ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Т.А. Ратникова, К.К. Фурманов

Анализ панельных данных и данных о длительности состояний

Учебное пособие

Издательский дом Высшей школы экономики
Москва 2014

УДК 303.7.023(075)
ББК 65в6
Р25

Рецензент:
доктор физико-математических наук, профессор,
заведующий кафедрой эконометрики и математических методов
экономики МШЭ МГУ им. М.В. Ломоносова *С.А. Айвазян*

Р25 **Ратникова, Т. А., Фурманов К. К.** Анализ панельных данных и данных о длительности состояний [Текст] : учеб. пособие / Т. А. Ратникова, К. К. Фурманов ; Нац. исслед. ун-т «Высшая школа экономики». — М. : Изд. дом Высшей школы экономики, 2014. — 373, [3] с. — 1000 экз. — ISBN 978-5-7598-1093-3 (в обл.).

Учебное пособие охватывает темы эконометрики продвинутого уровня. В нем изложены теория и практика применения актуальных методов вероятностно-статистического анализа экономических и социологических данных, используемых для оценивания зависимостей по пролонгированным выборкам объектов, в роли которых могут выступать индивиды, семьи, фирмы, регионы, страны и т.п. Наличие последовательного ряда наблюдений позволяет учитывать индивидуальные особенности различных единиц наблюдения и их эволюции, а также изучать продолжительность пребывания объектов в том или ином состоянии (например, длительность периодов бедности для домохозяйств или периодов безработицы для индивидов).

Излагаются базовые теоретические концепции анализа панельных данных и данных о длительности состояний, а также принципы построения наиболее востребованных моделей. Примеры использования рассмотренных методов на практике строятся по реальным российским панельным данным РМЭЗ (Российского мониторинга экономического состояния и здоровья населения).

Применение изученных методов к реальным российским статистическим данным позволит глубже понять цели и задачи экономической политики государства (или фирмы), а также научиться оценивать результаты этой политики.

Пособие полезно магистрантам, аспирантам и исследователям, специализирующимся в областях математических методов анализа экономики, микро- и макроэкономического анализа, экономики фирм, анализа потребительского поведения населения, рынка труда, экономики здравоохранения, демографии.

УДК 303.7.023(075)
ББК 65в6

ISBN 978-5-7598-1093-3

© Ратникова Т.А., Фурманов К.К., 2014
© Оформление. Издательский дом Высшей школы экономики, 2014

Содержание

Часть II. Моделирование длительности состояний

От авторов

Это учебное пособие — переработанный и значительно дополненный вариант книги Т.А. Ратниковой «Введение в эконометрический анализ панельных данных», изданной в 2010 г.

В то время лишь в нескольких учебниках по эконометрике, изданных на русском языке, существовали разделы, более или менее подробные, посвященные анализу панельных данных («Эконометрика» И.И. Елисеевой, «Эконометрика. Начальный курс» Я. Магнуса, П.К. Катышева, А.А. Пересецкого, перевод учебника М. Вербика «Путеводитель по современной эконометрике» под редакцией С.А. Айвазяна, «Эконометрика» В.П. Носко). Обстоятельные иностранные монографии [Hsiao, 1986; Baltagi, 1995; Matyas, Sevestre, 1996] и сейчас еще не переведены на русский язык и доступны студентам далеко не повсеместно. Кроме того, методический арсенал эконометриста, занимающегося анализом панелей, значительно расширился за последние десятилетия и продолжает расти; новые методы нашли свое место как в академических журналах, так и в программном обеспечении — их описание заслуживает отдельной книги, и настоящее пособие — попытка заполнить этот пробел.

Почему исследователи обращаются к панельным данным? Например, потому что такие данные позволяют учесть не измеримые, не наблюдаемые статистикой различия между обследуемыми объектами (регионами, фирмами, индивидами). В настоящее время данные такого рода встречаются нередко, так что говорить об анализе панельных данных как об узкой или маловостребованной отрасли науки не приходится. Обращаясь к пространственной выборке, аналитик изучает различия между наблюдаемыми объектами, а исследуя временные ряды, — изменение состояния отдельного объекта с течением времени. Использование панельных данных позволяет подступить к решению обеих задач и построить модель, объясняющую динамику состояний множества объектов.

Первая часть настоящего пособия выросла из курса лекций, читаемого студентам-старшекурсникам бакалавриата и магистрантам факультета экономики НИУ ВШЭ с 2001 г. В 2004 г. материалы лекций, работа над которыми была поддержана грантом НФПК (Национального фонда подготовки кадров) в рамках программы

«Совершенствование преподавания социально-экономических дисциплин в вузах» инновационного проекта развития образования, появились в электронном виде на сайте университета. В 2006 г. они были существенно дополнены и опубликованы в разделе «Лекционные и методические материалы» в «Экономическом журнале ВШЭ». Еще через четыре года они переросли в пособие [Ратникова, 2010], о котором уже говорилось.

Вторая, полностью новая, часть настоящей книги — «Модели длительности состояний» посвящена теме, которая практически не рассматривается в учебной литературе по эконометрике. В известных отечественных учебниках [Айвазян, Мхитарян, 1998; Магнус, Катышев, Пересецкий, 2004] этой теме уделено всего несколько страниц, иностранные пособия [Greene, 2012; Cameron, Trivedi, 2005] содержат немногословные параграфы. И те и другие книги могут создать впечатление, будто модели длительности — это причудливые вариации на тему классической регрессионной модели, обвешанные мишурой из непривычных терминов. В действительности речь идет о существенно обособленном подходе к анализу, а необычные термины и понятия — способы адекватно передать суть исследуемых процессов, плохо укладывающуюся в рамки привычных для эконометристов средств описания данных. Поэтому, в отличие от первой части, где материал излагается в предположении, что читатель твердо усвоил базовый курс эконометрики, вторая часть содержит изложение почти с нуля. В значительной мере она доступна для студента, изучившего курс теории вероятностей и математической статистики и знакомого с регрессионным анализом.

При составлении второй части мы столкнулись с проблемой выбора терминов — одно и то же понятие исследователи трактуют по-разному. Для важнейшего понятия анализа длительностей нами использован термин «функция риска» вместо более распространенного — «функция опасности отказов». Последний вариант, применяемый специалистами по теории надежности, несет на себе слишком явный отпечаток конкретной области применения. По той же причине был отклонен и вариант «сила смертности», распространенный среди демографов и актуариев. В математической статистике есть и другая функция риска, используемая при анализе критериев для про-

верки гипотез, но вряд ли совпадение терминов приведет к путанице — слишком уж различна суть этих функций.

Мы признательны всем, кто оказывал содействие в подготовке и совершенствовании этой книги. Учебник не смог бы состояться без его идейного вдохновителя — Э.Б. Ершова, без Г.Г. Канторовича, который потратил немало своих времени и сил, помогая совершенствовать текст на ранних стадиях разработки; без участия французских коллег — Ф. Гарда, Б. Дормонт и М. Морель, которые охотно делились своим опытом; без ценных советов и рекомендаций С.А. Айвазяна, В.А. Бессонова, П.К. Катышева, Е.В. Коссовой, А.А. Пересецкого, И.Г. Поспелова; без весьма полезного опыта работы с реальными данными в Центре трудовых исследований под руководством В.Е. Гимпельсона и Р.И. Капелюшникова; без В.С. Автономова, который принял решение о выделении средств факультета на издание пособия в 2010 г.; без кропотливого и самоотверженного труда сотрудников «Экономического журнала» и Издательского дома НИУ ВШЭ, которым пришлось немало повозиться с многочисленными формулами. Отдельную благодарность мы выражаем своим ученикам, в особенности Анне Гладышевой, Ирине Чернышевой и Татьяне Барладян, чьи исследовательские работы легли в основу приведенных в этой книге примеров.

Вся ответственность за недостатки, содержащиеся в пособии, целиком лежит на авторах. Мы будем признательны, если читатель, обнаруживший ту или иную ошибку, будет не только внимателен, но и великодушен и сообщит нам о своей находке, дав возможность ее исправить.

Часть I

Методы анализа панельных данных

1. Введение

1.1. История создания микроэконометрики

Сравнительно недавно эмпирические исследования в эконометрике были обогащены возможностью анализа новых источников данных: пространственных выборок объектов (индивидуумов, домохозяйств, предприятий и т.п.), наблюдаемых в течение некоторого периода времени. Такие пролонгированные пространственные выборки, где каждый объект наблюдается многократно (например, ежегодно) на протяжении отрезка времени, получили название ***панельных данных***.

По словам лауреата Нобелевской премии 2000 г. Джеймса Хекмана [Heckman, 2001], создание подобных баз данных — это главное достижение XX в. Использование этих источников открыло новые перспективы в развитии экономической науки и математических методов, обслуживающих ее.

Смысл высказывания Хекмана состоит в следующем.

Ранние эконометрические модели, опиравшиеся на данные пространственных выборок, или временных рядов, носили агрегированный характер и описывали поведение усредненных объектов, для которых Альфред Маршалл ввел специальные термины: «репрезентативный потребитель» или «репрезентативная фирма». Со временем выяснилось, что эти модели часто оказывались не слишком эффективными инструментами для анализа экономических явлений и выработки рекомендаций по социально-экономической политике. Очень часто ни значения, ни знаки коэффициентов, посчитанных по регрессиям для агрегированных временных рядов, не соответствовали предположениям экономической теории, так как возникало серьезное смещение агрегирования. Об этом писали и Тейл в 1954 г., и Грин в 1964-м, и Фишер в 1969-м.

Одним из решений проблемы виделась разработка программы сбора комбинированных макро- и микроданных, с которой выступил Оркутт в 1964 г. Усилия Оркутта послужили импульсом, кото-

рый привел в движение силы, создавшие современную ***микроэконометрику*** и один из ее разделов — ***анализ панельных данных***.

Основным источником микроэкономических данных служат национальные репрезентативные опросы, проведение которых весьма дорогостоящая акция. Чтобы инициировать такого рода деятельность, необходимо наличие серьезных мотивов. Этими мотивами стали, во-первых, потребность в исследованиях, выявляющих причины социально-экономических проблем, способных нарушить стабильность уклада общественной жизни, а во-вторых, спрос на социальные программы, адресованные непосредственно тем или иным специфическим проблемным группам.

Основной целью моделей, создаваемых на базе микроданных в 1960–1970-х годах, было изучение старой политики в новых условиях или предсказание возможных эффектов новой, никогда ранее не проводимой политики. Особенно пристальное внимание экономистов в эти годы занимал рынок труда. Попытки использования неоклассической теории для его описания вызвали потребность в данных индивидуального уровня и методах анализа и интерпретации зависимостей, получаемых на их основании.

Когда быстро растущий уровень развития вычислительной техники позволил оперативно оценивать сотни разнообразных регрессионных моделей, появился спрос на методы выявления среди множества этих часто взаимно противоречивых результатов таких, которые поддавались бы прозрачной экономической интерпретации. Помимо этого отобранные модели должны были при минимальной размерности вмещать все богатство и разнообразие информации, поставляемой новым типом данных.

Теперь, в начале XXI в. можно констатировать (опять же по словам Хекмана), что развитие микроэконометрики привело к ряду важных эмпирических открытий.

Наиболее важное открытие — это очевидность того, что неоднородность и многообразие (экономических агентов и явлений) пронизывают экономическую жизнь, и, следовательно, они должны непременно учитываться в эконометрических моделях.

Второй важный результат — появление новых моделей экономических явлений — моделей анализа панельных данных, которые предоставляют разнообразные возможности учета неоднородности.

1.2. Описание наиболее употребимых источников панельных данных

Панельные обследования в той или иной форме проводятся практически во всех экономически развитых странах, однако впервые сбор панельных данных начался в США.

В настоящее время наиболее востребованными можно назвать базы NLS (National Longitudinal Surveys of Labor Market Experience) и PSID (University of Michigan's Panel Study of Income Dynamics). О них следует сказать несколько слов, поскольку примеры анализа этих данных часто используются в различных учебниках и научных публикациях.

База NLS содержит данные по различным сегментам рабочей силы: мужчины и юноши в возрасте от 45 до 59 лет и от 14 до 24-х в 1966 г., женщины и девушки от 30 до 44 лет в 1967 г. и от 14 до 24-х в 1968 г., и молодежь обоих полов, которым исполнилось от 14 до 21 года в 1979 г. Первые четыре сегмента периодически опрашивали в течение 15 лет, последний сегмент продолжает наблюдаться. Перечень наблюдаемых характеристик насчитывает 1000 наименований с точки зрения рыночного предложения рабочей силы.

База PSID возникла с ежегодного сбора данных репрезентативной национальной выборки, охватывающей около 6000 семей и 15 000 индивидуумов, в 1968 г. и пополняется до сих пор. Данные содержат около 5000 характеристик, включая занятость, доход, переменные человеческого капитала, жилищные условия, мобильность и т.п.

В России сбор панельных данных начался в 90-е годы XX в.

Примерами панельных данных о российской экономике являются RLMS (Russia Longitudinal Monitoring Survay) или в русской аббревиатуре РМЭЗ (Российский мониторинг экономического положения и здоровья населения), Российский экономический тренд (доступные бесплатно по Интернету) и Российский экономический барометр — платная база данных. На РМЭЗ имеет смысл остановиться особо, поскольку эти данные очень широко используются исследователями и в России, и за рубежом.

РМЭЗ представляет собой серии общенациональных, репрезентативных опросов, регулярно проводимых с 1992 г. с целью систематического наблюдения воздействия российских реформ на дина-

мику экономического благосостояния домохозяйств и отдельных индивидов. Опросы проводятся международным консорциумом организаций при участии Института социологии РАН. Подробная информация о РМЭЗ и первичные данные представлены на сайте: <http://www.cpc.unc.edu/projects/rlms/home.html>. В базе данных РМЭЗ приведены результаты опросов свыше 10 000 человек. Информация, собранная в РМЭЗ, касается размеров, источников и структуры доходов и расходов домохозяйств, занятости, распределения времени, уровня образования, состояния здоровья и других характеристик (всего свыше 500 показателей).

Собираемая информация имеет двухуровневую структуру.

1. Информация индивидуального уровня — индивидуальные файлы:

- данные из всех взрослых и детских анкет;
- общая статистическая информация (регион, тип населенного пункта и т.п.) для каждого человека, участвовавшего в исследовании;
- некоторые сводные индивидуальные индексы (образование, профессиональная группа и т.п.);
- показатели участия данного человека в предыдущих и последующих исследованиях.

2. Информация уровня домохозяйства (семьи) — семейные файлы:

- данные семейных анкет;
- общая статистическая информация (регион, тип населенного пункта и т.п.) для каждой семьи;
- показатели участия данной семьи в предыдущих волнах исследования.

А вот как выглядит неполный перечень исследований, в которых были использованы данные РМЭЗ [Материалы конф. РМЭЗ, 2003]:

- Анализ сберегательного поведения российских домохозяйств.
- Незанятость в России: вынужденная или добровольная.
- Субъективные и объективные оценки здоровья населения.
- Бедность в России: масштабы и структурные особенности.
- Измерение продолжительности бедности в России.
- Экономический анализ причин вторичной занятости.
- Микроэкономический анализ динамических изменений на российском рынке труда.

- Распространенность курения в России.
- Проблема алкоголизма в России.
- Рабочее время как ресурс благосостояния.
- Динамика среднего класса в России 1990-х годов.
- Экономическая эффективность высшего образования.
- Финансовое поведение домохозяйств: сбережение, инвестирование, кредитование, страхование.
- Толерантность и динамика социального самочувствия в современном российском обществе.
- Гендерные аспекты инвестиций в человеческий капитал в современной России.
- Мобильность населения по доходам как механизм изменения неравенства.
- Роль государства и семьи в экономической поддержке пожилых людей в Российской Федерации.
- Человеческий капитал в России: модели текущих и пожизненных расходов.
- Сравнительная ценность различных форм человеческого капитала в России.
- Эволюция социального самочувствия россиян и особенности социально-экономической адаптации.
- Трудовая незащищенность и задолженность по заработной плате в Российской Федерации.
- Социально-экономические факторы феминизации бедности в России.
- Женщины в сфере занятости и на рынке труда в российской экономике.
- Анализ затрат домохозяйств на здравоохранение.
- Экономический статус и здоровье человека.
- Интерпретация скачка смертности в России.
- Доходы и занятость.

1.3. Преимущества использования панельных данных

Пролонгированная, или панельная, совокупность данных представляет собой пространственную выборку объектов, прослеживае-

мую во времени, и таким образом предоставляет множество наблюдений над каждым отдельным объектом. Панели можно создавать, объединяя вместе готовые временные ряды (как правило, так строятся панели стран и регионов).

Основные преимущества данных этого типа в следующем, они:

1) предоставляют исследователю большое количество наблюдений, увеличивая число степеней свободы и снижая зависимость между объясняющими переменными и, следовательно, стандартные ошибки оценок;
2) позволяют анализировать множество экономических вопросов, которые не могут быть адресованы к временным рядам и пространственным данным в отдельности;
3) позволяют предотвратить смещение агрегированности, неизбежно возникающее как при анализе временных рядов (где рассматривается временная эволюция усредненного «репрезентативного» объекта), так и при анализе перекрестных данных (где не учитываются ненаблюдаемые индивидуальные характеристики объектов и предполагается ***однородность***, всех коэффициентов регрессии);
4) дают возможность проследить индивидуальную эволюцию характеристик всех объектов выборки во времени;
5) решают проблему поиска «хороших» инструментов при оценивании моделей с эндогенными (т.е. коррелированными со случайными ошибками) регрессорами;
6) дают возможность избежать ошибок спецификации, возникающих от невключения в модель существенных переменных.

Поясним все вышесказанное следующими примерами.

Трудности с выводами о динамике изменения каких-либо объектов из пространственных наблюдений хорошо иллюстрируются следующей ситуацией на рынке труда. Рассмотрим влияние профсоюзных объединений на экономическое поведение рынка.

Одна группа экономистов, которая намеревается интерпретировать наблюдаемые различия между фирмами, где есть профсоюз и где его нет, полагает, что союзы и коллективно осуществляемые процессы фундаментально меняют ключевые аспекты соотношений занятости: компенсацию, внутреннюю и внешнюю мобильность труда, порядок работы и окружение. Другая группа экономистов рассма-

тривает эффекты от объединения как иллюзорные попытки противостояния совершенной конкуренции, условиям которой достаточно близко удовлетворяет реальный мир. Эти экономисты полагают, что наблюдаемые различия существуют главным образом благодаря различиям, предшествующим объединению или возникшим после. Профсоюзы не способствуют повышению заработной платы в долгосрочном периоде, потому что фирмы на это повышение реагируют повышением требований к качеству работников. Если одни полагают, что коэффициент при фиктивной переменной, отражающей статус участия в профсоюзе в уравнении заработной платы, есть мера эффекта от объединения, то другие считают, что этот коэффициент просто отражает уровень квалификации работника.

Модели, основанные только на пространственных данных, обычно не могут позволить выбрать верную гипотезу из этих двух, так как оценки отражают межиндивидуальные различия только в данный момент. При использовании панельных данных можно различить эти две ситуации, изучая разницу в заработной плате работника, движущегося от фирмы без профсоюза к фирме с профсоюзом. Если эффекта от участия в профсоюзе нет, то не будет меняться и заработная плата, и наоборот. Прослеживая данные фирмы до и после создания на ней профсоюза, можно сконструировать модель, измеряющую эффект от деятельности профсоюза.

Рассмотрим пример абстрактной модели с распределенными лагами:

$$y_t = \sum_{\tau=0}^{n} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, ..., T. \qquad (1.3.1)$$

Как правило, в таких моделях возникает проблема квазимультиколлинеарности между $n + 1$ объясняющими переменными $x_t, x_{t-1}, ..., x_{t-n}$. Таким образом, нет достаточной информации, чтобы получить точные оценки некоторых коэффициентов при лаговых переменных без априорного предположения о том, что они являются функциями небольшого числа параметров.

Когда есть панельные данные, мы можем использовать индивидуальные различия в величинах x, чтобы снизить проблему мультиколлинеарности. Более того, доступность пространственных массивов данных позволяет использовать различные предварительные ограничения на коэффициенты при лаговых регрессорах $\{\beta_\tau\}$.

Помимо того, что панельные данные позволяют конструировать и тестировать более сложные поведенческие модели, чем чистые пространственные данные или временные ряды, использование панельных данных позволяет снижать размерность моделей и дает средство разрешения некоторых ключевых эконометрических проблем. Например, такой проблемой является понять, заключается ли причина наблюдаемого эффекта в пропущенных (неверно измеренных, ненаблюдаемых) переменных, которые коррелированны с объясняющими переменными?

Рассмотрим в качестве примера простую модель:

$$y_{it} = \alpha + X'_{it}b + Z'_{it}\gamma + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \qquad (1.3.2)$$

где X'_{it} и Z'_{it} — векторы-строки объясняющих переменных; β, γ — векторы коэффициентов; случайная ошибка u_{it} подчиняется обычным предположениям теоремы Гаусса — Маркова.

Если модель (1.3.2) верно специфицирована, то метод наименьших квадратов (МНК) дает несмещенную и состоятельную оценку α, β и γ.

Предположим, что переменные X'_{it} — наблюдаемы, а Z'_{it} — ненаблюдаемы, и $\mathrm{cov}(X_{it}, Z_{it}) \neq 0$. Тогда оценки коэффициентов регрессии y на X будут смещены. Однако, если доступны повторяющиеся наблюдения для групп индивидуумов, они могут позволить нам выявить нежелательный (смещающий оценки при переменных X_{it}) эффект от невключения Z и устранить его. Пусть $Z_{it} = Z_i$ для $\forall t$. Мы можем перейти к 1-м разностям по времени:

$$y_{it} - y_{i,t-1} = (X'_{it} - X'_{i,t-1})\beta + (u_{it} - u_{i,t-1}), \quad i = 1, ..., N, \quad t = 2, ..., T.$$

Можем также взять отклонение от среднего:

$$y_{it} - \overline{y}_t = (X'_{it} - \overline{X}'_t)\beta + (u_t - \overline{u}_t), \quad i = 1, ..., N, \quad t = 1, ..., T,$$

где $\overline{y}_t = \dfrac{1}{N}\sum_{i=1}^{N} y_{it}$ и т.д.

Теперь оценка МНК $\hat{\beta}$ будет несмещенной (и этому не препятствует автокоррелированность случайных ошибок в преобразованных моделях).

Если бы мы имели только пространственные данные ($T = 1$) для ($Z_{it} = Z_i$) или только временной ряд ($N = 1$) для ($Z_{it} = Z_t$), такое бы было невозможным. Часто в этих случаях приходится использовать метод инструментальных переменных с инструментом, коррелирующим с X, но некоррелированным с Z и u. Найти такой инструмент, как правило, довольно сложно.

В работе Маккарди [MaCurdy, 1981] по жизненным циклам в предложении труда мужчин приведена хорошая иллюстрация вышеизложенного. При определенных упрощающих предположениях Маккарди показал, что функция предложения труда может быть записана в виде (1.3.2), где y — логарифм рабочих часов, X — логарифм реальной ставки заработной платы, Z — логарифм предельной полезности начального благосостояния работника. Z является ненаблюдаемой переменной и обусловливается суммарной величиной заработной платы работника и дохода от собственности за всю его жизнь к моменту начала наблюдения. Поэтому $Z_{it} = Z_i$. В этой задаче не только X коррелирует с Z, но и любая другая экономическая переменная (образование и т.п.). Следовательно, нельзя оценить β состоятельно из пространственных данных, но переходом к 1-м разностям по времени в панельных данных получаются состоятельные оценки.

1.4. Проблемы использования панельных данных

1.4.1. Гетерогенное смещение

Привлекательность панельных данных обусловливается теоретической возможностью элиминировать в регрессионной модели влияние некоторых специфических трудно измеряемых факторов, например, политики.

Если данные генерируются простым контролируемым экспериментом, то могут быть применены стандартные статистические методы. К несчастью, большая часть панельных данных поступает из очень сложных процессов повседневной экономической жизни. Типичное предположение, что y генерируется параметрической функцией распределения вероятностей $P(y|\theta)$, где θ — m-мерный

действительный вектор, один и тот же для всех индивидуумов и во все времена, может быть нереальным. Игнорирование таких гетерогенных параметров может привести к несостоятельности оценок.

Рассмотрим следующую модель:

$$y_{it} = \alpha_i + \beta_i X_{it} + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \qquad (1.4.1)$$

где X — единственная экзогенная переменная; случайная ошибка u_{it} подчиняется обычным предположениям теоремы Гаусса — Маркова.

Параметры α_i и β_i могут быть различны для различных индивидуумов, хотя и могут оставаться постоянными во времени. Следовательно, будут встречаться различные выборочные распределения, которые могут серьезно смещать регрессию y_{it} на X_{it}, оцененную по всем NT-наблюдениям и игнорирующую индивидуальную неоднородность коэффициентов модели (1.4.1).

Вышесказанное можно проиллюстрировать следующими примерами:

1. Гетерогенный (неодинаковый) для различных индивидуумов свободный член и гомогенный (одинаковый) наклон: $\alpha_i \neq \alpha_j$, $\beta_i = \beta_j$ для $\forall\, i, j$ (рис. 1.1).

Во всех этих ситуациях сквозная регрессия[1], игнорирующая гетерогенность константы, является смещенной, причем направление смещения не может быть диагностировано априорно.

2. И свободный член, и наклон гетерогенны: существуют такие i, j, для которых $\alpha_i \neq \alpha_j$, $\beta_i \neq \beta_j$ (рис. 1.2).

На рис. 1.2*а* изображена ситуация, когда сквозная регрессия приводит к бессмысленному результату, так как индивидуальные направления (коэффициенты наклона) существенно различаются. На рис. 1.2*б* некий смысл сквозной регрессии имеется, но приводит к ложным результатам о криволинейности сквозного соотношения.

[1] Здесь и далее мы так будем переводить англоязычный термин «pooled», под которым подразумевается регрессия, оцененная без учета особой (панельной) структуры данных.

Рис. 1.1

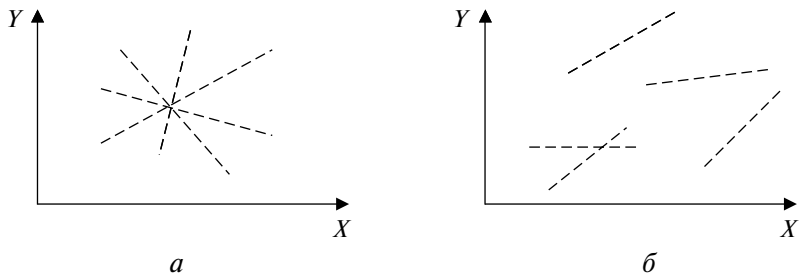

Рис. 1.2

Аналогичные примеры можно привести в случае, когда свободный член и наклон изменяются со временем и одинаковы для индивидуумов.

1.4.2. Смещение самоотбора

Другой распространенный источник смещения — неслучайная выборка. Например, известный факт, что в данных РМЭЗ практически нет наблюдений, относящихся к индивидуумам из высокодоходных групп населения. Когда такие неполные данные используются в качестве зависимой (объясняемой) переменной, это может повлечь за собой смещение самоотбора. Чтобы это продемонстри-

ровать, рассмотрим пример с пространственными данными. Пусть модель сформулирована так:

$$y_i = X_i'\beta + u_i, \quad i = 1,\ldots,N, \quad E(u_i) = 0, \quad D(u_i) = \sigma_u^2 I,$$

где y — заработная плата; X — набор экзогенных переменных, включая образование, интеллект и т.д.; I — единичная диагональная матрица.

Причем при $y_i = X_i'\beta + u_i \le L$ — индивидуумы включаются в выборку, при $y_i > L$ — исключаются.

Для простоты теперь предположим, что все экзогенные переменные принимают одни и те же значения для всех наблюдений, кроме образования (которое измеряется как продолжительность обучения) (рис. 1.3).

Из приведенного схематического рисунка видно, что линия регрессии, построенная по усеченным данным, будет иметь меньший угол наклона, чем ее аналог, который мог быть получен по полной выборке. Таким образом, влияние образования оказывается недооцененным. Это происходит потому, что в данных выборок такого типа появляется корреляция между объясняемой переменной y_i и случайной ошибкой u_i, что ведет к недооценке или переоценке влияния зкзогенных переменных.

Смещение самоотбора при анализе панельных данных часто является следствием истощения выборки, т.е. постепенного убывания числа объектов наблюдения. Истощение панели — это типичное явление. Панели домохозяйств могут истощаться из-за перемещений, распадов семей, а также отказов участвовать в опросах в дальнейшем. Если выбытие происходит по случайным причинам, смещения самоотбора может и не быть, но если существуют некие скрытые закономерности, то смещение неизбежно. Например, при повышении уровня доходов у домохозяйства могут пропасть стимулы участвовать в опросе, и тогда в выборке будут оставаться низкодоходные слои населения, что сделает выборку нерепрезентативной.

Перечисленные проблемы могут быть разрешены с помощью некоторых специальных приемов, которые подробно будут изложены в главе 10. Это может быть переход или к несбалансированным панелям, где разные индивидуумы наблюдаются в течение

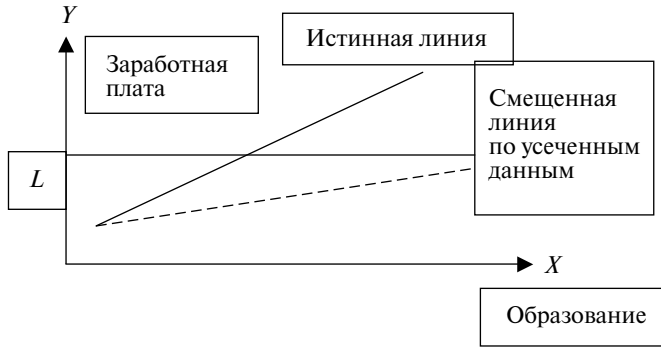

Рис. 1.3

различного числа тактов времени, или к панелям с замещением, где выбывшие объекты заменяются новыми, или использованием псевдопанелей, где в качестве объектов наблюдения выступают не отдельные индивидуумы, а группы индивидуумов со схожими (в некотором смысле) характеристиками. Хотя, конечно, это осложняет процесс оценивания.

Для решения проблемы самоотбора при исследовании пространственных выборок используют модель Хекмана. В настоящее время появились разработки, обобщающие эту модель для анализа панельных данных.

К часто встречающимся недостаткам панелей можно отнести также немногочисленность наблюдений, составляющих временные ряды для отдельных индивидуумов.

Конец ознакомительного фрагмента
Полная версия книги доступна на litres.ru