой — в числе степеней свободы: $2N - k_1 - k_2$ вместо $N - k_1$ и $N - k_2$. Впрочем, еще три степени свободы пропадают, поскольку при использовании обобщенного метода наименьших квадратов приходится оценивать дисперсии σ_1^2, σ_2^2 и ковариацию σ_{12} до получения окончательных оценок коэффициентов регрессии.

Таким образом, вся статистическая процедура выглядит следующим образом.

На первом шаге уравнения оцениваются раздельно обычным методом наименьших квадратов, и находятся остатки $\hat{\varepsilon}_{(1)}$, $\hat{\varepsilon}_{(2)}$. С помощью остатков строятся оценки

$$\hat{\sigma}_1^2 = \frac{1}{N - k_1}\hat{\varepsilon}_{(1)}^T\hat{\varepsilon}_{(1)},$$

$$\hat{\sigma}_2^2 = \frac{1}{N - k_2}\hat{\varepsilon}_{(2)}^T\hat{\varepsilon}_{(2)},$$

$$\hat{\sigma}_{12} = \frac{1}{N - k}\hat{\varepsilon}_{(1)}^T\hat{\varepsilon}_{(2)}.$$

Обычно предполагается, что $k_1 = k_2 = k$. При неравных k_1 и k_2 выбор k — отдельная задача, которую мы рассматривать не будем.

На втором шаге оценки $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\sigma}_{12}$ используются в процедуре обобщенного метода наименьших квадратов для нахождения вектора оценок $\hat{\beta}$. Поскольку матрица ковариаций V имеет специальную структуру, вычисление обратной матрицы несколько упрощается. Для описания этого упрощения удобно использовать понятие произведения Кронекера двух матриц (см. [9]):

$$V = \Sigma \otimes \mathbf{1},$$

где

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

а $\mathbf{1}$, как и раньше, единичная матрица порядка N. Из свойств произведения Кронекера получаем

$$V^{-1} = \Sigma^{-1} \otimes \mathbf{1},$$

$$X^T V^{-1} X = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 X_{(1)}^T X_{(1)} & -\sigma_{12} X_{(1)}^T X_{(2)} \\ -\sigma_{12} X_{(2)}^T X_{(1)} & \sigma_1^2 X_{(2)}^T X_{(2)} \end{pmatrix}$$

и

$$X^T V^{-1} Y = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 X_{(1)}^T Y_{(1)} & -\sigma_{12} X_{(1)}^T Y_{(2)} \\ -\sigma_{12} X_{(2)}^T Y_{(1)} & \sigma_1^2 X_{(2)}^T Y_{(2)} \end{pmatrix}.$$

Интересно отметить нетривиальный частный случай, когда наши оценки сводятся к обычным оценкам наименьших квадратов (собственно, ради этого и выписывались приведенные выше формулы). Предположим, что регрессионные матрицы $X_{(1)}$ и $X_{(2)}$ совпадают. Тогда, как легко заметить,

$$X^T V^{-1} X = \Sigma^{-1} \otimes X_{(1)}^T X_{(1)},$$

$$(X^T V^{-1} X)^{-1} = \Sigma \otimes (X_{(1)}^T X_{(1)})^{-1},$$

и для наших оценок получаем формулу

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y =$$

$$\frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_1^2 (X_{(1)}^T X_{(1)})^{-1} & \sigma_{12} (X_{(1)}^T X_{(1)})^{-1} \\ \sigma_{12} (X_{(1)}^T X_{(1)})^{-1} & \sigma_2^2 (X_{(1)}^T X_{(1)})^{-1} \end{pmatrix} \begin{pmatrix} X_{(1)}^T [\sigma_2^2 Y_{(1)} - \sigma_{12} Y_{(2)}] \\ X_{(1)}^T [\sigma_1^2 Y_{(2)} - \sigma_{12} Y_{(1)}] \end{pmatrix} =$$

$$= \begin{pmatrix} (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y_{(1)} \\ (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y_{(2)} \end{pmatrix}.$$

Конечно, есть и другой, очевидный, случай совпадения указанных оценок: $\sigma_{12} = 0$.

В заключение параграфа отметим, что та же идея учета корреляций между ошибками различных уравнений системы может быть использована и в том случае, когда уравнения не кажутся несвязанными. Сначала каждое уравнение системы оценивается двухшаговым методом наименьших квадратов, находятся остатки и, с их помощью, оценка матрицы ковариаций между ошибками, а затем на третьем шаге с

помощью обобщенного метода наименьших квадратов, так же как и выше, заново оцениваются коэффициенты всех уравнений сразу. Эта процедура, трехшаговый метод наименьших квадратов, дает выигрыш в эффективности по сравнению с двухшаговым методом, но в вычислительном плане чрезвычайно трудоемка — требуется обращение матриц значительно более высокого порядка. Кроме того, она, разумеется, не применима, если хотя бы одно из уравнений системы содержит неидентифицируемые коэффициенты.

8.6 Тестирование системы

Рассмотрим, наконец, вопрос об обнаружении корреляции между регрессорами и ошибками. Напомним сначала некоторые обстоятельства, связанные с этой проблемой. Мы выделяем одно уравнение из структурной системы и собираемся оценивать его коэффициенты. Какой метод оценивания выбрать.

Простейший подход — обычный метод наименьших квадратов. Его применимость определяется возможностью трактовать величины в правой части уравнения экзогенным образом (в этом случае корреляции между ними и ошибками не должно быть). Однако в контексте систем регрессионных уравнений некоторые из этих величин будут внутренними (по отношению ко всей системе). Такие величины, скорее всего, будут коррелировать с ошибками (влияние остальных уравнений), а тогда выбранный подход (OLS) несостоятелен.

Другой подход — использование инструментов (т.е. 2-SLS) — пригоден в обоих случаях, но при отсутствии корреляций менее эффективен, чем обычный метод наименьших квадратов. Коротко, OLS может привести к полностью ошибочным выводам, а 2-SLS — к потере в эффективности.

Тем самым, выбор метода оценивания, или, что, по существу, то же самое, модели (отдельное уравнение или система) оказывается одним из ключевых этапов исследования. Задача тестирования (т.е. выбора) модели оказывается при этом задачей тестирования некоторых величин на экзогенность.

Простой и наглядный тест на экзогенность (Hausman-Wu exogeneity test, см. [25]) состоит в следующем.

Прежде всего выделяется группа величин, экзогенность которых следует тестировать (подразумевается, что экзогенность остальных сомнения не вызывает). Для каждой из них строится целевой инструмент

обычным образом. Затем сравниваются две регрессии — короткая, в которую включены все первоначальные величины рассматриваемого уравнения, и длинная, в которую дополнительно включены построенные целевые инструменты.

Экзогенность (отсутствие корреляций с ошибками) означает, что дополнительно оцениваемые коэффициенты при инструментах на самом деле нулевые. Проверка подобных гипотез обсуждалась в параграфе 6.8, так что мы можем не выписывать соответствующую F-статистику.

ПРИЛОЖЕНИЯ

A. Гамма-функция и гамма-распределение

В этом приложении мы кратко напоминаем несколько полезных фактов, относящихся к гамма-функции, а также их применения к выводу свойств гамма-распределения.

Для положительных значений аргумента (другие нам не потребуются) гамма-функция определяется равенством

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x}\, dx \qquad (p > 0).$$

Ключевое свойство, постоянно использующееся в вычислениях с гамма-функцией, имеет вид

$$\Gamma(p+1) = p\Gamma(p) \tag{A.1}$$

(она легко доказывается интегрированием по частям). Поскольку $\Gamma(1) = 1$, из формулы (A.1) сразу получаем $\Gamma(n+1) = n!$ $(n = 0, 1, 2, \dots)$. Еще одно полезное частное значение $\Gamma(1/2) = \sqrt{\pi}$ (см. также параграф 1.6) обсуждается ниже.

Рассмотрим теперь некоторые детали, связанные с определением плотности гамма-распределения:

$$g(x) = \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, \qquad x > 0.$$

Начнем с проверки условия нормировки:

$$\int_0^\infty g(x)\, dx = \ (y = \alpha x)\ = \frac{\alpha^p}{\Gamma(p)} \int_0^\infty \left(\frac{y}{\alpha}\right)^{p-1} e^{-y} \frac{dy}{\alpha} = \\ = \frac{1}{\Gamma(p)} \int_0^\infty y^{p-1} e^{-y}\, dy = 1.$$

Для плотности гамма-распределения $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$ мы имеем два выражения — общее

$$\frac{\left(\frac{1}{2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)} x^{-1/2} e^{-x/2}$$

и частное, выведенное в параграфе 1.6:

$$\frac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2}.$$

Сравнение их дает упомянутое выше значение $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$. С помощью формулы (A.1) легко получаем также

$$\Gamma\left(k+\frac{1}{2}\right)=\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\cdots\frac{1}{2}\Gamma\left(\frac{1}{2}\right)=\frac{(2k)!}{2^{2k}k!}\sqrt{\pi}.$$

Вычислим теперь непосредственно свертку гамма-плотностей $\Gamma(\alpha,p_1)$ и $\Gamma(\alpha,p_2)$:

$$(g_1*g_2)(x)=$$
$$\int_0^x\frac{\alpha^{p_1}}{\Gamma(p_1)}y^{p_1-1}e^{-\alpha y}\frac{\alpha^{p_2}}{\Gamma(p_2)}(x-y)^{p_2-1}e^{-\alpha(x-y)}dy=$$
$$\frac{\alpha p_1+p_2}{\Gamma(p_1)\Gamma(p_2)}e^{-\alpha x}\int_0^x y^{p_1-1}(x-y)^{p_2-1}dy=\quad(y=xz)\quad=$$
$$\frac{\alpha^{p_1+p_2}}{\Gamma(p_1)\Gamma(p_2)}e^{-\alpha x}\int_0^1(xz)^{p_1-1}(x-xz)^{p_2-1}xdz=$$
$$\frac{\alpha^{p_1+p_2}}{\Gamma(p_1)\Gamma(p_2)}x^{p_1+p_2-1}e^{-\alpha x}\int_0^1 zp_1-1(1-z)^{p_2-1}dz.$$

Сравнивая полученное выражение с основной формулой для плотности $\Gamma(\alpha,p_1+p_2)$, получаем

$$\frac{1}{\Gamma(p_1)\Gamma(p_2)}\int_0^1 z^{p_1-1}(1-z)^{p_2-1}dz=\frac{1}{\Gamma(p_1+p_2)}.$$

Из этого равенства вытекает, что

$$\int_0^1\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}z^{p_1-1}(1-z)^{p_2-1}dz=1$$

— условие нормировки для плотности бета-распределения.

Перейдем теперь к вычислению моментов гамма-распределения.

$$\mathsf{E}(X^k)=\int_0^\infty x^k g(x)dx=\int_0^\infty\frac{\alpha^p}{\Gamma(p)}x^{k+p-1}e^{-\alpha x}dx=$$
$$=\frac{\alpha^p}{\Gamma(p)}\frac{\Gamma(p+k)}{\Gamma(p)\alpha^k}=\frac{\Gamma(p+k)}{\Gamma(p)\alpha^k}.$$

При натуральном k с помощью формулы(A.1) легко получаем

$$\mathsf{E}(X^k) = \frac{p(p+1)\cdots(p+k-1)}{\alpha^k}.$$

В частности,

$$\mathsf{E}X = \frac{p}{\alpha}, \quad \mathsf{E}(X^2) = \frac{p(p+1)}{\alpha^2}.$$

Из последней формулы следует, что

$$\mathsf{V}X = \frac{p(p+1)}{\alpha^2} - \frac{p^2}{\alpha^2} = \frac{p}{\alpha^2}.$$

Аналогично проверяется, что

$$\mathsf{E}[(x - \mathsf{E}X)^3] = \frac{2p}{\alpha^3}.$$

В частном случае показательного распределения ($p = 1$)

$$\mathsf{E}(X^k) = \frac{k!}{\alpha^k}, \quad \mathsf{E}X = \frac{1}{\alpha}, \quad \mathsf{V}X = \frac{1}{\alpha^2}.$$

Моменты бета-распределения $B(p_1, p_2)$ вычисляются аналогичным образом:

$$\mathsf{E}(X^k) =$$

$$= \int_0^1 x^k p(x)dx = \frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} \int_0^1 x^{k+p_1-1}(1-x)^{p_2-1}dx =$$

$$= \frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \frac{\Gamma(p_1+k)\Gamma(p_2)}{\Gamma(p_1+p_2+k)} = \frac{\Gamma(p_1+k)\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_1+p_2+k)}.$$

При натуральном k формула (A.1) позволяет вывести отсюда

$$\mathsf{E}(X^k) =$$

$$= \frac{p_1(p_1+1)\cdots(p_1+k-1)}{(p_1+p_2)(p_1+p_2+1)\cdots(p_1+p_2+k-1}.$$

В частности,

$$\mathsf{E}X = \frac{p_1}{p_1+p_2}, \quad \mathsf{E}(X^2) = \frac{p_1(p_1+1)}{(p_1+p_2)(p_1+p_2+1)},$$

$$\mathsf{V}X = \frac{p_1 p_2}{(p_1+p_2)^2(p_1+p_2+1)}.$$

При $p_1 = p_2 = 1$ получаем моменты равномерного распределения на $\langle 0, 1\rangle$:

$$\mathsf{E}X = \frac{1}{2}, \quad \mathsf{E}(X^k) = \frac{1}{k+1}, \quad \mathsf{V}X = \frac{1}{12}$$

(впрочем, их проще вычислить непосредственно).

B. Многомерное нормальное распределение

Начнем с определения. Случайный вектор X размерности r имеет нормальное распределение, если для любого $z \in \mathbb{R}^r$ одномерная случайная величина $z'X$ нормально распределена. Мы при этом придерживаемся соглашения, упоминавшегося в параграфе 1.5, о том, что вырожденное распределение считается нормальным.

Из приведенного определения следует, что любая компонента X_j (и любой подвектор) нормально распределенного вектора X также имеет нормальное распределение (в качестве z' берем координатные векторы $(1,0,\cdots,0)$,...,$(0,\cdots,0,1)$). Обратное утверждение неверно: если каждая компонента случайного вектора X нормально распределена, сам вектор не обязан иметь нормальное распределение[2].

В учебниках теории вероятностей доказывается, что многомерное нормальное распределение однозначно определяется вектором средних значений $a = \mathsf{E}X$ и матрицей ковариаций $C = \mathsf{E}[(X - \mathsf{E}X)(X - \mathsf{E}X)']$. Более точно, если $a \in \mathbb{R}^r$ — произвольный вектор, а C — произвольная симметричная неотрицательно определенная матрица r-го порядка, то существует (единственное) нормальное распределение в $\mathbb{R}^r$, имеющее этот вектор и эту матрицу в качестве вектора средних и матрицы ковариаций.

Если C — строго положительно определенная матрица (в этом случае она невырождена, и существует обратная матрица C^{-1}), то нормальное распределение задается плотностью

$$p(x) = (2\pi)^{-r/2} \frac{1}{\sqrt{\det C}} \exp\{-\frac{1}{2}(x-a)'C^{-1}(x-a)\}.$$

Если C вырождена, то соответствующее нормальное распределение сосредоточено на некотором линейном многообразии меньшей размерности. Вводя в нем систему координат, можно в этих координатах

[2]К сожалению, подобную ошибку можно встретить и в популярных учебниках

записать плотность нормального распределения аналогичной формулой. Пример подобной ситуации можно получить, если нормально распределенный вектор X, имеющий плотность, вложить в пространство большей размерности. В этом объемлющем пространстве у него уже не будет плотности.

Важный частный случай многомерного нормального распределения возникает при рассмотрении независимых нормально распределенных величин. Если $X_1, \cdots, X_r$ независимы, причем $X_i \in \mathbf{N}(a_i, \sigma_i^2)$ $(i = 1, \cdots, r)$, то вектор X, составленный из величин $X_1, \cdots, X_r$, всегда имеет многомерное нормальное распределение с параметрами $a = (a_1, \cdots, a_r)'$, $C = \mathsf{diag}(\sigma_1^2, \cdots, \sigma_r^2)$.

В отдельных случаях оказывается полезной формула для двумерной нормальной плотности (частный случай общей формулы):

$$p(x_1, x_2) = \\ = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\ \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-a_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-a_1)(x_2-a_2)}{\sigma_1\sigma_2} + \frac{(x_1-a_1)^2}{\sigma_2^2}\right]\right\}$$

(в этой формуле использован вместо ковариации коэффициент корреляции ρ между компонентами двумерного нормально распределенного вектора).

C. Закон больших чисел для зависимых случайных величин

Для простоты формулировки мы ограничимся случаем центрированных величин: $\mathsf{E}X_n = 0$, однако не будем предполагать их одинаковой распределенности. Условия теоремы, приводимой ниже, используют корреляционные характеристики, поэтому мы предположим существование дисперсий $\sigma_n^2 = \mathsf{E}(X_n^2) \neq 0$. Обозначим через ρ_{mn} коэффициент корреляции между X_m и X_n.

ТЕОРЕМА. *Пусть $\{X_n\}$ — последовательность центрированных случайных величин с конечными ненулевыми дисперсиями. Предположим, что последовательность их дисперсий ограничена: $\sigma_n^2 \leq c < \infty$, а коэффициенты корреляции удовлетворяют условию*

$$\rho_{mn} \to 0 \quad \textit{при} \quad |m-n| \to \infty.$$

Тогда для последовательности $\{X_n\}$ справедлив закон больших чисел:

$$\frac{X_1 + \cdots + X_N}{N} \to 0 \quad \textit{при} \quad N \to \infty$$

по вероятности.

Теорема относительно несложно доказывается при помощи неравенства Чебышёва: достаточно проверить, что

$$\frac{1}{N^2}\mathsf{var}(X_1 + \cdots + X_N) \to_{N\to\infty} 0. \qquad (C.1)$$

Зафиксируем малое положительное число ε и найдем по ε натуральное K, такое, что

$$|\rho_{mn}| \leq \varepsilon \quad \text{при} \quad |m-n| > K.$$

Тогда при $N > K$ можно написать оценку

$$\frac{1}{N^2}\mathsf{var}(X_1 + \cdots + X_N) = \frac{1}{N^2}\left(\sum_{m,n=1}^{N} \sigma_m \sigma_n \rho_{mn}\right) \leq$$

$$\leq \frac{c}{N^2}\left(\sum_{1\leq m,n\leq N:|m-n|\leq K} |\rho_{mn}| + \sum_{1\leq m,n\leq N:|m-n|>K} |\rho_{mn}\right) \leq$$

$$\frac{c}{N^2}\left(N^2 - \alpha_N(K) + \varepsilon\alpha_N(K)\right).$$

Здесь $\alpha_N(K)$ — число слагаемых во второй сумме; легко сосчитать, что $\alpha_N(K) = (N-K-1)(N-K)$. Подставляя это значение в нашу оценку и переходя к верхнему пределу при $N \to \infty$, получаем

$$\limsup_{N\to\infty} \frac{1}{N^2}\mathsf{var}(X_1 + \cdots + X_N) \leq c\varepsilon.$$

Ввиду призвольности ε это неравенство доказывает (C.1) и теорему.

D. Условные математические ожидания

Напомним сначала простейшее определение условного математического ожидания.

Пусть H — событие ненулевой вероятности, X — случайная величина с конечным математическим ожиданием. Тогда число

$$\mathsf{E}(X|H) = \frac{1}{\mathsf{P}(H)}\mathsf{E}(X1_H) \qquad (D.1)$$

называется условным ожиданием X при условии H. Наглядный смысл выражения в правой части состоит в том, что оно является усреднением величины X по множеству H. В частном случае, когда $X = 1_A$, мы получаем обычное элементарное определение условной вероятности:

$$\mathsf{P}(A|H) = \mathsf{E}(1_A|H) = \frac{\mathsf{E}(1_A 1_H)}{\mathsf{P}(H)} = \\ = \frac{\mathsf{E}(1_{AH})}{\mathsf{P}(H)} = \frac{\mathsf{P}(AH)}{\mathsf{P}(H)}.$$

Обобщением определения (D.1) является определение условного ожидания относительно разбиения.

Пусть $\mathcal{H} = \{H_1, H_2, \cdots\}$ — полная группа событий, т.е. разбиение пространства элементарных исходов Ω на непересекающиеся части:

$$H_1 \cup H_2 \cup \cdots = \Omega, \quad H_i \cap H_j = \emptyset \quad (i \neq j).$$

Предположим еще, что все вероятности $\mathsf{P}(H_i)$ ненулевые, и составим из условных математических ожиданий $\mathsf{E}(X|H_i)$ функцию

$$\hat{X}(\omega) = \mathsf{E}(X|H_i), \quad \omega \in H_i. \qquad (D.2)$$

Эта функция $\hat{X}$ называется условным математическим ожиданием величины X относительно разбиения $\mathcal{H}$ и обозначается $\mathsf{E}(X|\mathcal{H})$.

Подчеркнем, что здесь условное ожидание перестает быть числом и становится случайной величиной. Ее наглядный смысл — "локальное"усреднение величины X , т.е. усреднение по отдельным множествам H_i. Случайная величина $\hat{X}$ постоянна на событиях H_i и в этом смысле измерима относительно σ-алгебры $\sigma(\mathcal{H})$, порожденной разбиением $\mathcal{H}$: для каждого промежутка $\langle a, b\rangle$ его прообраз $\hat{X}^{-1}(\langle a, b\rangle)$ является объединением каких-то из множеств H_i, т.е. элементом σ-алгебры $\sigma(\mathcal{H})$.

Перечислим некоторые основные свойства условного математического ожидания $\mathsf{E}(X|\mathcal{H})$, легко вытекающие из определения:

1. (линейность)
$$\mathsf{E}(\alpha_1 X_1 + \alpha_2 X_2|\mathcal{H}) = \alpha_1 \mathsf{E}(X_1|\mathcal{H}) + \alpha_2 \mathsf{E}(X_2|\mathcal{H}).$$
2. (формула полного математического ожидания)
$$\mathsf{E}(\mathsf{E}(X|\mathcal{H})) = \mathsf{E}X.$$
3. Множитель Z, измеримый относительно σ-алгебры $\sigma(\mathcal{H})$ (т.е. локально постоянный = постоянный на множествах H_i), можно выносить за знак условного математического ожидания
$$\mathsf{E}(ZX|\mathcal{H}) = Z\mathsf{E}(X|\mathcal{H}).$$
4. Если Z измерима относительно $\sigma(\mathcal{H})$, то
$$\mathsf{E}(ZX) = \mathsf{E}(Z\mathsf{E}(X|\mathcal{H})) \qquad (D.3)$$

Для получения последней формулы надо приравнять математические ожидания обеих частей формулы свойства 3 и упростить левую часть по формуле полного ожидания.

Перейдем, наконец, к самому общему определению условного математического ожидания. Пусть $\mathcal{S}$ — какая-нибудь σ-алгебра, состоящая из событий (не обязательно всех), X — случайная величина с конечным математическим ожиданием. Определим новую случайную величину $\hat{X} = \mathsf{E}(X|\mathcal{S})$ — условное математическое ожидание X относительно $\mathcal{S}$, перечислив свойства, которыми она должна обладать. Таких свойств всего два:

I) $\hat{X}$ измерима относительно $\mathcal{S}$, т.е. прообразы промежутков лежат в $\mathcal{S}$: для любого $\langle a, b\rangle$
$$\hat{X}^{-1}(\langle a, b\rangle) \in \mathcal{S}.$$

II) Если Y измерима относительно $\mathcal{S}$, то

$$\mathsf{E}(YX) = \mathsf{E}(Y\hat{X})$$

(точнее, если левая часть определена, то и правая определена, и они равны между собой почти всюду).

Как указано выше, условное ожидание относительно разбиения этими свойствами обладает. В общем случае, т.е. когда σ-алгебра $\mathcal{S}$ не порождается разбиением, справедлива теорема существования и почти единственности, которую мы примем без доказательства:

Величина $\hat{X}$, обладающая свойствами I) и II), существует. Любые две такие величины совпадают с вероятностью 1 (они называются вариантами условного математического ожидания).

Из приведенного определения и теоремы существования вытекают свойства 1 — 4, в которых равенство случайных величин следует понимать как равенство с вероятностью 1.

Свойство 1 приобретает такой вид: $\alpha_1\hat{X}_1 + \alpha_2\hat{X}_2$ — один из вариантов условного математического ожидания для $\alpha_1X_1 + \alpha_2X_2$.

В самом деле, формула II) приобретает вид

$$\mathsf{E}[Y(\alpha_1X_1 + \alpha_2X_2)] = \mathsf{E}[Y(\alpha_1\hat{X}_1 + \alpha_2\hat{X}_2)]$$

и вытекает из аналогичных соотношений для $\hat{X}_1$ и $\hat{X}_2$.

Измеримость линейной комбинации $\alpha_1\hat{X}_1 + \alpha_2\hat{X}_2$ (свойство I)) мы проверять не будем (эвристически она почти очевидна, а формальное рассуждение несколько тяжеловесно).

Свойство 2 — частный случай формулы II) при $Y \equiv 1$.

Наконец, свойство 3 можно переформулировать так: $Y\hat{X}$ — один из вариантов условного математического ожидания для YX. Для проверки формулы II) выберем случайную величину Z, измеримую относительно $\mathcal{S}$. Нужно доказать, что

$$\mathsf{E}(ZY\hat{X}) = \mathsf{E}(ZYX).$$

Это вытекает из того, что произведение ZY измеримых относительно $\mathcal{S}$ величин также измеримо (мы не проверяем это свойство). Измеримость $Y\hat{X}$ (свойство I)) следует из тех же соображений.

Приведем еще одно полезное свойство условных математических ожиданий:

5. Если $\mathcal{S}_1 \subset \mathcal{S}_2$, то

$$\mathsf{E}(\mathsf{E}(X|\mathcal{S}_1)|\mathcal{S}_2) = \mathsf{E}(\mathsf{E}(X|\mathcal{S}_2)|\mathcal{S}_1) = \mathsf{E}(X|\mathcal{S}_1)$$

(это доказывается похожими рассуждениями).

Наиболее важным для большинства приложений является случай, когда σ-алгебра $\mathcal{S} = \sigma(U)$ порождается некоторой случайной величиной U (или несколькими величинами), т.е. порождается событиями вида $U^{-1}(\langle a, b\rangle)$. Принято писать такое условное математическое ожидание в виде

$$\mathsf{E}(X|U).$$

Всякая функция, измеримая относительно указанной σ-алгебры, представляется в виде $g(U)$. В частности, это верно для условного математического ожидания. При этом функция g почти единственна. Для ее значений $g(u)$ иногда пишут выражение

$$g(u) = \mathsf{E}(X|U = u),$$

которое с некоторыми оговорками можно трактовать, как условное ожидание относительно события $(U = u)$, даже если последнее имеет нулевую вероятность.

Условные вероятности относительно $\mathcal{S}$ определяются как частный случай: по определению

$$\mathsf{P}(A|\mathcal{S}) = \mathsf{E}(1_A|\mathcal{S}).$$

В таком общем контексте условные вероятности являются случайными величинами и определены лишь почти единственным образом. В ряде случаев (см.[12]) удается выбрать варианты этих условных вероятностей для разных A так, чтобы на некотором едином множестве полной вероятности выполнялось главное свойство обычных вероятностей — аддитивность (даже счетная аддитивность). В частности, это удается сделать для событий, связанных с некоторой случайной величиной X. На этом пути получается условное распределение X:

$$\mathsf{P}(X \in \langle a, b\rangle|U = u).$$

Если совместное распределение величин X и U задается плотностью $p(x, u)$, то условное распределение задается условной плотностью

$$p(x|u) = \frac{p(x, u)}{p_U(u)} = \frac{p(x, u)}{\int_{\mathbb{R}} p(x, u)dx}.$$

При этом условное математическое ожидание получается интегрированием по условной плотности:

$$\mathsf{E}(X|U = u) = \int_{\mathbb{R}} xp(x|u)dx.$$

Полное изложение теории условных математических ожиданий и условных распределений можно найти в [12].

Подчеркнем, что наше обсуждение этих вопросов весьма схематично и далеко не полно. Мы лишь обозначаем некоторые ключевые определения и формулы.

Литература

[1] Боровков А.А. Математическая статистика. М.: Наука, 1984 (имеется также более позднее, переработанное, издание: Новосибирск, Наука, 1997).

[2] Ван-дер-Варден Б. Математическая статистика. М.: ИЛ, 1960.

[3] Воинов В.Г., Никулин М.С. Несмещенные оценки и их приложения. М.: Наука, 1989.

[4] Джонстон Дж. Эконометрические методы. М.: Статистика, 1980.

[5] Елисеева И.И., Юзбашев М.М. Общая теория статистики. М.: Финансы и статистика, 1996.

[6] Кендалл М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973.

[7] Кокрен У. Методы выборочного исследования. М.: Статистика, 1976.

[8] Крамер Г. Математические методы статистики. М.: Мир, 1975 (первое издание на русском языке: М.: ИЛ, 1948).

[9] Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. 5-е изд., М.: Дело, 2001.

[10] Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.

[11] Тутубалин В.Н. Теория вероятностей. М.: МГУ, 1972.

[12] Ширяев А.Н. Вероятность. 2-е изд., М.: Наука, 1989.

[13] Bernardo J.M., Smith A.F.M. Bayesian Theory. Wiley, 1993.

[14] Breusch T.S., Pagan A.R. The Lagrange Multiplier test and its applications to model specification tests in econometrics. Review of Economic Studies (1980), v.47, p.239 – 253.

[15] Charemza W.W., Deadman D.F. New Directions in Econometric Practice. General to Specific Modelling, Cointegration and Vector Autoregression. Second edition. Edward Elgar Publishing, Inc., 1997.

[16] Durbin J., Watson G.S. Testing for serial correlation in least squares regression.I. Biometrika (1950), v.37, p.409 – 428.

[17] Durbin J., Watson G.S. Testing for serial correlation in least squares regression.II. Biometrika (1951), v.38, p.159 – 178.

[18] Durbin J., Watson G.S. Testing for serial correlation in least squares regression.III. Biometrika (1971), v.58, p.1 – 19.

[19] Greene W.H. Econometric Analysis. Fourth edition. Prentice Hall International, Inc., 2000.

[20] Hamilton J. Time Series Analysis. Princeton: Princeton University Press, 1994.

[21] Hendry D.F. Dynamic Econometrics. Oxford University Press, 1995.

[22] Intriligator M.D. Econometric Models, Technics, and Applications. Prentice-Hall, Inc., 1978.

[23] Johnston J., DiNardo J. Econometric Methods. Fourth edition. McGraw-Hill, 1997.

[24] Pindyck R.S., Rubinfeld D.L. Econometric Models and Economic Forecasts. Third edition. McGraw-Hill, 1991.

[25] Stewart J., Gill L. Econometrics. Second edition. Prentice Hall Europe, 1998.

Сергей Сергеевич Валландер

ЛЕКЦИИ ПО СТАТИСТИКЕ И ЭКОНОМЕТРИКЕ

Утверждено к печати
Ученым советом Европейского университета
в Санкт-Петербурге

Компьютерная верстка автора

Издательство Европейского университета в Санкт-Петербурге
198187, Санкт-Петербург, ул. Гагаринская, 3
e-mail: books@eu.spb.ru

Лицензия ИД № 03435 от 05.12.2000. Сдано в набор
Подписано к печати 20.12.05 Формат 60 x 90 1/16. Гарнитура Таймс.
Бумага офсетная. Усл.-печ. л.16 п.л.
Тираж 300 экз.

Отпечатано с оригинал-макета
на ризографе Европейского университета в Санкт-Петербурге
191187, Санкт-Петербург,
Гагаринская ул., д. 3