

Прикладной эконометрический анализ  
в статистическом пакете Stata

*Станислав Колеников*

E-mail: skolenik@nes.ru

Российская Экономическая Школа, 2000

© С. О. Колеников

In theory, theory and practice are the same. In practice, they are not.

Теоретически, теория и практика — это одно и то же, но на практике так не получается.

*(Подмечено на стене в Лаборатории математического моделирования в экологии и медицине ВЦ РАН п/р В.В.Шакина.)*

# Оглавление

<b>1</b>	<b>Введение</b>	<b>7</b>
<b>2</b>	<b>Регрессионные модели</b>	<b>12</b>
2.1	Применение статистических методов в экономических исследованиях . . . . .	12
2.2	Классическая модель линейной регрессии . . . . .	16
2.2.1	Обозначения и формулировки . . . . .	16
2.2.2	Метод наименьших квадратов . . . . .	18
2.2.3	Проверка гипотез . . . . .	19
2.3	Нарушения предположений классической модели . . . . .	21
2.3.1	Нецентральность . . . . .	21
2.3.2	Стохастичность регрессоров . . . . .	21
2.3.3	Гетероскедастичность остатков . . . . .	23
2.3.4	Автокоррелированность остатков . . . . .	26
2.3.5	Мультиколлинеарность . . . . .	27
2.3.6	* Проблема робастности . . . . .	31
2.3.7	Преобразование к нормальности и линейности . . . . .	34
2.4	Прочие отклонения от модели . . . . .	35
2.4.1	Спецификация модели: выбор нужных переменных . . . . .	35
2.4.2	Нелинейность . . . . .	38
2.4.3	Идентификация резко выделяющихся наблюдений . . . . .	39
2.4.4	Визуальный анализ . . . . .	43

2.4.5	Множественная проверка гипотез . . . . .	46
2.4.6	Данные с пропусками . . . . .	47
2.5	Диагностика регрессий . . . . .	51
2.5.1	Сводка методов диагностики . . . . .	52
2.5.2	Пример анализа регрессии . . . . .	55
2.6	Альтернативные спецификации статистических зависимостей . . . . .	60
2.6.1	Данные особой структуры и обобщенный МНК . . . . .	61
2.6.2	Системы одновременных уравнений . . . . .	63
2.6.3	Модели с дискретными и другими ограниченными зависимыми переменными . . . . .	63
2.6.4	Квантильные регрессии . . . . .	66
2.6.5	Непараметрические регрессии . . . . .	67
<b>3</b>	<b>Краткое описание пакета Stata</b>	<b>70</b>
3.1	Обозначения . . . . .	71
3.2	Установка и запуск пакета Stata . . . . .	72
3.3	Интерфейс Stata . . . . .	73
3.4	Общий вид команд Stata . . . . .	75
3.5	Помощь . . . . .	76
3.6	Условные модификаторы . . . . .	77
3.7	Работа с файлами . . . . .	77
3.8	Работа с данными . . . . .	79
3.9	Основные статистические средства . . . . .	82
3.10	Функции . . . . .	85
3.11	Повторяемые фрагменты . . . . .	85
3.12	Результаты работы . . . . .	86
3.13	Программы . . . . .	87
3.14	Графика . . . . .	90
3.15	Информационные команды . . . . .	92
3.16	Internet-возможности Stata . . . . .	93
3.17	Расширение возможностей Stata . . . . .	95

3.18	Сообщения об ошибках . . . . .	96
3.19	Прочее . . . . .	98
3.20	С чего начать? . . . . .	99
<b>4</b>	<b>Мониторинг экономического положения и здоровья населения России</b>	<b>101</b>
<b>5</b>	<b>Заключение</b>	<b>106</b>
<b>6</b>	<b>Домашние задания</b>	<b>107</b>
	<b>Литература</b>	<b>110</b>

# Глава 1

## Введение

Данный текст — это материалы к семинарам по прикладной эконометрике, проведенным весной 2000 г. на экономических факультетах Воронежского Государственного Университета и Уральского Государственного Университета (Екатеринбург) в рамках программы повышения квалификации преподавателей экономических вузов на базе Центра дополнительного профессионального образования Российской Экономической Школы (<http://www.nes.ru/english/outreach/outreach.htm>).

Основной акцент изложения сделан на прикладных аспектах эконометрического анализа. В частности, освещаются такие проблемы, как выбор спецификации эконометрической модели, нарушения предположений классической модели множественной линейной регрессии, методы диагностики регрессий, а также приводятся дополнительные сведения о наиболее часто используемых в литературе методах анализа экономических зависимостей. Никаких теорем не доказывается, хотя ссылки на теоретическую литературу в нужных местах приводятся. Нестрогость изложения не должна вводить в заблуждение: корректное применение даже достаточно простых эконометрических методов невозможно без достаточного знания теории, поэтому данная книга не может служить введением в эконометрику<sup>1</sup>.

Практическая реализация обсуждаемых методов выполнена в пакете Stata Stata-

---

<sup>1</sup>В качестве подобного введения автор может порекомендовать Я. Магнус, П. К. Катыхев, А. А. Пересецкий. Эконометрика. Начальный курс. М., “Дело”, 1997; более продвинутое изложение можно найти в С. А. Айвазян, В. С. Мхитарян. Прикладная статистика и основы эконометрики. М., ЮНИТИ, 1998.

Corp. (1999, 2001), Kolenikov (forthcoming). Параллельно с изложением теоретических результатов и подходов приводятся ссылки на соответствующие команды пакета. Этот пакет популярен среди прикладных экономистов как в России, так и за рубежом, благодаря его открытости и обширному набору средств эконометрического анализа. На практических занятиях, а также в экзаменационных работах использовались данные Мониторинга здоровья и экономического положения домохозяйств России (RLMS). Основные сведения, необходимые для работы с этой базой данных, также приводятся в этой книге. Эти данные ценны тем, что они являются практически единственным открытым источником микроэкономических данных по России (<http://www.cpc.unc.edu/rlms/>).

#### Stata

Врезки, в которых будут указываться и описываться необходимые команды Stata, будут оформлены так, как этот абзац. Как правило, описание будет весьма кратким. Более подробную информацию о любой команде Stata можно получить через встроенную систему помощи. Для этого надо войти в меню Help/Search или Help/Command или набрать на клавиатуре `whelp имя команды`, например, `whelp regress`. Идеалом, безусловно, является обращение к первоисточникам — руководствам пользователя.

Анализ данных — это скорее искусство (или по меньшей мере ремесло), нежели точная наука, и автор надеется, что рекомендации, даваемые в этой книге, не будут возведены в ранг абсолютной истины. Практика показывает, что данные могут вести себя как угодно, и тесты, хорошо работающие в одних условиях, будут совершенно бесполезны в других, и разные тесты, пытающиеся уловить один и тот же эффект (например, гетероскедастичность), могут давать совершенно противоположные результаты.

Книга построена следующим образом. В главе 2 приводятся основные понятия и результаты вводных курсов эконометрики, связанные с концепцией линейной регрессии и метода наименьших квадратов, рассматриваются возможные варианты развития и дополнения этой базовой концепции. В главе 3 приводятся основные команды пакета Stata и пользовательские приемы, упрощающие работу с пакетом. Далее в главе 4 дается краткое введение в базу данных RLMS — ее основные характеристики и базовые ориентиры для работы. Небольшое заключение в главе 5 подводит основные итоги книги. И, наконец, в главе 6 приводятся домашние задания, выдававшиеся участником семинара. Читатель может использовать их для самоконтроля.

Возможны разные варианты прочтения этой книги. Читатель, пользующийся дру-

гим эконометрическим или статистическим пакетом, вряд ли нуждается в гл. 3, и, скорее всего, захочет просмотреть основные методы диагностики регрессий в параграфах 2.3–2.4, сведенные в удобную табличку тестов на стр. 53, с которой пользователи, возможно, будут консультироваться весьма часто. Наиболее любопытные читатели доберутся до параграфа 2.6, посвященного эконометрическим моделям, выходящим достаточно далеко за рамки модели линейной регрессии. В частности, в этом разделе дается краткое введение в модели логит, пробит, в анализ панельных данных, и пр. Этот материал содержит минимальную информацию как о сути упоминаемых методов, так и об их реализации в пакете Stata.

Напротив, читатель, перед которым стоит задача как можно быстрее разобраться, “как же работает эта чертова программа”, сосредоточит свое внимание на главе 3. Она дает общее представление о том, что и как надо делать, чтобы ввести данные, преобразовать их к нужному виду, оценить свою статистическую модель и перенести результаты в любимый редактор для подготовки публикации. Совершенно необходим для дальнейшего чтения вводный раздел обозначений 3.1. Следующий по важности и общности материал — как записываются вообще команды Stata (параграфы 3.3–3.6). Далее команды и элементы синтаксиса Stata сгруппированы по основным видам (работа с файлами, преобразование данных, вывод результатов, средства программирования, графика). Список команд (примерно на полторы страницы), соответствующих основным эконометрическим моделям, приводится в разделе 3.9 (с. 83). В разделе 3.20 предложены средства самообучения и начала работы в пакете Stata.

Исследователям-практикам, а также преподавателям, придумывающим задачи и курсовые работы для студентов, будет полезна глава, посвященная RLMS — основному источнику экономических микроданных по домохозяйствам России.

При чтении книги может создаться впечатление, что она перегружена отдельными деталями, при том, что многие концепции и методы упомянуты лишь вскользь. Автор намеренно шел на это: учитывая низкую насыщенность рынка эконометрической литературы на русском языке, я счел полезным предоставить хотя бы минимальную информацию о моделях и методах, вообще не упоминаемых в начальных курсах теоретической эконометрики, но встречающихся достаточно часто в прикладной работе и научных публикациях, в надежде, что исследователь, пользующийся этой книгой, сможет найти о них более подробную информацию и применить в своей работе метод,



адекватный задаче.

Автор выражает благодарность всем тем, без кого эта книга не появилась бы, появилась бы позже или в значительно худшем виде: Сергею Гуриеву, руководителю Центра дополнительного профессионального образования РЭШ, за идею по проведению этого курса и написания книги, а также за помощь в подготовке текста; Сергею Артемьевичу Айвазяну, моему научному руководителю в аспирантуре Центрального экономико-математического института, за ценные замечания и научную поддержку; Эрику Берглофу, директору Российско-Европейского центра по экономической политике, за поддержку в ходе работы над семинарами и книгой; Анне Хмелевской, Ирине Щепиной и Инне Мальцевой за организацию семинаров в Воронеже и Екатеринбурге; Сергею Голованю за помощь с `TeX`ом; всем читателям этой книги и слушателям семинаров за их вопросы и замечания; компании Stata Corporation за замечательный пакет; Университету Северной Каролины, компании Paragon и Институту социологии РАН за проведение и публикацию данных Мониторинга здоровья и экономического положения домохозяйств России. Работа была профинансирована в рамках проекта поддержки кафедр программы “Высшее образование” Мегaproекта “Развитие образования в России” Института Открытое Общество, гранты N НВС 807, 808.

21 декабря 2000 г. начались поставки новой — седьмой — версии пакета Stata. Эта версия сохраняет совместимость с предыдущими версиями, однако содержит и много новых и приятных особенностей. На том уровне изложения, который был в целом принят в этой брошюре, самые заметные отличия — поддержка более длинных имен переменных и программ (до 32 символов), улучшенные средства поиска в Интернет; объединение функций окна подсказки и вывода результатов (“кликабельность” окна результатов) при помощи внутреннего языка SMCL (Stata Markup and Control Language), родственного с другими языками разметки (HTML, SGML); улучшенная (наконец-то) графика, в т.ч. разные стили линий (пунктирные и т.п.); новые средства кластерного анализа; дальнейшее усиление средств анализа панельных данных; наконец, общее ускорение работы за счет использования новых компиляторов. Описание новых возможностей имеется на корпоративном сайте по адресу <http://www.stata.com/stata7>.

Станислав Колеников,

РЭШ, ЦЭМИ, РЕЦЭП, ЦЭФИР, Университет Северной Каролины (Чапел Хилл), 1999-2001.

E-mail: [skolenik@nes.cemi.rssi.ru](mailto:skolenik@nes.cemi.rssi.ru), [skolenik@unc.edu](mailto:skolenik@unc.edu), [stas4stata@yahoo.com](mailto:stas4stata@yahoo.com)

# Глава 2

## Регрессионные модели

### 2.1 Применение статистических методов в экономических исследованиях

В настоящее время в России все большее признание находит подход к анализу экономических явлений, опирающийся на аналитические системы теоретической экономики и использующий математический аппарат как для построения теоретических моделей, так и для анализа данных.

Прикладные экономические исследования обязательно включают в себя обработку статистических данных — макроэкономических временных рядов, бюджетов домохозяйств, характеристик экономической деятельности предприятий и т. д. Статистика и эконометрика, понимаемые как научные методы обработки данных, могут при этом служить различным целям<sup>1</sup>:

1. *Исследование данных, разведочный анализ и диагностика*. При данном подходе к анализу данных исследователь позволяет данным направлять исследование (data-driven research). Отталкиваясь от данных (и пользуясь аппаратом мат. статистики и эконометрики) при самых минимальных модельных допущениях, исследователь делает вывод о наличии статистических соотношений (корреляций) между рядами экономических показателей, о наличии единичных корней в финансовых времен-

---

<sup>1</sup> Очень хорошее введение в проблематику статистического анализа зависимостей в эконометрике можно найти в Айвазян, Мхитарян (1998, гл. 10.)

ных рядах, о группировании данных в кластеры и т. д. — о наличии в данных внутренней структуры.

2. Достаточно близко к этому примыкают методы обработки данных, возникшие в 1990-х гг. и объединяемые названием *data mining* (что можно перевести на русский как “обогащение данных”, по аналогии с процессами обогащения руды в горном деле). Эта область находится на стыке информационных технологий и статистики и, как правило, имеет дело с объемами данных, исчисляемыми мега- и гигабайтами. Разрабатываемые в ее рамках алгоритмы направлены на поиск в данных повторяющихся фрагментов и шаблонов (*patterns*). В эконометрической практике эти методы пока что еще не встречаются. *Data mining* не ставит задачи оценки статистической достоверности получаемых результатов, что в определенной мере снижает их ценность для научных исследований.
3. *Верификация теоретических моделей*. Здесь во главу угла ставится теоретическая модель, которую экономист хочет проверить на практике. Она должна быть представима в виде, допускающем эконометрическую проверку — например, сформулированы результаты сравнительной статики, временной ряд разложен в соответствии с предполагаемой лаговой структурой, производственная функция или функция полезности потребителя представлены в удобном аналитическом виде, и т. п. Иногда в качестве подтверждения теоретической модели исследователи довольствуются корреляциями (частными корреляциями, свободными от (линейного) вклада прочих переменных, в многомерных задачах), т. е. знаками коэффициентов регрессионной модели.

В подавляющем большинстве случаев приходится довольствоваться ретроспективными (т. е. уже наблюдаемыми) данными, а не планировать и проводить эксперимент, как это возможно в естественнонаучных отраслях; при этом данные, которыми располагает исследователь, могут не вполне точно соответствовать переменным теоретической модели, а некоторые переменные могут и вовсе быть ненаблюдаемы, и исследователю приходится изобретать те или иные приближения (проху) к нужным параметрам (например, квалификация работника сама по себе может не быть наблюдаема, однако в качестве аппроксимации квалификации могут выступать уровень образования — среднее, высшее, техникум, и т.п. — или общая

продолжительность обучения). Модель теоретическая, таким образом, достаточно жестко обуславливает модель эконометрическую, предписывая определенные спецификации, включающие в себя требуемые переменные.

После того, как все необходимые предварительные действия проведены — построена теоретическая модель, сформулирована эконометрическая спецификация, выработаны проверяемые гипотезы исследования, собраны и подготовлены данные — исследователь с помощью эконометрических и статистических методов принимает или отвергает гипотезы о наличии и виде зависимости между экономическими переменными, о значениях определенных параметров модели, и т.п.

4. *Построение и идентификация моделей.* Часто возникают ситуации, когда перед исследователем стоит задача выбора какой-то одной модели из ряда имеющихся. Например, на основную исследуемую переменную может влиять много факторов, и исследователь хочет выделить наиболее существенные. Так, цена на жилье определяется в первую очередь его размером — количеством комнат, общей площадью, однако есть дополнительные факторы: наличие телефона, лифта, совмещенный или раздельный санузел, этаж дома, тип дома, недавний ремонт, престижный район и т.п. Другим примером выбора модели из нескольких возможных может служить выбор автокорреляционной структуры временного ряда (ARMA модель). В таких задачах исследователь оценивает (идентифицирует) каждую из моделей и по определенным критериям сравнивает полученные модели.

Для дотошного читателя сделаем следующие ремарки. Следует иметь в виду, что теоретические свойства оценок коэффициентов в выбираемых таким образом моделях отличаются от свойств оценок, характерных для заранее фиксированных моделей, и точных результатов в данной области пока что нет.

С выбором “лучших” вариантов связано явление *publication bias* (смещенность публикуемых результатов), которое заключается в том, что для публикации в научном журнале скорее будет выбрана работа, в которой показаны статистически значимые результаты, чем работа, в которой эксперимент не привел к значимым результатам. Эти и подобные эффекты исследуются в рамках *мета-анализа* — дисциплины, исследующей связь различных публикаций и возможности извлечения информации за счет объединения статистических результатов, полученных в

разных исследованиях на одну и ту же тему.

5. *Построение прогнозов.* Для построения хороших прогнозов нужно иметь (вычислительно) хорошую модель прогнозируемых процессов, и для решения данной задачи естественно привлекать лучшее из вышеупомянутых подходов. Далеко не всякая теоретическая модель хорошо описывает реальные данные; более того, для достаточно сложных процессов реального мира теоретических моделей может вообще не существовать. Поэтому для построения прогнозов (и, соответственно, для выбора прогнозирующих моделей) используются меры и критерии, связанные с качеством подгонки под данные (*goodness of fit*), зачастую без явного выдвижения статистических гипотез или анализа взаимосвязей между факторами (переменными), подразумеваемых выбранной прогностической моделью, и даже без формирования параметрической модели (т.е. непараметрическими методами, среди которых можно упомянуть ядерные оценки плотностей и линий регрессии или модели нейронных сетей).

Эта задача в определенной мере перекликается с предыдущей — в частности, если в качестве критериев отбора моделей используются критерии *goodness of fit* или перекрестной проверки (*cross-validation*).

Каждый из этих подходов имеет свои критерии “качества” конструируемых ими моделей. При разведочном анализе критерии обычно достаточно субъективны: обнаружены убедительные связи в данных или нет. *Data mining* в основном оперирует понятиями типа частот правильной классификации шаблонов. Выбор и идентификация моделей обычно базируются на информационных критериях или мерах качества подгонки, основанных на остаточных суммах квадратов. Прогнозные модели должны обеспечивать хорошее качество приближения при прогнозировании вне выборки (*out of sample prediction*).

Математически наиболее обоснованными являются статистические процедуры, опирающихся на результаты математической статистики, т.е. область анализа данных, названная выше “верификацией теоретических моделей”. Конечным результатом таких процедур обычно является мера достоверности статистических выводов — *уровень значимости*, или *доверительная вероятность*. В классических курсах статистики обычно проводится проверка строго сформулированных нулевых гипотез при уровне значимо-

сти 10%, 5% или 1%. Более интересная и более универсальная формулировка приводится в классической книге по математической статистике Кендалла и Стюарта ((Кендалл, Стюарт 1973)): "Любой критерий с уровнем значимости вплоть до [указанная цифра] отвергнет данную нулевую гипотезу".

Современная трактовка понятия доверительной вероятности в эконометрической литературе — это (условная) вероятность получить такие (или еще хуже, в контексте нулевой гипотезы) наблюдения в реальном эксперименте, если верна нулевая гипотеза. Для нулевой гипотезы эта вероятность должна быть вычислима аналитически, и именно поэтому в качестве нулевой гипотезы  $H_0$  в подавляющем большинстве случаев выступает простая гипотеза.

Одним из удобных и в то же время достаточно простых, а потому интенсивно используемых в прикладных эконометрических исследованиях, способов описания статистических зависимостей между (количественными) экономическими переменными является линейная регрессия.

## 2.2 Классическая модель линейной регрессии

### 2.2.1 Обозначения и формулировки

По определению, *регрессия* — это зависимость среднего значения случайной величины от некоторой другой величины или нескольких величин, или условное математическое ожидание Мат. энциклопедия (1984):

$$E[y|x] = f(x). \quad (2.1)$$

Таким образом, модель регрессии описывает вероятностное соотношение между *объясняющими переменными (регрессорами, независимыми переменными)* и *зависимой (результатирующей) переменной*. Естественным первым приближением для функции регрессии является ее линейаризация, и соответствующая модель носит название *модель линейной регрессии*. Предлагается следующее функциональное соотношение между реализовавшимся значением зависимой переменной и регрессорами:

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \dots, n \quad (2.2)$$

где  $y_i$  — зависимая переменная,  $x_i$  — вектор объясняющих переменных,  $x_i \in \mathbb{R}^p$ ,  $\beta$  — вектор параметров соответствующей размерности,  $\varepsilon_i$  — ошибка,  $i$  — номер наблюдения и  $n$  — общее количество наблюдений. Если объединить в столбцы данные по всем наблюдениям, то модель (2.2) может быть записана в матричном виде следующим образом:

$$\mathbf{y} = \mathbf{X}^T \beta + \varepsilon, \quad (2.3)$$

где  $\mathbf{y} = (y_1, \dots, y_N)^T$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^T$ , и матрица плана  $\mathbf{X}$  представляет собой матрицу, в которой по строкам записаны наблюдения  $x_i$ ,  $i = 1, \dots, n$ , а по столбцам — объясняющие переменные  $X_j$ ,  $j = 1, \dots, p$ :

$$\begin{aligned} \mathbf{X} &= \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \text{наблюдение}_1 \\ \text{наблюдение}_2 \\ \vdots \\ \text{наблюдение}_n \end{pmatrix} \\ &= (X_1, X_2, \dots, X_p) \end{aligned} \quad (2.4)$$

Чаще всего полагается, что  $x_{i1} = 1$ , тогда коэффициент  $\beta_1$  — это константа, или свободный член регрессионной модели.

В классической модели линейной регрессии, помимо функционального соотношения (2.2) (или (2.3)), накладываются дополнительные (и весьма жесткие) предположения о стохастической структуре модели:

$$\mathbf{E}\varepsilon_i = 0 \quad (2.5)$$

$$\mathbf{E}\varepsilon_i^2 = \sigma^2 \quad (2.6)$$

$$\mathbf{E}\varepsilon_i \varepsilon_j = 0 \quad \forall i \neq j \quad (2.7)$$

$$\text{rk } \mathbf{X} = p < n \quad (2.8)$$

$$X_j \quad \text{детерминированы.} \quad (2.9)$$

Часто бывает полезным предположение о явной форме ошибок:

$$\varepsilon_i \sim N(0, \sigma^2) \quad (2.10)$$



## 2.2.2 Метод наименьших квадратов

При подобных предположениях основным (и, как будет упомянуто ниже, наиболее качественным, в определенном смысле) способом оценки параметров модели  $\beta$  является метод наименьших квадратов:

$$\hat{\beta}_{\text{МНК}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - x_i^T \beta)^2 \quad (2.11)$$

Решением данной минимизационной задачи является *оценка наименьших квадратов* (англ. OLS, ordinary least squares), записываемая в матричном виде как

$$\hat{\beta}_{\text{МНК}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (2.12)$$

По результатам оценивания регрессионной модели можно построить *прогнозные значения* (fitted values)  $\hat{y}_i = x_i^T \hat{\beta}$  и *остатки* (residuals)  $e_i = y_i - \hat{y}_i$ ,  $i = 1, \dots, n$ .

### Stata

Команда пакета Stata, производящая оценку по методу наименьших квадратов, носит естественное название **regress**. После команды **regress** можно получить достаточно большое количество диагностических статистик (см. ниже), а также создать переменные, содержащие прогнозные значения, остатки и т. п., отдав команду **predict** “новая переменная”, опция, где опция — это вид статистики, которую надо построить: **predict** ... , **residuals** для получения остатков, **predict**, ... **xb** — для получения прогнозных значений  $\hat{y}$  и т. д. Более подробное описание возможностей команды **regress** и связанных с ней команд можно получить во встроенном мини-уроке **tutorial regress**.

Теоретическим обоснованием метода наименьших квадратов служит теорема Гаусса-Маркова:

**Теорема 2.1 (Гаусс, Марков)** *МНК-оценки являются несмещенными линейными оценками с минимальной дисперсией при выполнении условий (2.2)–(2.9), имеющими нормальное распределение при дополнительном предположении (2.10).*

Иными словами, в классе несмещенных линейных оценок МНК-оценки имеют наименьшую ковариационную матрицу<sup>2</sup>, которая равна

$$\text{Var } \hat{\beta}_{\text{МНК}} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (2.13)$$

---

<sup>2</sup> На множестве положительно определенных матриц отношение частичного порядка вводится следующим образом:  $A > B$ , если матрица  $(A - B)$  положительно определена.

Естественная оценка этой матрицы получается подставлением естественной оценки  $\sigma^2$ :

$$s^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2, \quad (2.14)$$

$$\widehat{\text{Var}} \hat{\beta}_{\text{МНК}} = s^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (2.15)$$

Несмещенность и эффективность (минимальная, в определенном смысле, точнее, в определенном классе оценок, дисперсия) — вполне приятные свойства, и именно поэтому МНК заслужил большую популярность в прикладной статистике. Заметим также, что МНК-оценки являются оценками максимального правдоподобия, если сделать дополнительное предположение о нормальности ошибок (2.10).

Прочие свойства оценок МНК, прогнозных значений и остатков можно найти в любой вводной книге по эконометрике.

### 2.2.3 Проверка гипотез

Почти всегда в прикладных исследованиях следующим шагом после оценивания регрессии является проверка тех или иных гипотез. Наиболее явно эта задача ставится при верификации теоретических моделей, хотя и в других задачах статистического анализа данных результаты проверки определенных гипотез могут служить дополнительным доводом в пользу рассматриваемой модели.

Наиболее часто проверяются линейные гипотезы относительно коэффициентов, т.е. гипотезы вида

$$H_0 : C\beta = r \quad \text{vs.} \quad H_a : C\beta \neq r, \quad (2.16)$$

где  $C$  — матрица  $q \times p$  полного ранга по строкам ( $\text{rk } C = q < p$ ), а  $r$  — вектор  $q \times 1$ . Иными словами, гипотеза  $H_0$  накладывает на коэффициенты  $q$  ограничений. Примером такой гипотезы может служить  $H_0 : \beta_2 = \dots = \beta_p = 0$ , или проверка того, что регрессионная модель в целом значима (т.е. описывает данные лучше, чем фраза “В среднем,  $y = \bar{y}$ ”). Для такой гипотезы  $C = I_{p-1}$ ,  $r = \mathbf{0}$ ,  $q = p - 1$ .

Статистикой для проверки гипотез такого вида является  $F$ -статистика:

$$F = \frac{(SSE_R - SSE_U)/q}{SSE_U/(n-p)} = \frac{(C\beta - r)^T (C(\mathbf{X}^T \mathbf{X})^{-1} C^T)^{-1} (C\beta - r)/q}{SSE_U/(n-p)}, \quad (2.17)$$

где  $SSE_R$  = sum of squared errors of the restricted model — сумма квадратов остатков модели с ограничениями (т.е. модели, оцененной при  $H_0$ ),  $SSE_U$  = sum of squared errors of the unrestricted model — сумма квадратов остатков в модели без ограничений. При нулевой гипотезе  $F$ -статистика имеет (центральное) распределение Фишера  $F(q, n - p)$ .

В частных случаях проверки гипотезы о значении одного из коэффициентов  $H_0 : \beta_k = \beta_k^{(0)}$  vs.  $H_a : \beta_k \neq \beta_k^{(0)}$  используется  $t$ -статистика<sup>3</sup>

$$t_{\beta_k} = \frac{\hat{\beta}_k - \beta_k^{(0)}}{\widehat{\text{Var}}(\hat{\beta}_k)^{1/2}} \sim t(n - p)|_{H_0}, \quad (2.18)$$

имеющая при  $H_0$  распределение Стьюдента с  $n - p$  степенями свободы, где оценка дисперсии  $\widehat{\text{Var}}(\hat{\beta}_k)$  — соответствующий диагональный элемент матрицы (2.15).

В классическом подходе к проверке гипотез, гипотеза  $H_0$  должна быть отвергнута, если  $F$ - или  $t$ -статистика превосходит соответствующий квантиль заранее зафиксированного критического уровня. Более современный вариант с использованием доверительных вероятностей предлагает считать статистической мерой достоверности получаемых результатов условную вероятность наблюдать такой же или худший исход при условии  $H_0$ . Например, если в качестве нулевой выступает гипотеза о независимости от определенного фактора (наиболее часто проверяемая гипотеза, которая обычно встраивается в результаты оценивания регрессии статистическими пакетами):

$$H_0 : \beta_k = 0 \quad \text{vs.} \quad H_a : \beta_k \neq 0, \quad (2.19)$$

то (эмпирическим) уровнем значимости (в англоязычной литературе — observed significance, или p-value) будет условная вероятность

$$\mathbf{P} \left[ |\hat{\beta}_k| > |\hat{\beta}_k \text{ наблюдаемое}| \mid H_0 \right]. \quad (2.20)$$

Большие значения (скажем, больше 10%) считаются свидетельством того, что не так уж маловероятно было бы наблюдать подобный исход, если бы данные действительно были порождены распределением, заданным нулевой гипотезой, и поэтому  $H_0$  не должна быть отвергнута. Напротив, значения ниже 1% говорят о том, что данные, скорее всего, несовместимы с нулевой гипотезой.

---

<sup>3</sup>  $t$ -статистика аналогична  $F$ -статистике в том смысле, что  $t^2(n - p) = F(1, n - p)$

**Stata**

Проверка линейных гипотез в пакете Stata выполняется командой `test`, отдаваемой после оценивания модели (командой `regress` или любой другой командой оценивания; см. раздел 3.9).

## 2.3 Нарушения предположений классической модели

Приведенная выше классическая модель достаточно проста и допускает достаточно простое решение (оценку параметров модели) по методу наименьших квадратов. Однако, в то же время, она достаточно хрупка по отношению к нарушениям базовых предположений, которые сводят на нет полезные свойства МНК-оценок, устанавливаемые теоремой Гаусса-Маркова.

Рассмотрим, к чему приводят нарушения отдельных условий теоремы 2.1.

### 2.3.1 Нецентральность

Условие (2.5), вообще говоря, не является существенным ограничением, если в число регрессоров входит (может входить) константа (столбец единиц в матричной записи). В этом случае смещение математического ожидания ошибки может быть поглощено свободным членом регрессионной модели.

### 2.3.2 Стохастичность регрессоров

Условие детерминированности регрессоров (2.9) существенно упрощает анализ и верно, вообще говоря, только в случае запланированных экспериментов, в которых исследователь полностью контролирует входные параметры (независимые переменные). В том случае, если регрессоры стохастические, т.е. являются случайными величинами, условия на моменты (2.5)–(2.7) заменяются условными математическими ожиданиями при условии  $x$ . При этом сама задача должна быть переформулирована в терминах случайной выборки, и необходимость в условии (2.7) отпадает по определению последней<sup>4</sup>. Необходимо так-

---

<sup>4</sup> Естественно, происхождение данных должно допускать подобную переформулировку. Классом задач, в которых такая переформулировка невозможна (или, во всяком случае, требует довольно за-

же переформулировать ранговое условие (2.8) в терминах невырожденного предела по вероятности для матрицы  $\mathbf{X}^T\mathbf{X}$ :

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^T \mathbf{X} = M > \mathbf{0}_{p \times p} \quad (2.21)$$

Наиболее вероятное дальнейшее нарушение предположений модели — коррелированность регрессоров и ошибки, когда

$$E[\varepsilon|x] \neq 0 \quad (2.22)$$

Основные эконометрические примеры, в которых ошибки и регрессоры могут быть коррелированы — это *модели с ошибками измерения* (measurement error models), рассматриваемые ниже в этом параграфе, и *одновременные уравнения* (simultaneous equations, см. параграф 2.6.2).

Можно показать, что в случае (2.22) МНК-оценки оказываются смещенными и несостоятельными (т. е. смещение не стремится к нулю в асимптотике). Чтобы избавиться от смещения, используется техника *инструментальных переменных* (англ. IV, instrumental variables): регрессоры проецируются в подпространство некоторых других переменных (инструментов), про которые известно, что они не коррелированы с ошибкой  $\varepsilon$ , но хорошо отражают регрессоры  $X$  (имеют с ними тесную корреляцию). Данная процедура является вариантом *двухшагового метода наименьших квадратов* (англ. 2SLS, two-stage least squares). IV-оценки являются несмещенными, однако по эффективности они существенно уступают МНК. Обобщенный метод моментов (generalized method of moments — GMM, ((Greene 1997), (Matyas 1999)), развивающий идеи оценки минимума  $\chi^2$  ((Neuman, Pearson 1928))) позволяет получить оценки, эффективные в классе IV-оценок, использующих данный фиксированный набор инструментов.

Выбор инструментов можно производить только из априорных предположений о том, какие переменные, *скорее всего*, некоррелированы с ошибкой, а какие — *неизбежно* коррелированы. Проверка на необходимость применения инструментальных переменных проводится с помощью теста Хаусмана ((Hausman 1978)). При нулевой гипотезе о некоррелированности ошибок и регрессоров и МНК-оценка, и IV-оценка являются метных усилий, является анализ временных рядов, для которого имеются свои собственные методы. См. Айвазян, Мхитарян (1998, гл. 16). Кроме того, условие независимости данных нарушается и для стратифицированных выборок, о которых будет рассказано ниже (см. раздел 2.6.1)

несмещенными, при этом первая эффективна, а вторая — нет, однако предел по вероятности их разности равен нулю. При альтернативе (ошибки и регрессоры коррелированы) МНК-оценка, в отличие от IV-оценки, несостоятельна, и предел по вероятности нулю не равен. Тогда при нулевой гипотезе квадратичная форма специального вида от разности оценок коэффициентов будет иметь (центральное) распределение  $\chi^2$  с числом степеней свободы, равным количеству сравниваемых коэффициентов / налагаемых линейных ограничений.

Тест Хаусмана является общим тестом на корректность спецификации модели. Так, он применяется для проверки корректности модели случайного эффекта против модели фиксированного эффекта для панельных данных.

**Stata** Команда пакета Stata, выполняющая регрессию с инструментальными переменными, называется `ivreg`. Тест Хаусмана выполняется командой `hausman`, для которой необходимо оценить менее эффективную, но заведомо состоятельную модель, сохранить результаты (`hausman, save`), затем оценить модель более эффективную, но несостоятельную при нарушении нулевой гипотезы, и оценить разницу коэффициентов (`hausman` без параметров).

Возможен другой вариант отказа от детерминированности регрессоров. Регрессоры сами по себе могут быть детерминированы, но измеряться с ошибкой, и тогда модель приобретает вид:

$$y_i = x_i^{*T} + \varepsilon_i \quad (2.23)$$

$$x_i = x_i^* + \delta_i \quad (2.24)$$

где измеряемыми величинами являются  $x_i$ , однако данные ( $y_i$ ) порождаются ненаблюдаемыми  $x_i^*$ . Это приводит к коррелированности регрессоров и ошибок, что вызывает смещение оценок. Как и в предыдущем случае, для получения несмещенных оценок используется метод инструментальных переменных, причем инструменты должны выбираться некоррелированными с ошибками  $\delta_i$ .

### 2.3.3 Гетероскедастичность остатков

Нарушение условий на вторые моменты (2.6) (*гомоскедастичность*) и (2.7) (*независимость*) приводит к тому, что МНК-оценки перестают быть эффективными в своем

классе. Еще хуже, однако, что “наивная” МНК-оценка ковариационной матрицы оценок коэффициентов оказывается смещенной и несостоятельной, из-за чего тесты на значения коэффициентов будут показывать неверный уровень значимости. Как правило, оценки дисперсии оценок коэффициентов занижаются, т.е. наивные оценки оказываются слишком “оптимистическими”.

Оказывается, что можно найти линейное преобразование переменных, сводящее задачу к МНК. Если ввести ковариационную матрицу ошибок регрессии

$$\Omega = \text{Var } \varepsilon \quad (2.25)$$

то можно построить оценки *обобщенного МНК* (англ. GLS, generalized least squares) следующего вида:

$$\hat{\beta}_{\text{ОМНК}} = (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Omega^{-1} \mathbf{y} \quad (2.26)$$

Аналогом теоремы Гаусса-Маркова в случае нарушений условий на вторые моменты является теорема Айткена.

**Теорема 2.2 (Айткен (Aitken))** *Если в классической модели линейной регрессии нарушены предположения (2.6)–(2.7), то оценка ОМНК является наиболее эффективной в классе линейных несмещенных оценок.*

При этом дисперсия этой оценки равна

$$\text{Var } \hat{\beta}_{\text{ОМНК}} = (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1}, \quad (2.27)$$

а дисперсия “наивной” оценки МНК —

$$\text{Var}(\hat{\beta}_{\text{МНК}}) = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \Omega^{-1} \mathbf{X}) (\mathbf{X}^T \Omega^{-1} \mathbf{X}) > (\mathbf{X}^T \Omega^{-1} \mathbf{X})^{-1} \quad (2.28)$$

Идентификация нарушения условий на вторые моменты ошибок не так уж тривиальна. Есть, однако, ряд задач, в которых эти условия можно считать априорно нарушенными. В первую очередь, это задачи анализа временных рядов, а также анализ стратифицированных и панельных обследований, о чем будет рассказано в разделах 2.3.4 и 2.6.1.

Что касается гетероскедастичности, при которой сохраняется независимость наблюдений (2.7) (но нарушается постоянство дисперсий ошибок (2.6)), то ее можно обнаружить, дополнительно сделав предположение об определенной функциональной форме

этой зависимости. Так, тест Гольдфельда-Куандта (Goldfeld-Quandt) предполагает зависимость дисперсии ошибок от одной из переменных, а тест Бройша-Пагана (Breusch-Pagan) — линейную зависимость дисперсии от некоторых дополнительных переменных.

**Stata**

В пакете Stata реализована следующая версия теста на гетероскедастичность (Кука-Вайсберга, Cook-Weisberg) которая вызывается командой `hettest`, отдаваемой после `regress`:

$$\begin{cases} \ln e_i^2 = z^T \gamma + \text{ошибка}_i \\ H_0 : \gamma = \mathbf{0} \end{cases}$$

где  $z$  может быть прогнозными значениями зависимой переменной или матрицей заданных переменных.

В общем случае гетероскедастичность без дополнительных предположений выявить, учесть и побороть невозможно: ковариационная матрица ошибок содержит  $\frac{N(N-1)}{2}$  неизвестных, оценить которые по  $N$  наблюдениям невозможно. Поэтому для оценивания ковариационной матрицы ошибок  $\Omega$  делаются разнообразные предположения о параметрической зависимости  $\Omega$  от некоторого малого числа параметров  $\theta$  известного вида:  $\Omega = \Omega(\theta)$ , где вектор параметров  $\theta$  должен быть (состоятельно) оценен по выборочным данным. В силу этого, оценивание с помощью *доступного обобщенного МНК* (feasible generalized least squares) состоит из (как минимум) двух этапов: состоятельного оценивания  $\theta$  (например, при помощи обычного МНК, являющегося состоятельным даже при нарушении условий на вторые моменты), а затем, с использованием состоятельной оценки  $\hat{\theta}$  (и, соответственно, состоятельной оценки  $\widehat{\Omega}(\hat{\theta})$ ), самой регрессионной модели. Для уточнения оценок процедуру “оценивание  $\theta \rightarrow$  оценивание регрессионной модели с ковариационной матрицей  $\Omega(\hat{\theta})$ ” можно повторять до достижения сходимости; при определенных условиях получаемые в пределе оценки будут эквивалентны оценкам МНК.

Альтернативный способ борьбы с гетероскедастичностью — оценивать ковариационную матрицу оценок коэффициентов из условий второго порядка минимума суммы квадратов остатков, пользуясь разложением Тейлора. Такие поправки известны в эконометрической практике как оценка ковариационной матрицы в форме Уайта (White):

$$\hat{V}(\hat{\beta}) = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n e_i^2 x_i x_i^T \right) \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \quad (2.29)$$



Вид этой оценки ковариационной матрицы оценок параметров провоцирует назвать ее “оценкой бутерброда” (sandwich estimator), и это название также встречается в статистической литературе. Встречается также название “оценка Хьюбера” (Huber), который независимо предложил эту оценку. В случае независимости наблюдений эта матрица является состоятельной оценкой искомой ковариационной матрицы; обобщения на случай зависимых данных в следующем разделе.

**Stata** В пакете Stata оценка этой матрицы вызывается не слишком, на мой взгляд, удачно названной опцией **robust** команды **regress**. Кроме того, в пакете Stata имеется возможность оценивания регрессии с весами (в данном случае, веса должны быть обратно пропорциональны стандартному отклонению для данного наблюдения) — **regress [weight=exp]**, где квадратные скобки для указания весов *обязательны*. Stata различает несколько типов весов (см. **help weights**); в данном случае необходимо указать **aweight** — аналитические веса. Наконец, есть специальная команда для оценивания с весами, учитывающими дисперсию отдельных наблюдений — **vwls**.

### 2.3.4 Автокоррелированность остатков

Вопрос об автокоррелированности остатков имеет смысл ставить тогда, когда данные упорядочены во времени (и отстоят друг от друга на равные промежутки). В этом случае можно применять средства анализа временных рядов.

**Stata** Пакет Stata версии 6 и выше имеет достаточно большое количество встроенных команд для анализа временных рядов (команды с префиксом **ts**), в т.ч. операторы лага (сдвига назад по оси времени на единицу) **L.**, разности **D.**, сглаживания сезонных колебаний **S.**. Общая справка по этим командам находится по ключевому слову **time**.

В контексте анализа временных рядов тестом на простейшую автокорреляцию (первого порядка) ошибок является тест Дарбина-Уотсона (Durbin-Watson), статистикой которого является

$$D = \frac{\sum_{i=2}^N (e_i e_{i-1})}{\sum_{i=1}^N e_i^2} \quad (2.30)$$

Если ошибки некоррелированы, статистика Дарбина-Уотсона должна принимать значения, близкие к 2. Значения, близкие к 0 или 4, должны служить тревожным сигналами.

лом. К сожалению, распределение этой статистики зависит от распределения ошибок, поэтому процентные точки для теста на автокоррелированность ошибок получаются исключительно вычислительным экспериментом. Таблицы критических значений статистики Дарбина-Уотсона приводятся в Айвазян, Мхитарян (1998). Для выявления лаговой структуры более высокого порядка необходимо по полной программе привлекать средства анализа временных рядов.

**Stata** В пакете Stata статистика Дарбина-Уотсона выводится командой `dwstat`, отдаваемой после `regress`.

Как и в случае с гетероскедастичностью, можно сформулировать поправки к матрице ковариации оценок коэффициентов, чтобы та была состоятельна при автокоррелированности остатков. Один из вариантов такой поправки был предложен Ньюи и Вестом ((Newey, West 1987)):

$$\hat{\text{Var}}(\hat{\beta}) = \sum_{l=-k}^k \left(1 - \frac{|l|}{k+1}\right) \left(\frac{1}{n} \sum_{i=1}^n x_{i-l} x_i^T\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n e_i e_{i-l} x_{i-l} x_i^T\right) \left(\frac{1}{n} \sum_{i=1}^n x_i x_{i-l}^T\right)^{-1}, \quad (2.31)$$

Напомним, что  $x_i$  обозначает столбец, соответствующий  $i$ -му наблюдению. Такая оценка ковариационной матрицы состоятельна при автокорреляции ошибок с числом лагов, не превышающим  $k$ . Убывающие веса при более отдаленных лагах использованы для того, чтобы гарантировать положительную определенность получаемой матрицы. При  $k = 0$  оценка Ньюи-Веста сводится к оценке Уайта (2.29).

**Stata** В пакете Stata регрессия с поправками к ковариационной матрице в форме Ньюи-Веста вызывается командой `newey`. Для того, чтобы корректно использовать временную структуру данных, необходимо предварительно отдать команду `tsset`, либо указать в опции `newey`, `t()`, какая переменная соответствует времени.

### 2.3.5 Мультиколлинеарность

Нарушение условия (2.8) носит название *мультиколлинеарность*, т.е. что-то вроде множественной совместной линейности. Точная коллинеарность означает, что регрессоры не являются линейно независимыми. В этом случае линейно зависимые коэффициенты

оценить невозможно, хотя можно оценить те линейные комбинации, которые друг от друга линейно не зависят.

Очевидно, на практике встретиться с точной мультиколлинеарностью вряд ли возможно <sup>5</sup> (за исключением досадных оплошностей типа включения в набор регрессоров всех 0/1-переменных, порождаемых одним и тем же фактором, например, индикаторов *и* мужского, *и* женского пола).

**Stata** К счастью (или к несчастью), Stata умеет обрабатывать подобные ситуации и выбрасывать, на свое усмотрение, переменные, которые она сочтет коллинеарными. К счастью — потому что процесс выполнения задания не будет прерван, а к несчастью — потому что контролировать, какие переменные будут выброшены, нельзя (а вообще-то исследователь должен был предусмотреть это на этапе выбора спецификации модели). Для корректной работы с категориальными переменными у пакета Stata есть собственное средство создания бинарных переменных — команда `xi`. Наконец, можно задать регрессию с “поглощением” одного качественного фактора — `areg`, где префикс `a` означает `absorb`, т.е. “поглотить”. Для поглощаемого фактора будет выведена F-статистика. Возможно, для моделей со сложными категориальными структурами удобнее использовать средства дисперсионного анализа — команду `anova` (см. также `help anova`, `tutorial anova`), позволяющую задавать количественные факторы с помощью опции `anova ... , continuous`.

Однако и неполная мультиколлинеарность способна доставить немало хлопот. Из-за близости матрицы  $\mathbf{X}^T\mathbf{X}$  к вырожденной дисперсии оценок коэффициентов убегают к бесконечности. Типичные признаки подобной ситуации — незначимость отдельных коэффициентов при значимости регрессии в целом, значительное изменение оценок коэффициентов (например, изменение знаков) при изменении состава регрессоров.

Мультиколлинеарность можно выявить и напрямую — например, визуально проанализировав матрицу выборочных корреляций, или, что более корректно в статистическом смысле, проведя анализ главных компонент.

**Stata** Анализ главных компонент является, в некотором смысле, частным случаем факторного анализа, поэтому соответствующая команда Stata носит название `factor ... , pc`, где опция `pc` показывает, что нас интересуют главные компоненты (`principal components`).

---

<sup>5</sup> Хотя именно такая постановка задач характерна для задач дисперсионного анализа.

На языке вычислительных методов линейной алгебры проблема мультиколлинеарности связана с понятием “плохая обусловленность”. Критерием плохой обусловленности является высокая величина отношения  $\lambda_{max}/\lambda_{min}$  — максимального и минимального собственных чисел матрицы  $\mathbf{X}^T\mathbf{X}$ , — называемого *показателем обусловленности* (condition number). Это соотношение также позволяет судить о степени серьезности проблем мультиколлинеарности: показатель обусловленности в пределах от 10 до 100 свидетельствует об умеренной коллинеарности, свыше 1000 (бывает и такое) — об очень серьезной коллинеарности.

Наиболее детальным показателем наличия проблем, связанных с мультиколлинеарностью, является *коэффициент увеличения дисперсии* (англ. variance inflation factor, VIF; см. Fox (1997), Smith and Young (2001)), определяемый для каждой переменной как

$$\text{VIF}(\beta_j) = \frac{1}{1 - R_j^2}, \quad (2.32)$$

где  $R_j^2$  — коэффициент множественной детерминации в регрессии  $X_j$  на прочие  $X$  (здесь  $X_j$  обозначает  $j$ -ю переменную, т.е.  $j$ -й столбец матрицы  $\mathbf{X}$ ). Этот коэффициент фигурирует в выражении для дисперсии выборочной оценки коэффициентов линейной регрессии:

$$\text{Var } \beta_j = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n - 1) \text{Var } X_j^2} \quad (2.33)$$

и показывает, во сколько раз дисперсия оценки больше “идеальной”, если бы мультиколлинеарности не было<sup>6</sup>. Поводом для беспокойства следует считать значения VIF от 4 и выше, что соответствует  $R_j^2 \simeq 0.75$ .

**Stata** Значения коэффициентов увеличения дисперсии выводятся командой `vif`, отдаваемой после `regress`.

Мультиколлинеарность возникает напрямую, если в регрессию включен набор 0/1-переменных, порождаемых одним качественным фактором с несколькими категория-

---

<sup>6</sup> Стандартная ошибка оценки, очевидно, увеличивается в  $\sqrt{\text{VIF}}$  раз. Эта величина имеет смысл диагностический, а не практический: нельзя делить на VIF для того, чтобы получить “правильную” дисперсию!

ми<sup>7</sup>: сумма таких бинарных переменных будет чаще всего давать единицу, если доля наблюдений, попадающих в базовую категорию, меньше  $1/2$ , и поэтому эти переменные в совокупности коллинеарны с константой. В реальных задачах при количестве объясняющих переменных более десяти, мультиколлинеарность возникает с очень большой вероятностью.

Наконец, если какая-либо переменная принимает такие значения, что ее стандартное отклонение много меньше, чем абсолютное значение среднего (например, среднее равно 70, а стандартное отклонение — 5, так что переменная в основном принимает значения от 60 до 80), то такая переменная будет также коллинеарна с константой. Другими словами, вариабельность переменной недостаточна, чтобы точно оценить соответствующий коэффициент: член  $\text{Var } X_j^2$  в выражении (2.33) мал, и поэтому дисперсия оценки коэффициента велика. В этом случае простым и естественным способом борьбы с высокой дисперсией оценки коэффициента будет отцентрировать соответствующую переменную, т.е. от переменной  $X_j$  перейти к переменной  $X_j^* = X_j - \bar{X}_j$ .

В более общем случае есть несколько способов ослабить эффекты мультиколлинеарности, но они, естественно, связаны с определенными потерями (по сравнению с хорошими свойствами МНК-оценок). Один из возможных путей — исключение некоторых из коллинеарных регрессоров (что означает невозможность оценить коэффициенты при выкидываемых регрессорах, т.е. определенную потерю информации; процедуры выбора переменных будут рассмотрены в параграфе 2.4.1) или переход к главным компонентам исходных переменных (что затрудняет интерпретацию получаемых коэффициентов, а также анализ значимости отдельных переменных).

Другой подход к решению проблемы мультиколлинеарности заключается в *смещенном* оценивании параметров. Идея этого подхода состоит в том, чтобы попытаться найти оценку, минимизирующую среднее квадратическое отклонение, или среднее квадратич-

---

<sup>7</sup> В свете этого заявления, которые делаются при пояснении результатов регрессии, вроде: “Наблюдается значимый эффект энергетической отрасли, а металлургия и химия незначимы”, выглядят несколько наивно. Во-первых, фактор “отрасль” имеет смысл рассматривать как единое целое. Во-вторых, оценки коэффициентов в конкретной регрессионной модели зависят от того, какая категория была выбрана в качестве базовой. В-третьих, из-за мультиколлинеарности  $t$ -статистики отдельных коэффициентов говорят не так уж много.

ческий риск оценки:

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{B}} \mathbf{E}(\hat{\beta} - \beta)^2 = (\text{смещение } \hat{\beta})^2 + \text{Var}(\hat{\beta}) \quad (2.34)$$

где класс оценок  $\mathcal{B}$  — более широкий, чем рассматриваемые обычно несмещенные линейные по  $\mathbf{y}$  оценки.

В рамках такого подхода матрицу  $\mathbf{X}^T \mathbf{X}$  можно *регуляризовать*, или сделать “более обратимой” путем добавления заведомо регулярной матрицы — например, вида  $\nu I_p$ , где  $I_p$  — единичная матрица размера  $p$ . Тогда оценка будет иметь вид:

$$\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \nu I_p)^{-1} \mathbf{X}^T \mathbf{y} \quad (2.35)$$

Эти оценки называются *ридж-оценками* (от англ. ridge — гребень; в русской литературе встречается также вариант “гребневая регрессия”. Происхождение этого термина, по всей видимости, связано с тем, что функция правдоподобия в случае мультиколлинеарности представляет собой не пик, а нечто вроде гребня; см. Демиденко (1981)). В английской литературе встречается также вариант shrinkage estimator, показывающий, что ридж-регрессия “стягивает” оценки коэффициентов к нулю. При этом с ростом  $\nu$  дисперсия оценок уменьшается, хотя увеличивается их смещение. Можно показать, что существует  $\nu$  такое, что среднеквадратическая ошибка из (2.34) смещенной оценки ниже, чем у несмещенной оценки МНК, т.е. можно подобрать  $\nu$  таким образом, чтобы достигнуть компромисса между смещением и дисперсией.

**Stata**

Ридж-регрессия реализована командой `rxridge`, имеющейся в официальных дополнениях к Stata, STB-28. Эта команда была изначально написана для весьма древней версии Stata, и у меня были проблемы с этой командой в 6-й версии Stata. Корректная версия находится на сайте компании, и ее можно найти командой `webseek rxridge`.

### 2.3.6 \* Проблема робастности

Наконец, одним из самых сложных случаев для анализа чувствительности оценок является нарушение предположения о том, что мы имеем дело с “хорошим” распределением ошибок (например, нормальным, как в (2.10)). Иными словами, как меняются результаты анализа, если стохастические компоненты (в случае регрессии — ошибки  $\varepsilon$ ) ведут себя не так, как нам бы хотелось их промоделировать?

Может оказаться, что отклонение от модельных допущений о стохастической природе ошибок меняет не только интерпретацию результатов, но и требует применения принципиально иной методологии анализа данных. Так, при сильной асимметричности распределений интерпретация обычной линейной регрессии затрудняется: среднее, в отличие от симметричных распределений, не является хорошим показателем того, где в основном лежат значения наблюдаемой величины. Асимметрия часто присуща данным, в которых наблюдения отличаются друг от друга масштабом — например, в финансовых данных по однородным предприятиям, характеризуемых размером — числом занятых, объемом производства, капиталом, и т.п. Весьма странные распределения имеют доли (например, доля аутсайдеров среди владельцев акций, или доля расходов на питание в бюджете домохозяйства) и отношения экономических величин вообще. Для анализа таких данных стоит использовать методы, свободные от распределения — такие, как знаковые и ранговые тесты Уилкоксона-Манна-Уитни на равенство медиан (`signrank` и `ranksum`) вместо  $t$ -теста на равенство средних.

Некоторые из вопросов такого рода находятся в ведении *робастной статистики* Хьюбер (1984), главной задачей которой является выяснение влияния отклонений формы распределений стохастических компонент от предполагаемой (заданной) на результаты статистического анализа и построение статистических процедур (оценок, тестов, критериев), которые как можно слабее зависели бы от предположений о распределениях. В этом жанре оценки параметров регрессионной модели рассматриваются как функционалы от распределений ошибок, и одной из характеристик робастности является кривая влияния (англ. *influence function* или *influence curve*) — производная этого функционала в заданной точке пространства регрессоров на заданном распределении. Значение этой производной определяет, насколько изменится значение оценки при изменении (возможно, бесконечном) наблюдаемого значения зависимой переменной при фиксированных значениях остальных наблюдаемых значений.

Точный анализ показывает, что оценка МНК не является робастной. На качественном уровне, при появлении в выборке выбросов, обусловленных тяжелыми хвостами распределений ошибок, метод наименьших квадратов стремится провести поверхность отклика через крайние точки, а не через основную массу точек. Это и не удивительно, учитывая линейность МНК-оценок по  $y$ : если в каком-то  $i$ -м наблюдении  $y_i \rightarrow \infty$ , то и  $\hat{\beta}_{\text{МНК}} \rightarrow \infty$ .

Более удачными, с точки зрения робастности, являются М-оценки, получаемые как решения экстремальной задачи

$$\sum_{i=1}^N \rho(z_i; \beta) \rightarrow \min_{\beta}, \quad (2.36)$$

где функция  $\rho(\cdot)$  асимптотически растет по первому аргументу медленнее, чем  $z^2$  и тем самым придает меньшие веса далеко отстоящим наблюдениям<sup>8</sup>. Примером функции, обеспечивающей робастность оценок, является  $\rho(z, \beta) = |z|$ . Получаемая при этом регрессия называется *медианной*, поскольку получаемая линия соответствует условной медиане.

Еще одна часто используемая спецификация — функция Хьюбера (Huber)

$$\rho_c^{Huber}(z) = \begin{cases} z^2/2, & |z| < c \\ c|z| - c^2/2, & |z| \geq c \end{cases} \quad (2.37)$$

Параметр  $c > 0$  играет роль настроечного параметра, отвечающего за робастность: если  $c \rightarrow \infty$ , то мы получаем метод наименьших квадратов; если, напротив,  $c \rightarrow 0$ , то мы получаем робастную медианную регрессию.

Другая спецификация функции  $\rho(\cdot)$ , которая практически игнорирует слишком далекие выбросы — бивесовая функция Тьюки (Tukey):

$$\rho_c^{biweight}(z) = \begin{cases} \frac{c^2}{6} \left[ 1 - \left( 1 - \left( \frac{z}{c} \right)^2 \right)^3 \right], & |z| < c \\ \frac{c^2}{6}, & |z| \geq c \end{cases} \quad (2.38)$$

Здесь  $c$  — также параметр робастности. При  $c \rightarrow \infty$  бивесовая функция вырождается в обычную параболу метода наименьших квадратов.

**Stata**

Похожий алгоритм реализован в команде `rreg` — робастная регрессия — в пакете Stata. В нем на начальных стадиях алгоритма используется функция Хьюбера, а затем — функция Тьюки.

Естественно, что, приобретая робастность оценки, мы должны где-то потерять. Обычно компромисс происходит за счет эффективности: если ошибки действительно имеют нормальное распределение, то робастные оценки теряют в эффективности

<sup>8</sup>  $z$  соответствует стохастическим компонентам, т.е. остаткам регрессии:  $z = y - x^T \beta$ .



5–10% при  $H_0 : \varepsilon_i \sim N(0, \sigma^2)$ . Эти оценки, впрочем, превосходят по эффективности МНК даже при долях загрязнения тяжелыми хвостами на уровне малых процентов.

Тема идентификации выбросов, связанная с проблемами робастности, будет еще раз поднята в разделе 2.4.3.

### 2.3.7 Преобразование к нормальности и линейности

Иногда отклонение от нормальности можно компенсировать за счет преобразования зависимых и/или объясняющих переменных. Наиболее популярным классом преобразований является однопараметрическое *преобразование Бокса-Кокса* (Box-Cox):

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases} \quad (2.39)$$

где  $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$  — среднее геометрическое  $y_i$ . Оценка необходимой степени преобразования  $\lambda$  можно произвести методом максимального правдоподобия<sup>9</sup>. Оказывается, что преобразование Бокса-Кокса не только позволяет прийти к нормальности, но и, в ряде случаев, стабилизировать дисперсию ошибок, а также избавиться от нелинейности (см. также раздел 2.4.2)

Самым типичным случаем является логарифмическое преобразование, применяемое тогда, когда ошибки имеют *мультипликативный* характер (приводящий к логарифмически нормальному распределению), а не аддитивный (приводящий к обычному нормальному распределению). Очень многие экономические данные имеют распределение, близкое к логнормальному (доходы населения, объем производства, занятость, капитал промышленных предприятий, параметры бюджетов разных стран или регионов, и т. п.). Еще одним аргументом в пользу логарифмирования в экономических задачах можно считать то, что логарифмическое преобразование производственной функции Кобба-Дугласа приводит ее к линейному виду.

Следует, впрочем, иметь в виду, что при использовании преобразования Бокса-Кокса (как и любого другого преобразования) могут возникнуть сложности с интерпретацией регрессионной модели, ее ошибок или коэффициентов. В случае с логарифмическим

---

<sup>9</sup> Нормировка на  $\dot{y}$  делается именно для того, чтобы получать корректные отношения правдоподобия.

преобразованием коэффициенты имеют вполне понятную экономисту интерпретацию эластичностей зависимой переменной по объясняющей.

**Stata**

Преобразование Бокса-Кокса выполняется командой `boxcox`. Опция `boxcox ... , graph` позволяет вывести график итераций процедуры максимального правдоподобия. Преобразованные значения можно получить командой `predict ... , tyhat` или опцией `boxcox ... , generate`. Задав, помимо преобразуемой переменной, список регрессоров, можно получить оценку регрессии

$$y^{(\lambda)} = \mathbf{X}^T \beta + \text{ошибки}, \quad (2.40)$$

результаты которой можно востребовать командой `regress` без параметров. Более мощный вариант преобразования Бокса-Кокса дается командой `boxcox2`, доступной в официальном дополнении STB-54.

## 2.4 Прочие отклонения от модели

Помимо отклонений от допущений (2.5)–(2.9), в реальной жизни нарушается и условие (2.2) на сам вид модели, что также необходимо уметь диагностировать и исправлять.

### 2.4.1 Спецификация модели: выбор нужных переменных

В регрессию, анализируемую исследователем, могут быть как включены переменные, не связанные с зависимой, так и пропущены переменные, существенные для ее объяснения. В первом случае точность оценивания, вообще говоря, снижается: оценки “зашумляются”, хотя и остаются несмещенными. Кроме того, включение дополнительных переменных несет риск возникновения или усиления мультиколлинеарности, что также сопряжено с увеличением дисперсии. Во втором случае оценки коэффициентов могут быть смещенными, а в силу недостаточной точности модели остатки будут слишком велики (т. е. оценка дисперсии ошибок будет смещена вверх).

К сожалению, однозначных рецептов выбора переменных, которые надо оставить в регрессии, не существует. В силу вышесказанного предпочтительнее изначально включать в регрессию как можно больше переменных (увеличение дисперсии все-таки не так плохо, как смещение оценок).

Если же необходимо, из тех или иных соображений, ограничить размерность модели, то обычно используемые процедуры включают в себя методы пошагового отбора или удаления переменных, основанные на тестах отношения правдоподобия или информационных критериях, в которых одни члены учитывают точность приближения, а другие штрафуют за излишне большое число подгоночных параметров.

**Stata**

Решение задачи выбора регрессоров в пакете Stata выполняется метакомандой `sw` (англ. *stepwise*). Полный синтаксис процедуры выбора регрессоров в линейной модели будет иметь вид `sw regress depvar varlist, опции`, где *опции* описывают параметры включения в модель и исключения из нее объясняющих переменных из списка `varlist`. Критерием, на основе которого делается решение о включении или исключении переменной из списка регрессоров, является статистика отношения правдоподобия.

Популярной мерой, характеризующей качество приближения модели (*goodness of fit*), является доля объясненной дисперсии  $R^2$ : чем выше, т.е. ближе к 1, статистика  $R^2$ , тем лучше. Эта статистика настолько популярна, что для целого ряда моделей были придуманы квази- $R^2$ , принимающие значение 0, если модель не имеет никакой объясняющей силы, и 1, если данные объяснены полностью. Следует, однако иметь в виду, что:

- статистика  $R^2$  возрастает с добавлением новых регрессоров, а при количестве регрессоров, равному количеству наблюдений, гарантированно достигает единицы (что, однако, не означает, что данные хорошо и полностью описаны: дисперсия прогнозных значений будет равна бесконечности).
- статистика  $R^2$  не робастна: при наличии выбросов  $R^2 \rightarrow 1$ .
- квази- $R^2$  могут в действительности иметь максимальное значение намного меньше 1, и в силу этого их ценность, мягко говоря, невелика.
- статистика  $R^2$  характеризует только прогностические возможности модели (*goodness of fit*). Анализ причинных связей — задача гораздо более тяжелая и требующая применения весьма мощных вероятностных концепций (причинность по Грэнжеру, *Granger causality test* ((Handbook 1983, 1984, 1986, 1994)).

Модификацией  $R^2$ , учитывающей первый из указанных эффектов, является статистика  $R_{adj}^2$ , в которой более тонко учитывается число степеней свободы модели:

$$R_{adj}^2 = 1 - \frac{\mathbf{e}^T \mathbf{e} / n - p}{\mathbf{y}^T \mathbf{y} / n - 1}, \quad (2.41)$$

где  $\mathbf{e}$  — вектор регрессионных остатков, а  $\mathbf{y}$  — (центрированный) вектор значений зависимой переменной.

Более удачны, в статистическом смысле, *информационные критерии*, соотносящие информацию, предоставляемую моделью, и информацию, имеющуюся в данных. Их идея состоит в том, что “качество модели” достигается как баланс качества приближения к реальным данным и статистической сложности модели, связанной со слишком большим числом параметров (overparametrization), поэтому статистика критерия состоит из штрафа за недостаточную подгонку и штрафа за излишнее число параметров<sup>10</sup>. Исторически первым, а потому наиболее популярным информационным критерием является *критерий Акайке* (AIC, Akaike information criteria):

$$\text{AIC} = -2 \ln L(\hat{\theta}) + 2p, \quad (2.42)$$

где  $L(\hat{\theta})$  — значение функции правдоподобия (ее логарифм сводится к остаточной сумме квадратов в нормальном случае), а  $p$  — количество регрессоров. “Оптимальная” в смысле данного критерия регрессия будет доставлять минимум критерию AIC. Другой вариант, байесовский критерий Шварца (Schwarz Bayesian information criterion, SBIC, BIC), использует в качестве штрафа за параметры  $p \ln n$ , где  $n$  — число наблюдений:

$$\text{SBIC} = -2 \ln L(\hat{\theta}) + p \ln n, \quad (2.43)$$

Поскольку критерий Шварца сильнее штрафует за лишние параметры, он выбирает модели меньшей размерности.

**Stata**

К сожалению, в пакете Stata нет встроенных команд, посвященных информационным критериям. Есть, однако, программа `fittest`, находящаяся в архиве SSC-IDEAS

<sup>10</sup> Формально, информационные критерии являются более точными оценками ожидаемой информации модели, или математического ожидания функции правдоподобия, чем само максимальное значение функции правдоподобия, полученное в ходе оценивания по методу максимального правдоподобия. Оценки максимального правдоподобия оказываются ближе к данным, чем к истинной модели. См., напр., Konishi and Kitagawa (1996).

(<http://ideas.uqam.ca>), которая выдает также значения  $R^2, R_{adj}^2$ , информационных критериев Акайке и Шварца, а также ряд статистик, относящихся в основном к логистическим регрессиям. Другая программа, вычисляющая критерии Акайке, Шварца, а также критерий информационной сложности Боздогана, находится на web-страничке автора и называется `icomp`<sup>11</sup>.

## 2.4.2 Нелинейность

Другим возможным нарушением классической модели регрессии может быть случай, когда функция регрессии  $E[y|x]$  нелинейна. Игнорирование нелинейности может представлять определенную проблему, поскольку неучтенная нелинейность отзовется изменением свойств остатков. Они оказываются смещенными, у них возникает корреляционная структура, а значит, смещаются и ковариационные матрицы оценок коэффициентов и, в конечном итоге,  $t$ - и  $F$ -статистики. Эта проблема может быть сформулирована в терминах пропущенных переменных (можно считать, что в регрессии пропущены необходимые нелинейные члены), и один из вариантов теста на неучтенную нелинейность был предложен в 1960-х гг. Рамсеем. В этом тесте рассматривается полиномиальная регрессия вида

$$e_i = \sum_{k=1}^K \gamma_k \hat{y}_i^k + \text{ошибка}_i, \quad (2.44)$$

где  $\hat{y}_i$  — прогнозные значения из обычной линейной МНК-регрессии, а  $e_i$  — ее остатки, и проверяется гипотеза  $H_0 : \gamma = \mathbf{0}$ .

**Stata** Тест Рамсея осуществляется в пакете Stata командой `ovtest`. Stata использует первые четыре степени ( $K = 4$ ) регрессоров или предсказанных значений независимой переменной.

---

<sup>11</sup> К сожалению, эти программы дают разные результаты; могу только сказать в свое оправдание, что я пользовался именно приведенными выше формулами, которые, в свою очередь, выведены из первых принципов. В статистической и эконометрической литературе гуляют и другие определения индексов AIC и SBIC — например, через остаточные суммы квадратов, к которым эти критерии сводятся в нормальном случае при неизвестной дисперсии. Вследствие этого нет однозначности и в публикуемых статьях, в которых авторы выбирают с помощью информационных критериев ту или иную модель. Опасайтесь подделок!

Нелинейность может заключаться в том, что функция регрессии связана с известными нелинейными функциями регрессоров (например, в моделях вида  $y = a + bx^2 + \varepsilon$ ,  $y = a \sin x + \varepsilon$ ,  $y = ax^b e^\varepsilon$ , где  $\varepsilon$  — “хорошие” (центрированные, независимые, с конечной дисперсией) ошибки. В подобных случаях преобразование переменных задачу можно свести к классической модели линейной регрессии, где линейность понимается как линейность относительно *параметров*.

В более серьезных случаях нелинейность является существенной, т.е. не сводимой к линейной модели. Функция регрессии имеет общий вид

$$y_i = f(x_i, \beta) + \varepsilon_i, \quad (2.45)$$

где  $f(\cdot)$  — известная функция достаточно общего вида ( $y = a \sin(bx + c) + \varepsilon$ ,  $y = ax^b + \varepsilon$  — чем отличаются эти функции от приведенных выше?). Оказывается, что *нелинейный метод наименьших квадратов* (англ. NLS, non-linear least squares) обеспечивает наиболее эффективные, в определенном классе максимизационных задач, оценки искомых параметров.

**Stata**

Пакет Stata позволяет оценивать и такие нелинейные регрессии с помощью команды `nl`. Чтобы воспользоваться этой командой, необходимо написать небольшую программу с достаточно жестко зафиксированным синтаксисом, которая будет вычислять значение функции регрессии  $f(\cdot)$  и передавать на оптимизацию `nl`.

### 2.4.3 Идентификация резко выделяющихся наблюдений

В связи с тем, что МНК-оценки неробастны, возникает естественный вопрос: не получится ли так, что малое число выделяющихся наблюдений будет задавать такую поверхность регрессии, которая будет иметь мало общего с поверхностью, проходящей через большинство точек? Например, в случае парной регрессии — может ли случиться, что прямая регрессии пройдет через одну точку и центр масс остальных? Увы, ответ положительный: наличие выделяющихся наблюдений (influential observations), или выбросов (outliers) — явление скорее типичное, нежели редкое, в прикладном анализе. Иногда это связано с тем, что отдельные наблюдения действительно сильно отличаются от остальных (например, Москва практически всегда выделяется при анализе данных по регионам России), а иногда может быть вызвано ошибкой во вводе данных — непра-

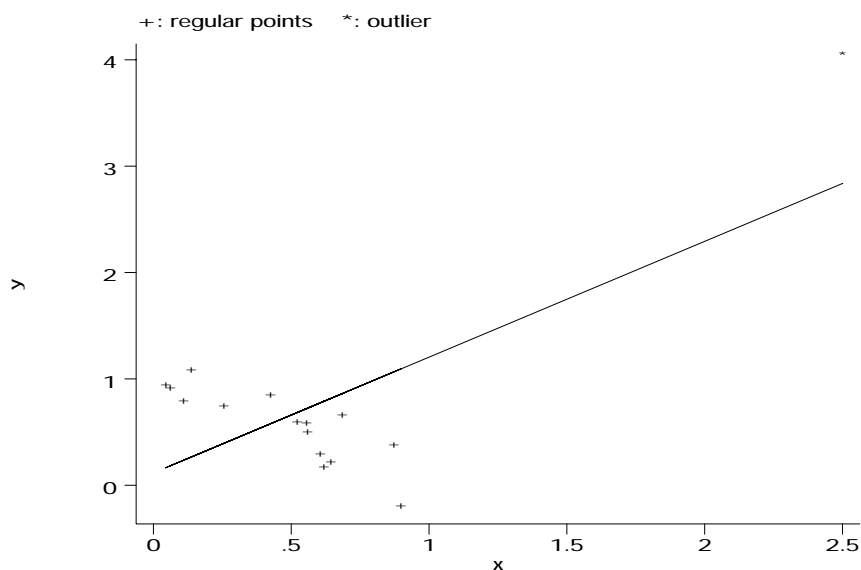


Рис. 2.1: Линия регрессии оттянута на себя выбросом. Истинная линия регрессии:  $y = 1 - x + \varepsilon$

вильно поставленная десятичная запятая, пропуск цифры при вводе данных или запись величины в миллионах рублей вместо тысяч (в результате деноминации 1997 г.), и т. п. Наконец, далеко отстоящие (в терминах стандартных отклонений) от основной массы данных точки могут появляться в асимметричных распределениях (логнормальное, гамма) или в распределениях с тяжелыми хвостами (распределение Стьюдента).

Чрезмерно высокое влияние отдельных наблюдений может быть связано с тем, что данное наблюдение отстоит далеко от остальных наблюдений в пространстве регрессоров (и, соответственно, обладает большим *плечом* (англ. leverage) в воздействии на данные), а может быть связано с большой ошибкой  $\varepsilon_i$  в данном наблюдении. Может быть, что оба фактора накладываются друг на друга, что может как усугубить (рис. 2.4.3), так и облегчить ситуацию.

Выявлять выделяющиеся наблюдения можно следующим образом<sup>12</sup>. Рассмотрим

<sup>12</sup> Данная тема, пожалуй, наименее типична для стандартных курсов по эконометрике, хотя стати-

прогнозные значения зависимой переменной:

$$\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \equiv Hy \quad (2.46)$$

Элементы матрицы  $H$  несут информацию о конфигурации точек в пространстве регрессоров  $\mathbf{X}$  и в то же время непосредственно задают влияние каждой точки  $y_i$  на все прогнозные значения  $\hat{y}$ . Можно показать, что  $h_{ii} = \sum_{j=1}^n h_{ij}^2$ , и поэтому мерой влияния  $i$ -точки можно положить  $h_i \equiv h_{ii}$  (англ. hat value, имеет смысл условной корреляции наблюдаемого и прогнозного значений при фиксированной остальной выборке). Далее,  $1/n \leq h_i \leq 1$ , причем среднее значение равняется  $p/n$ , и поэтому потенциально выделяющиеся наблюдения можно идентифицировать по высокому значению  $h_i$  — например, больше  $3p/n$ .

**Stata** hat-values можно получить командой `predict ... , hat`, отдаваемой после команды `regress`.

Помимо идентификации “опасных” точек в пространстве регрессоров, влияние на оценки МНК будут оказывать, как упоминалось выше, большие ошибки. Остатки регрессии как таковые, по всей видимости, не обязательно будут достаточно информативны, поскольку в совокупности они не являются независимыми, и, более того, МНК стремится провести поверхность регрессии как можно ближе к далеко отстоящим данным. Для получения независимых остатков необходимо исключить данное  $i$ -е наблюдение, прогнать регрессию заново и получить *стьюдентизированные остатки*<sup>13</sup>:

$$e_i^* = \frac{e_i}{s_e^{(i)} \sqrt{1 - h_i}}, \quad (2.47)$$

где  $s_e^{(i)}$  — оценка стандартного отклонения остатков при исключении  $i$ -го наблюдения, а появление коэффициента  $\sqrt{1 - h_i}$  связано с тем, что  $\text{Var } e_i | H_0 = (1 - h_i)\sigma^2$ . При нулевой гипотезе нормального распределения ошибок величина  $e_i^*$  имеет распределение Стьюдента с  $N - p - 1$  степенями свободы. Полностью аналогичной величиной будет стикам она известна не первый и даже не второй десяток лет. Любопытный читатель может найти развитие темы в Draper, Smith (1998), Fox (1997), Smith and Young (2001).

<sup>13</sup> Называемые также остатками по методу складного ножа, jack-knife, называемого также методом расщепления выборки. Его идея как раз и заключается в исключении отдельных наблюдений, оценивания статистической модели с исключенным наблюдением и сопоставления полученных оценок с оценками, полученными по полной выборке Эфрон (1988).



$t$ -статистика для коэффициента  $\gamma$  в регрессии  $y = \mathbf{X}^T \beta + \gamma D_i + \varepsilon_i$ , где  $D_i$  — бинарная переменная, равная единице в  $i$ -й точке и нулю в остальных.

Сочетание “большого плеча” и большого остатка выявляется при помощи  $D$ -статистики Кука (англ. Cook’s distance):

$$D_i = \frac{e_i^2}{p} \frac{h_i}{1 - h_i} \quad (2.48)$$

Самые высокие значения  $D$ -статистики свидетельствуют о том, что данное наблюдение достаточно заметно изменяет МНК-оценки коэффициентов. Эмпирическое значение порога “тревожности” —  $D_i > \frac{4}{N-p}$ .

Непосредственное влияние отдельных наблюдений на оценку коэффициента  $\hat{\beta}_k$  дается статистикой  $DFBETA_{k,i}$ :

$$DFBETA_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_k^{(i)}}{(\widehat{\text{Var}}\beta_k^{(i)})^{1/2}}, \quad (2.49)$$

где верхний индекс  $(i)$  показывает, что из расчетов исключено  $i$ -е наблюдение. Иными словами, мы получаем оценки коэффициентов и оценку их ковариационной матрицы по методу складного ножа и строим что-то вроде  $t$ -статистики, показывающей отклонение коэффициента при исключении данного наблюдения. В соответствии с этой интерпретацией, следует обращать внимание на наблюдения с  $|DFBETA_{k,i}| > 2/\sqrt{n-p}$ .

Еще одна статистика диагностики влияния наблюдений показывает, насколько сильно данное наблюдение оттягивает на себя линию регрессии:

$$DFFITS_i = e_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \quad (2.50)$$

Здесь  $h_{ii}$  в числителе учитывает, насколько далеко данная точка отстоит от основного массива, а  $1 - h_{ii}$  дает поправку на дисперсию остатков. Как и расстояние Кука, эта статистика учитывает и величину остатка, и его плечо в воздействии на линию регрессии. Если абсолютная величина статистики  $DFFITS_i$  в  $i$ -м наблюдении свыше  $2\sqrt{p/n}$ , то, возможно, это наблюдение заметно смещает всю линию регрессии.

#### Stata

Стюдентизированные остатки можно получить командой `predict ... , rstudent` после команды `regress`.  $D$ -статистика Кука вычисляется командой `predict ... , cooksd`, статистики  $DFBETA$  — `predict ... , dfbeta(имя переменной)` или отдельной командой `dfbeta`, статистики  $DFFITS$  — командой `predict ... , dfits`.

## 2.4.4 Визуальный анализ

Визуальный анализ часто является хорошим подспорьем в диагностике регрессий не очень больших размерностей и зачастую может помочь выявить большинство упомянутых выше нарушений классических предположений. Перечислим основные виды графиков, которые можно использовать для анализа “адекватности” регрессии.

**Stata** Практически вся графика Stata является вариантами команды `graph`, у которой имеется добрая сотня разнообразных опций на разнообразные случаи жизни. Наиболее часто используемые графики реализованы в виде отдельных команд. См. раздел 3.14.

- Перед началом анализа, еще до стадии оценивания регрессии, можно проанализировать распределение зависимой и независимых переменных. Сильная асимметрия может свидетельствовать о необходимости применения преобразований к нормальности, многомодальность — о наличии структуры групп наблюдений (которую можно учесть, введя бинарные переменные), и т. д.

**Stata** Общая сводка описательных статистик по одной или нескольким переменным выводится командой `summarize`. Графическое представление распределения отдельной переменной, т. е. гистограмму, можно получить командой `graph “имя переменной”`. Более продвинутые варианты анализа включают в себя использование ядерных оценок плотности (`kdensity`), нормальной бумаги (`qnorm`), а также прочие диагностические графики (описание которых можно найти по ключевому слову `diagplots`) и более совершенные средства создания гистограмм (программа `histplot`, загружаемая с архива программных компонентов SSE-IDEAS, находящегося в Бостонском Колледже: <http://ideas.uqam.ca>). Наконец, относительно простым тестом на нормальность является тест по третьему и четвертому моментам (которые, при соответствующей нормировке, равны нулю у нормального распределения, и совместное выборочное распределение которых является нормальным) — `sktest`, от англ. skewness-kurthosis test.

- Аналогичную процедуру можно выполнить в отношении регрессионных остатков ... <sup>14</sup>

---

<sup>14</sup> Следует, впрочем, иметь в виду, что большие *ошибки* (приводящие к регрессионным выбросам)

Stata

... которые можно получить командой `predict ... , residuals` после `regress`.

- Связь отдельных регрессоров с зависимой переменной можно проследить на диаграммах рассеяния. При помощи этих графиков уже можно выявить определенные недостатки регрессии. Так, если на диаграмме рассеяния большая часть данных группируется возле нуля, и есть несколько точек в оставшемся поле, то, скорее всего, данные необходимо трансформировать, чтобы снизить влияние удаленных точек.

Пример диаграммы рассеяния двух асимметричных распределений приводится на рис. 2.4.4.

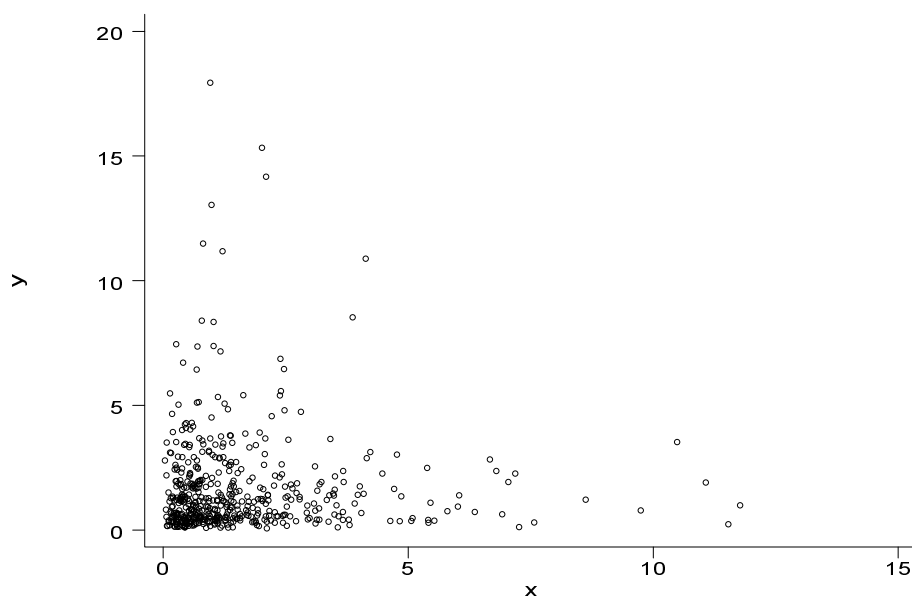


Рис. 2.2: Частные распределения обеих переменных асимметричны; график заполнен в основном около нуля и возле осей; необходимо преобразование к нормальности?

Более содержательным, в регрессионном контексте, графиком будет (частная) диаграмма рассеяния, очищенная от линейного вклада остальных переменных, не обязательно приводят к большим *остаткам*. Кроме того, остатки в совокупности не являются независимыми (так, их сумма равна нулю).

т. е. диаграмма рассеяния остатков регрессий

$$\mathbf{y} = \mathbf{X}^{(-k)T} \boldsymbol{\beta}^{(-k)} + \boldsymbol{\varepsilon}^{(-k)} \quad (2.51)$$

и

$$X_k = \mathbf{X}^{(-k)T} \boldsymbol{\gamma}^{(-k)} + \boldsymbol{\delta}^{(-k)}, \quad (2.52)$$

где верхний индекс  $(-k)$  означает отсутствие в составе регрессоров  $k$ -й переменной. Такой график называется *графиком добавленной переменной* (англ. added variable plot) или *графиком частной регрессии* (англ. partial regression plot). С его помощью можно выявлять гетероскедастичность (вида роста дисперсии ошибок с ростом какой-либо из переменных), нелинейность, а также находить возможные выбросы.

**Stata** График частной регрессии выводится командой `avplot`. К этой команде, как и к другим командам диагностики, выводящим двумерные графики, приложимы большинство опций диаграмм рассеяния.

- Общую скрытую нелинейность и/или гетероскедастичность можно обнаружить и на графике остатков в зависимости от прогнозных значений (т. е. по горизонтальной оси откладываются  $\hat{y}$ , а по вертикальной —  $e$ ). По построению, эти переменные некоррелированы, поэтому в общем и целом график должен лежать вокруг оси абсцисс.

**Stata** Соответствующая команда носит название `rvfplot` — англ. residual versus fitted. Аналитическими дополнениями являются диагностические тесты `hettest` и `ovtest`.

- Альтернативой графику частной регрессии (в особенности для диагностики нелинейности) может быть график *частных остатков*:

$$e^{(k)} = e + \beta_k X_k \quad (2.53)$$

**Stata** Соответствующие команды Stata — `cprplot` и `acprplot` (англ. component plus residual).

Возможно, какие-то из этих графиков можно включать в публикуемые материалы исследования — как свидетельство основательного анализа данных и адекватности статистических результатов.

### 2.4.5 Множественная проверка гипотез

Одним из простейших случаев проверки нескольких гипотез одновременно является  $F$ -тест на несколько линейных ограничений на параметры вида (2.16). Более тонким случаем является проверка гипотезы о значении (знаке) одного и того же коэффициента в нескольких регрессиях. Тонкостью, обычно игнорируемой, однако чрезвычайно важной, является корректная интерпретация получаемого совокупного уровня значимости. Действительно, если событие  $A_k$  состоит в том, что в  $k$ -й регрессии нулевая гипотеза не отвергнута (и, соответственно,  $\bar{A}_k$  — что отвергнута), то, очевидно,

$$P(\cup_k \bar{A}_k) \leq \sum_k P(\bar{A}_k) \quad (2.54)$$

а следовательно,

$$P(\cap_k A_k) \geq 1 - \sum_k P(\bar{A}_k) \quad (2.55)$$

В левой части (2.55) фигурирует вероятность принять нулевую гипотезу *во всех* регрессиях. Соответственно, если требуется, чтобы *совокупный* уровень значимости составлял  $\alpha$ , то самым простым способом гарантировать этот уровень значимости будет потребовать, чтобы правая часть (2.54) превосходила  $1 - \alpha$ . В свою очередь, простейший способ добиться этого — потребовать, чтобы уровень значимости в каждом из тестов  $P(\bar{A}_k)$  не превосходил  $\alpha/K$ , где  $K$  — общее количество тестов. Описанная выше процедура называется *процедурой Бонферрони* (Bonferroni adjustment) и является одним из примеров поправок на проверку множественных гипотез. Другие известные процедуры, зачастую более точные и менее консервативные — процедуры Шеффе (Sheffé), Тьюки (Tukey) и Воркинга-Хотеллинга (Working-Hotelling) ((Шеффе 1980), (Smith and Young 2001)).

Поправка на множественность — процедура методологическая, поэтому явно выраженной команды Stata для нее нет. Если исследователь собирается применять процедуру Бонферрони и ему заранее известно количество моделей, которые он будет оценивать, то можно задать уровень значимости для построения доверительных интервалов после оценивания моделей командой `set level ...`. По умолчанию устанавливается уровень значимости 95 (процентов). Текущее состояние можно выяснить командой `query` — см. раздел 3.15.

## 2.4.6 Данные с пропусками

Данные с пропусками — это проклятие исследований, в которых используются результаты выборочных обследований: зачастую, увы, невозможно гарантировать, что все респонденты дадут полную и точную информацию. Эта тема привлекла и привлекает значительное внимание в общественных науках, однако в эконометрике, как ни странно, эта тема известна только в рамках довольно узких моделей тобит-регрессии и выборочного отбора (sample selection — модель Хекмана). Данный раздел в значительной мере следует Little and Rubin (1987).

### Терминология

Возможность использования методов анализа разной степени сложности связана с тем, насколько простым или сложным является механизм, согласно которому данные оказываются пропущенными. Полезная терминология была введена в Rubin (1976). Говорится, что пропуски в данных *полностью случайны* (data are missing completely at random — MCAR), если  $P(X_j \text{ пропущено} | \text{прочие } X)$  не зависит ни от  $X_j$ , ни от прочих  $X$  (то есть эта вероятность постоянна для всех наблюдений, и наблюдаемые  $X_j$  являются случайной подвыборкой тех  $X_j$ , которые должны были получиться в эксперименте). Пропуски в данных *случайны* (missing at random — MAR), если  $P(X_j \text{ пропущено} | \text{прочие } X)$  не зависит от  $X_j$  (но могут зависеть от других  $X$ ). Оказывается, что в этих случаях механизм пропусков *несущественен* (ignorable), и к данным применимы вариации метода максимального правдоподобия. Наконец, если  $P(X_j \text{ пропущено} | \text{прочие } X)$  зависит от самого  $X_j$ , то механизм пропусков является *существенным* (non-ignorable), и для корректного анализа данных необходимо знать этот механизм. Введенные выше понятия относятся к отдельным переменным, и в пределах одной и той же базы данных можно наблюдать все эти варианты. Можно построить тесты, отличающие MAR от MCAR, однако по данным невозможно отличить, являются ли они MAR, или же механизм пропусков *существенен*.

В качестве пояснения чаще всего приводится пример ответов на вопросы, связанные с доходом респондентов. Если вероятность сообщить свой доход постоянна для всех респондентов (например, 15%), то данные следуют MCAR. Если эта вероятность связана с другими переменными (скажем, люди с более низким образованием реже указывают

свой доход), то данные следуют MAR. Наконец, если более богатые люди менее охотно указывают свой доход, то механизм пропусков является существенным, и это, увы, наиболее правдоподобный вариант.

Перейдем теперь к рассмотрению методов анализа, используемых на практике.

### Анализ имеющихся данных

Наиболее естественным способом анализа данных с пропусками кажется анализ по всем имеющимся данным, т.е. с использованием тех наблюдений, по которым наблюдаются все интересующие исследователя переменные (complete case analysis). В свете вышесказанного очевидно, что он дает несмещенные оценки только тогда, когда данные следуют MCAR. Иногда можно использовать для отдельных фрагментов анализа разные наблюдения на основании доступности тех или иных данных — например, для расчета корреляций использовать не только наблюдения, в которых наблюдаются *все* переменные, корреляции которых необходимо посчитать ...

**Stata** ... как это делает команда `correlate` ...

а и те наблюдения, по которым имеются наблюдения конкретной пары переменных

**Stata** ... как это делает `pwcorr`.

Такой метод можно назвать методом доступных случаев (available case analysis). Очевидный его недостаток — полученная таким образом корреляционная матрица может не быть положительно определенной. Естественно, оговорка относительно MCAR относится и к этому случаю.

Еще одним популярным способом скорректировать выборку при наличии пропусков является использование весов. Типичным примером являются пост-стратификационные веса в стратифицированных выборочных обследованиях. Эти веса соотносят количество запланированных наблюдений, которые должны были быть получены в данной страте, и количество реально наблюдавшихся выборочных единиц.

### “Пополнение” данных

Следующим по популярности подходом к анализу неполных данных является метод “вписывания”, или “пополнения” данных (imputation): на основании тех или иных со-

ображений сам исследователь или его программа вписывает на место пропущенных данных какие-то осмысленные, на взгляд исследователя или программы, цифры. В какой-то степени похожей задачей являются задачи интерполяции и экстраполяции, когда по известным значениям функции в нескольких точках необходимо построить значения функции в других точках.

**Stata**

Стандартный метод, предоставляемый пакетом Stata — детерминистическое пополнение данных на основе линейной регрессии. А именно: команда `impute` для каждого наблюдения (точнее, для каждой группы наблюдений с одинаковой структурой пропусков) оценивает линейную регрессию по имеющимся переменным в качестве регрессоров и пропущенными переменными в качестве зависимой переменной (дополнительно используя, естественно, все случаи, для которых эта переменная доступна наряду с остальными имеющимися переменными) и строит прогнозное значение по этой регрессии.

Метод пополнения данных по линейной модели вполне работоспособен тогда, когда данные следуют MAR, и когда линейная модель действительно адекватно описывает данные.

В стратифицированных обследованиях популярен другой метод, называемый методом “горячей колоды” (`hot deck imputation`). Он, как, впрочем, и восстановление по линейной модели, обыгрывает идею восстановления данных по условному распределению: если условием является категоричная переменная (возможно, многомерная), то пропущенные данные можно подставить из числа наблюдаемых в той же группе (или, в некотором более общем виде, подставить значение, наблюдаемое в “похожем” по прочим признакам наблюдении). В простейшем виде этот метод восстанавливает пропуски, пользуясь наблюдениями в той же страте. Теоретические свойства этой процедуры не вполне ясны.

**Stata**

Имеется пользовательская команда `hotdeck`, выполняющая пополнение данных по этому методу ((Mander and Clayton 1999)).

Наконец, “венцом творения” в области восстановления пропущенных данных на данный момент является метод множественного восстановления (`multiple imputation`), предложенный в конце 70-х Дональдом Рубином Rubin (1978). Его идея состоит в том, чтобы восстановить данные не один, а несколько раз, оценить требуемые модели с по-



мощью стандартных методов анализа полных данных, а затем подходящим образом обобщить результаты оценивания. Обычно обобщение сводится к усреднению точечных оценок и вычислению дисперсии полученной оценки как взвешенной суммы оценок дисперсий отдельных точечных оценок (within variance) и разброса между отдельными вычислительными экспериментами (between variance). В качестве модели происхождения данных используется многомерное нормальное распределение; число повторов обычно невелико — от трех до пяти. Ограничением данной модели является предположение о том, что данные следуют MAR.

**Stata**

Автору неизвестны программные модули Stata, которые выполняли бы множественное пополнение данных, хотя пользователи пакета неоднократно высказывали свои пожелания о том, что такие процедуры необходимо иметь.

## Методы на основе ММП

Принципиально иным подходом к анализу пропущенных данных является оценивание моделей на основе метода максимального правдоподобия, скорректированного на пропуски. Пусть данные, которыми располагает исследователь, имеют вид  $Y = (Y_{miss}, Y_{obs})$ , где  $Y_{obs}$  — это реально наблюдаемые величины, а  $Y_{miss}$  — пропущенные, которые исследователь мог бы наблюдать, если бы данные были полными.

Для стандартных моделей функция правдоподобия для всех данных, в т.ч. ненаблюдаемых, может быть сравнительно легко записана в виде  $L(\theta|Y) = f(Y|\theta)$ . Величина, к которой необходимо свести задачу —  $L(\theta|Y_{obs})$ . Сделав определенные предположения о механизме, согласно которому данные оказываются пропущенными  $R_{ij} = I(y_{ij} \text{ наблюдается})$  со своей функцией распределения  $g(R|Y, \psi)$ <sup>15</sup>, можно получить общую функцию правдоподобия в виде

$$L(\theta, \psi|Y_{obs}, R) = \int f(Y_{obs}, Y_{miss}|\theta)g(R|Y_{obs}, Y_{miss}, \psi)dY_{miss} \quad (2.56)$$

При определенных условиях интегрирование в правой части можно провести в явном виде, либо факторизовать задачу, разложив функцию правдоподобия на последовательно интегрирующиеся сомножители.

---

<sup>15</sup> Очевидно,  $R$  наблюдается всегда.

Эlegantным решением многих задач с пропущенными данными является EM-алгоритм, итеративно чередующий подстановку оценок вместо пропущенных данных (по определенной параметрической модели) и получение новых оценок параметров по пополненной таким образом выборке. Классической работой на эту тему, в которой доказаны теоретические свойства EM-алгоритма (сходимость алгоритма, сходимость к критической точке функции правдоподобия, скорость сходимости в зависимости от количества доступных данных), является Dempster et. al. (1977), однако Little and Rubin (1987) считают, что самые ранние аналоги EM-алгоритма были предложены еще в 1920-е гг. Оказывается, что довольно большое число задач может быть переформулировано в терминах EM-алгоритма за счет введения дополнительных переменных — например, в задаче кластерного анализа такой переменной является функция принадлежности, т.е. номер кластера, к которому принадлежит наблюдение.

Название “EM-алгоритм” связано с двумя его шагами, обрабатываемыми на каждой итерации. Шаг “E” (expectation) — это вычисление условного ожидания “пропусков” при условии наблюдающихся данных и текущих значений параметров. Во многих задачах (в частности, при анализе данных из экспоненциального семейства, включающего в себя такие распределения, как нормальное, биномиальное, Пуассона и Бернулли, возможно, в сочетаниях) этот шаг напрямую не выполняется, поскольку функция правдоподобия зависит от данных только через достаточные статистики ((Закс 1978)), и поэтому на шаге E можно посчитать условные ожидания этих достаточных статистик. Шаг “M” представляет собой максимизацию функции правдоподобия (в соответствии с методами анализа для полных данных), в которую подставлены оценки пропущенных данных (или достаточных статистик), полученные на шаге E. Обобщенные EM-алгоритмы ограничиваются тем, что просто увеличивают значение функции правдоподобия на каждом шаге. Итерации прекращаются, когда приращение функции правдоподобия на очередном шаге меньше заданного уровня (скажем,  $10^{-6}$ ).

## 2.5 Диагностика регрессий

Как можно обнаружить, что с регрессией “что-то не в порядке”? Выше были упомянуты тесты на нарушение предположений классической модели — гетероскедастичность, нелинейность и т. п., а также соответствующие им команды пакета Stata. Ниже будет

приведена сводка этих диагностических тестов, а сейчас рассмотрим более подробно, как находить *выделяющиеся наблюдения*, которые могут существенно исказить оценки коэффициентов.

**Stata**

В пакете Stata имеется достаточно обширный спектр средств диагностики регрессий, некоторые из которых уже упомянуты выше, а некоторые будут рассмотрены ниже. Справку по этим средствам можно найти по ключевым словам `regdiag` и `diagplots`.

### 2.5.1 Сводка методов диагностики

Сведем вышеперечисленные методы диагностики регрессий в единую таблицу.

**Stata**

После оценивания регрессии Stata сохраняет информацию об оцененной модели до следующей процедуры оценивания параметров (или до целенаправленного сброса результатов оценивания), поэтому можно, отдав один раз команду `regress`, после этого последовательно отдавать диагностические команды, проводить тесты на коэффициенты или получать прогнозные значения, не прогоняя регрессию заново. Все это объяснено в `tutorial regress` и авторском `tutorial aboutreg`.

Таблица 2.1: Диагностика регрессий

Название теста	Принцип	“Плохие” признаки	Stata
<i>Коррелированность ошибок</i>			
Тест Дарбина–Уотсона	$H_0 : E\varepsilon_t\varepsilon_{t-1} = 0$	Статистика DW ближе к 0 или к 4, чем к 2	<code>regress → dwstat</code>
<i>Гетероскедастичность: дисперсия не постоянна</i>			
Тест Кука–Вайсберга	$H_0 : \ln \sigma_i = \gamma^T z_i$	Значимость доп. регрессии: $F, \chi^2 \rightarrow \infty$	<code>regress → hettest</code>
Визуальный анализ	Графики частных регрессий и остатков-прогнозов	Четко выраженное увеличение разброса	<code>regress → avplot; rvfplot</code>
<i>Мультиколлинеарность</i>			
Главные компоненты	Выявление осей, возле которых группируются данные	Высокое отношение собственных значений ков. м-цы $\lambda_{max}/\lambda_{min} \gg 1$	<code>factor, pc</code>
VIF	Оценка увеличения дисперсии оценок коэффициентов из-за мультиколлинеарности	Индивидуальные значения $VIF > 4$ ( $\sqrt{VIF} > 2$ )	<code>regress → vif</code>
<i>Нелинейность</i>			
RESET-тест Рамсея	Регрессия зависимой переменной на степени объясняющих переменных или прогнозных значений	$F, \chi^2 \rightarrow \infty$	<code>regress → ovtest</code>
Визуальный анализ	Графики частных регрессий, остатков-прогнозов	Наличие четко выраженных кривых вместо случайного разброса точек	<code>regress → avplot; rvfplot; cprplot</code>

Название теста	Принцип	“Плохие” признаки	Stata
<i>Робастность, выбросы</i>			
Форма распределений	Информация о характеристиках распределения (асимметрия, тяжелые хвосты)	Значимо отличные от 0 значения коэффициентов асимметрии и эксцесса остатков, наличие тяжелых хвостов; несовпадение с прямой на нормальной бумаге	<code>summarize;</code> <code>sktest;</code> <code>graph</code> <code>переменная,</code> <code>norm;</code> <code>kdensity;</code> <code>qnorm</code>
<i>D</i> -статистика Кука, <i>DFFITS</i> , <i>DFBETA</i>	Идентификация выделяющихся наблюдений	Точки с высоким значением статистик влияния	<code>regress →</code> <code>predict,</code> <code>cooksdi;</code> <code>predict,</code> <code>dfit;</code> <code>predict,</code> <code>dfbeta</code>
Визуальный анализ	Графики частных регрессий и остатков-прогнозов	Отдельно отстоящие точки	<code>avplot;</code> <code>rvfplot</code>
<i>Стохастичность регрессоров</i>			
Тест Хаусмана	Сравнение эффективной (при $H_0$ ), но несостоятельной (при $H_a$ ) модели с состоятельной (при обеих гипотезах), но менее эффективной (при $H_0$ )	$\chi^2 \rightarrow \infty$	<code>hausman</code>

©С. О. Колеников

## 2.5.2 Пример анализа регрессии

В этом подразделе мы приведем пример “разбора полетов” с применением описанных выше средств диагностики.

В нашем примере будет использована регрессия 1 из обучающей программы `tutorial aboutreg`. В этом уроке, конечно, есть гораздо больше, чем эта регрессия, но для получения приводимой ниже таблицы результатов и ее обсуждения в Stata можно отдать команды:

```
. use auto, clear
. regress price mpg foreign weight
```

Stata выводит следующую таблицу результатов регрессии:

Таблица 2.2: Пример распечатки регрессии в пакете Stata

Source	SS	df	MS	Number of obs = 74		
Model	317252881	3	105750960	F( 3, 70) =	23.29	
Residual	317812515	70	4540178.78	Prob > F =	0.0000	
Total	635065396	73	8699525.97	R-squared =	0.4996	
				Adj R-squared =	0.4781	
				Root MSE =	2130.8	

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mpg	21.8536	74.22114	0.294	0.769	-126.1758	169.883
weight	3.464706	.630749	5.493	0.000	2.206717	4.722695
foreign	3673.06	683.9783	5.370	0.000	2308.909	5037.212
_cons	-5853.696	3376.987	-1.733	0.087	-12588.88	881.4931

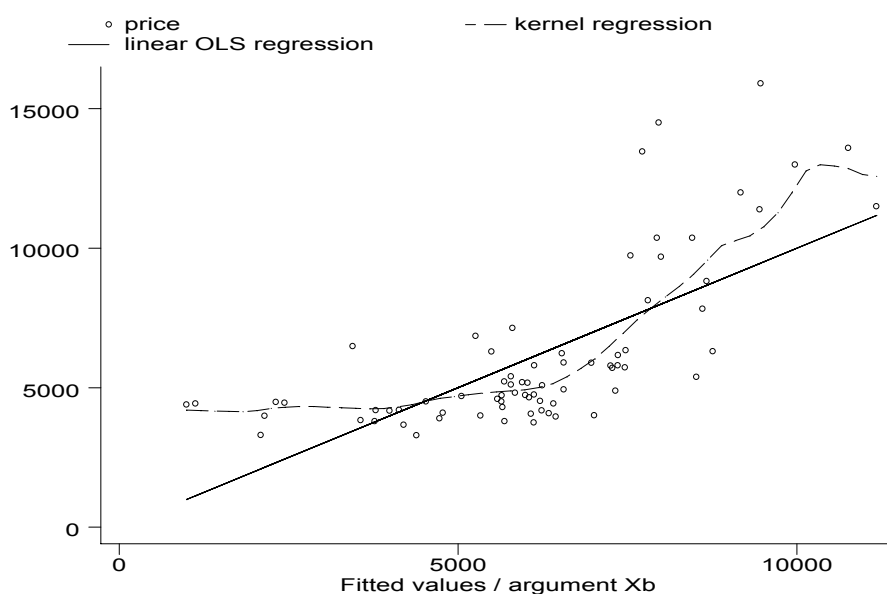
Здесь в левом верхнем углу — таблица дисперсионного анализа (с указанием суммы квадратов и доли дисперсии  $y$ , объясненных моделью, суммы квадратов остатков и их дисперсии, общая сумма квадратов и дисперсия  $y$ ), справа вверху — прочая информация, связанная с регрессией (количество наблюдений, общая  $F$ -статистика для гипотезы

$H_0$ : все коэффициенты равны нулю, кроме константы; статистики  $R^2$  и  $R_{adj}^2$  и оценка стандартного отклонения остатков). Наконец, в нижней части таблицы приведены оценки коэффициентов и их стандартных ошибок,  $t$ -статистики для гипотез  $H_0 : \beta_k = 0$  и доверительные интервалы.

Результаты аналитических тестов (таких, как `ovtest`, `hettest` и прочих) оставляются на научное любопытство читателя, а ниже будут приведены основные результаты визуального анализа.

Начнем с графика, представляющего проекцию облака точек на ось прогнозных значений (fitted values). На рис. 2.3 представлены, помимо самих точек, линейный прогноз (биссектриса графика) и непараметрическая ядерная оценка (`kernreg`, см. ниже раздел 2.6.5). На этом графике видно, что линейная аппроксимация функции регрессии не является адекватной, что и подтверждается тестом Рамсея на нелинейность (2.44).

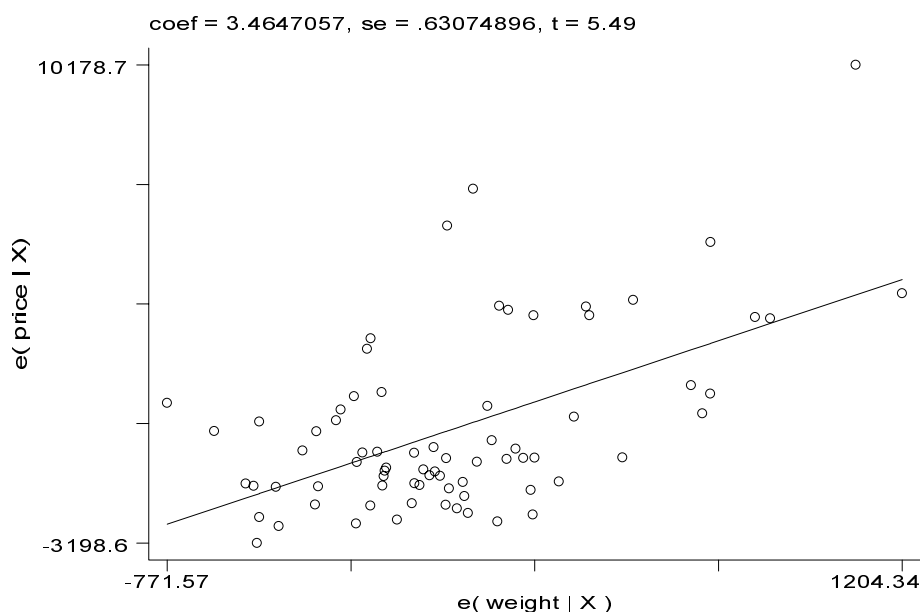
Рис. 2.3: Регрессия в пространстве прогнозных значений: прямая, полученная по МНК, и непараметрическая оценка кривой регрессии. Видно значительное расхождение.



Иногда нелинейность, а также гетероскедастичность, относительно отдельных переменных можно выявить с помощью графика частной регрессии (см. стр. 2.52). В данном случае (рис. 2.4), впрочем, ничего особенного не наблюдается.

Одним из наиболее важных и информативных графиков является график, связы-

Рис. 2.4: График частной регрессии для переменной `weight` (`avplot weight`).



вающий регрессионные остатки и прогнозные значения. В случае приведенной выше регрессии этот график, к счастью для пояснительных целей и к несчастью для научных, показывает едва ли не все дефекты данной регрессии из числа рассматриваемых в этой книге.

В простейшем представлении (рис. 2.5) мы видим, что остатки почти линейно связаны с прогнозными значениями в первых двух третях графика, после чего их дисперсия заметно возрастает, они смещаются вверх, и за счет этого их сумма равна нулю. Такое поведение, естественно, неудовлетворительно, поскольку в идеале мы рассчитываем увидеть “белый шум”, т.е. график без каких-либо очевидных зависимостей.

Более того, если приложить определенные усилия (см. подпись к рис. 2.6 по поводу использованного синтаксиса команды `rvfplot`), то можно построить красивый график, демонстрирующий нелинейность соотношения между прогнозными значениями и остатками.

Влияние отдельных наблюдений исследуется при помощи статистик, получаемых командой `predict` с такими опциями, как `rstudent`, `dfbeta`, `dffits`, `cooks` и `hat`<sup>16</sup>.

<sup>16</sup> Подчеркивания показывают минимально возможные сокращения; см. раздел 3.1



Рис. 2.5: Диаграмма рассеяния остатков (`rvfplot, yline(0)`).

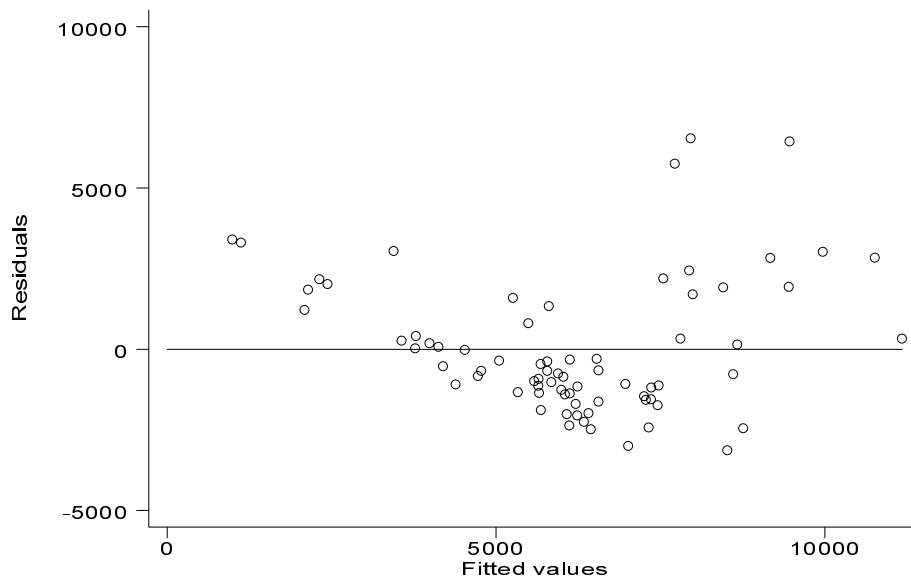
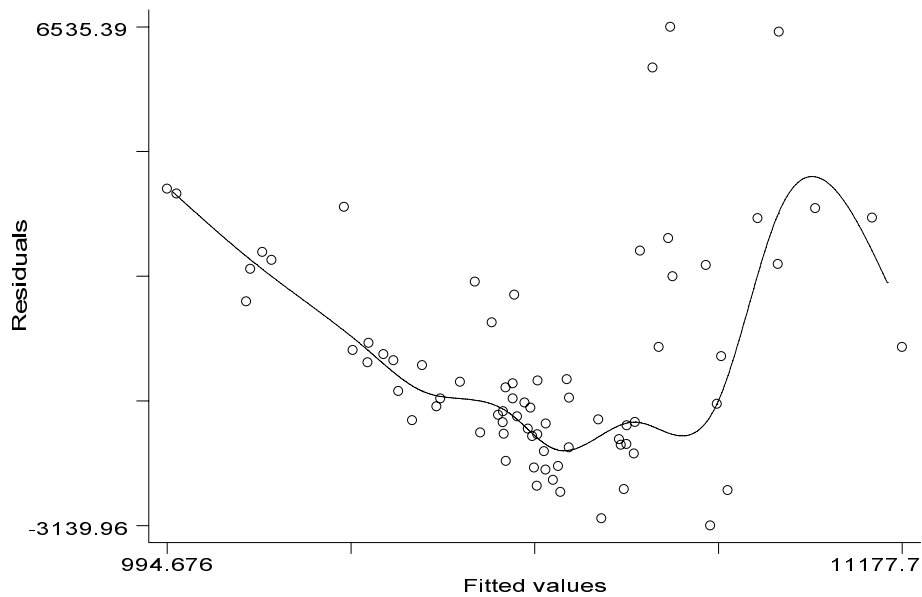
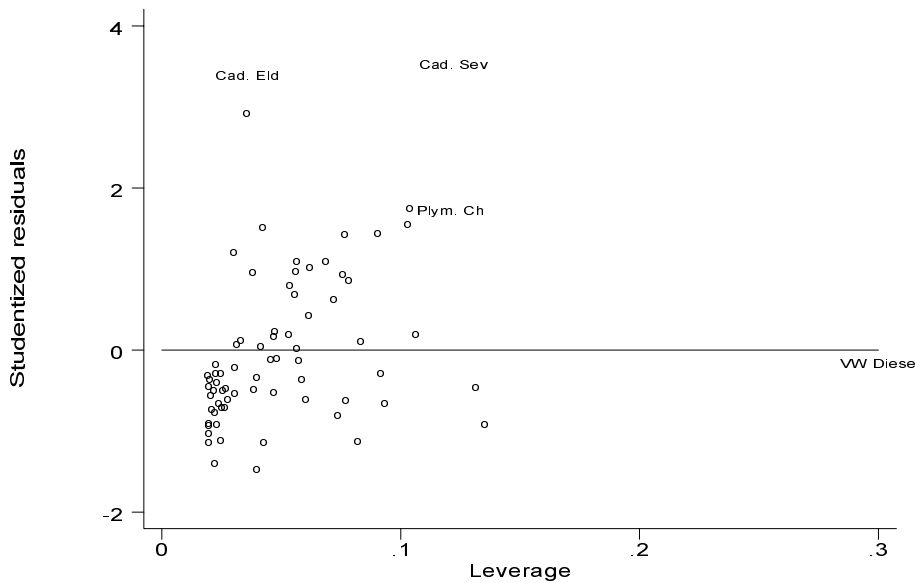


Рис. 2.6: Диаграмма рассеяния остатков (`rvfplot, c(s) bands(10) d(50)`).



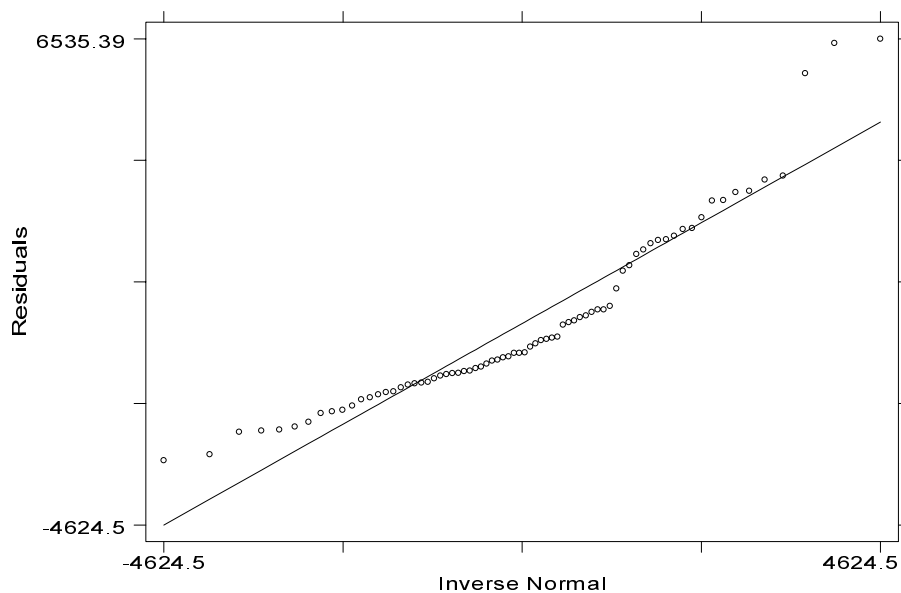
На рис. 2.7 приведен график, связывающий относительное влияние каждого наблюдения (leverage) и величину студентизированного остатка. Произведение этих величин составляет расстояние Кука  $D$ . Более подробное объяснение см. в разделе 2.4.3. Наблюдения, которые могут оказывать существенное влияние на коэффициенты, промаркированы названиями соответствующих автомобилей. Чтобы представить себе, насколько существенно могут сместиться оценки коэффициентов при воздействии выбросов, найдите в выборке наблюдение с максимальным значением  $D$  и проведите оценку параметров регрессионной модели без этого наблюдения (подсказка: `predict ... , cooks` и `regress ... , if ... < ...`, где вместо `...` вы подставите что-нибудь более осмысленное).

Рис. 2.7: Статистики, характеризующие влияние отдельных наблюдений.



Дополнительным подтверждением тому, что регрессионные остатки в данной модели не обладают хорошими статистическими свойствами, может служить график для диагностики отклонений распределения остатков от нормального. На рис. 2.8 отложены квантили распределения остатков и нормального распределения с аналогичным средним и дисперсией. Точки не лежат на хорошей и аккуратной прямой, а три точки в правой части графика означают тяжелые хвосты остатков: наблюдаемые квантили больше, чем соответствующие процентные точки нормального распределения.

Рис. 2.8: График квантилей нормального распределения для остатков регрессии (1) (`qnorm ...`).



На этом, безусловно, графические средства анализа данных в пакете Stata не исчерпываются. Автор призывает читателя углубить свои знания и закрепить практические навыки, изучив обучающие программы `tutorial regress`, `tutorial aboutreg` и `tutorial graphics`.

## 2.6 Альтернативные спецификации статистических зависимостей

В современной эконометрической практике применяется очень много различных вариантов описания зависимостей одних величин от других, объединяемых в общее понятие “регрессии”; МНК-оценки как таковые применяются далеко не всегда.

Выше упоминались такие модели, как временные ряды, робастные регрессии, ридж-оценки и др. Расскажем еще о нескольких видах регрессионных моделей, встречающихся в литературе.

## 2.6.1 Данные особой структуры и обобщенный МНК

Как уже упоминалось выше, учет структуры матрицы ковариации ошибок может дать выигрыш в эффективности оценок. Иногда этот выигрыш может даже быть “в разы”. Неверное же представление о стохастической структуре модели может приводить к смещению оценок дисперсии, что искажает выводы на основе  $t$ -,  $F$ - и  $\chi^2$ -статистик.

Одним из частных примеров моделей со сложной структурой ошибок являются *панельные модели*, насчитывающие три измерения данных: переменные – объекты (исследуемые единицы) – время. Для них разработаны специальные методы анализа ((Maddala 1993), (Baltagi 1995)). Как правило, индивидуальные эффекты выделяются в виде аддитивной составляющей:

$$y_{it} = x_{it}^T \beta + u_i + \varepsilon_{it} \quad (2.57)$$

Эти данные порождаются длительными обследованиями, в которых одни и те же индивидуумы (домохозяйства, фирмы и т. п.) опрашиваются последовательно через определенные интервалы времени (как правило, раз в год или в квартал).

### Stata

Команды пакета Stata для анализа панельных данных имеют префикс `xt`, обозначающий наличие как структурной стохастики `x`, так и временной компоненты `t`. Панельные регрессии вызываются командой `xtreg`: с фиксированным эффектом (англ. fixed effect) — с опцией `xtreg ... , fe`, со случайным эффектом (англ. random effect) — с опцией `xtreg ... , re`. Для использования этих команд данные должны быть приведены в “длинную” форму — см. `reshape`, с. 81.

Зависимость между наблюдениями возникает также в стратифицированных выборках, к которым относится большинство крупномасштабных экономических исследований (в т.ч. цитируемое далее обследование RLMS, гл. 4). Выборка для таких исследований разрабатывается следующим образом. Выбираются однородные (по социальным, экономическим, демографическим показателям, если речь идет о населении; по объему выпуска и занятости, по отраслевой принадлежности, если речь идет о предприятиях) группы объектов — *страты* (так, в RLMS стратой является административный район; область была сочтена разработчиками слишком крупным объектом). Из набора этих страт, полностью покрывающих интересующую исследователя совокупность, выбираются случайным образом с вероятностями, пропорциональными размеру страт,

некоторое малое число первичных единиц выборки (primary sampling units — PSU). Затем в пределах этих PSU процедура случайного выбора повторяется с использованием более мелких группировок (в RLMS — участки переписи населения, избирательные участки, почтовые отделения), и так далее, пока единицей случайного выбора не будут сами объекты — домохозяйства, предприятия и т.п. Процедура случайного отбора может быть модифицирована, с тем, чтобы в выборку не попали “слишком близкие” объекты (например, соседи по лестничной площадке).

Ввиду подобной структуры выборки, отдельные наблюдения, в отличие от истинно случайной выборки, не являются независимыми. Действительно, если в выборке присутствует объект из некоторого PSU данной страты, то условная вероятность (при указанном выше условии включения элемента в выборку) того, что другие элементы этого же PSU попадут в выборку, больше, чем условная вероятность того, что в выборку попадут элементы из других PSU этой страты. Индивиды, относящиеся к одной структурной единице выборки, могут находиться под воздействием специфических для данной единицы ошибок, что требует включения дополнительных членов в уравнение регрессии в стиле дисперсионного анализа:

$$y_{it} = x_{it}^T \beta + \nu_{PSU} + u_i + \varepsilon_{it} \quad (2.58)$$

Подобная зависимость наблюдений будет сказываться на всех оценках и статистических выводах, которые делаются на основе результатов анализа подобной стратифицированной выборки. В частности, наивные оценки вторых моментов (дисперсий) будут сильно занижены, поскольку основной вклад в дисперсию будет связан с самым первым уровнем стратификации.

**Stata**

Пакет Stata обладает весьма обширным набором средств, позволяющих учитывать стратификационный характер выборок — это около двух десятков команд с префиксом `svy`. Для использования этих команд необходимо указать, какие переменные несут в себе информацию о структуре выборке (`svyset` и `svydes`). Иногда вместо `svy`-команд можно воспользоваться опцией `, cluster()`, которую можно использовать с большинством команд Stata, оценивающих параметрические модели, в т.ч. с командой `regress`. Для уточнения оценок параметров и вторых моментов регрессионных моделей можно использовать веса (см. `help weights`), связанные с вероятностью включения в выборку отдельных наблюдений (т.е. веса, учитывающие стратифика-

ционное происхождение выборки) — `pweight` (сокр. от probability weights) — если такие веса входят в базы данных обследований.

## 2.6.2 Системы одновременных уравнений

Подобные модели описывают явления, в которых несколько переменных определяется одновременно, как некоторое равновесие экономической системы. Типичным примером СОУ является равновесие рыночных спроса и предложения.

Проблема одновременности тесно связана с уже упоминавшейся проблемой стохастичности регрессоров. Дело в том, что эндогенные переменные (т. е. переменные, определяемые в равновесии; сопутствующее понятие — экзогенные, или заданные извне, переменные) коррелированы с ошибками, и поэтому оценивание по методу наименьших квадратов приводит к смещенным и несостоятельным оценкам. В зависимости от структуры уравнений, коэффициенты при эндогенных переменных могут быть, а могут и не быть идентифицируемы.

Для разрешения проблемы эндогенности используются двух- и трехшаговый метод наименьших квадратов (3SLS).

**Stata** И соответствующая команда называется `reg3`.

## 2.6.3 Модели с дискретными и другими ограниченными зависимыми переменными

Часто возникает потребность в анализе моделей, в которых в качестве зависимой переменной фигурирует качественная величина, например, наличие-отсутствие или отказ-участие. Естественным образом такие величины кодируются как 0/1 и называются на статистическом жаргоне “успех-неуспех”. Они имеют (условное) биномиальное распределение. Метод наименьших квадратов, применяемый напрямую, будет как минимум страдать от гетероскедастичности: ошибки должны быть устроены так, чтобы в результате получилось значение 0 или 1. Возможно, что для каких-то наблюдений и в случае успеха, и в случае неуспеха ошибка должна быть отрицательной (или положительной), и тогда будет нарушаться и предположение об (условной) центральности ошибок.

Для разрешения подобных трудностей моделируется непосредственно вероятность успеха (т. е. регистрации 1 в принятой кодировке исходов). При дополнительном предположении наличия индексной функции, являющейся линейной комбинацией известных переменных,

$$\begin{aligned} P(y = 1|x) &= F(x^T \beta) \\ P(y = 0|x) &= 1 - F(x^T \beta) \end{aligned} \quad (2.59)$$

Эта величина должна лежать в промежутке  $[0, 1]$ , что накладывает ограничения на вид функции  $F$ . Чаще всего в качестве этой функции используется та или иная функция распределения. В подавляющем большинстве работ используется одна из двух функций распределения — стандартной нормальной величины или логистического распределения:

$$F(z) = \frac{1}{1 + \exp(-z)} \quad (2.60)$$

Соответствующие модели носят название *пробит*- и *логит*-моделей; для второй еще используется название *логистическая регрессия*. Существенных оснований предпочитать одну модель другой, видимо, нет. Обе функции распределения симметричны, а различия между ними не так велики:  $\sup_{x \in (-\infty, +\infty)} |F_{\text{logit}}(x) - F_{N(0,1)}(x)| < 0.02$ , но у логистического распределения более тяжелые хвосты. Пробит-модель привлекательна тем, что в ней используется самое типичное распределение в мире — нормальное, и поэтому она удобна для анализа моделей с многомерным нормальным распределением ошибок, если зависимых переменных несколько. В качестве примера можно привести модель Хекмана регрессии с внешним выбором наблюдений (Heckman sample selection model)<sup>17</sup>. С другой стороны, логит-модель допускает достаточно широкий спектр средств анализа качества приближения (goodness of fit).

Иногда встречается также асимметричная функция дополнительных логарифмов, называемая также функцией Гомперца (Gompertz, соответственно, гомпит/гомпит-модель):

$$F(z) = 1 - \exp[-\exp(z)] \quad (2.61)$$

---

<sup>17</sup> В этой модели вероятность попадания объекта в выборку зависит от известных факторов. В связи с непредставительностью выборки относительно исследуемой совокупности многие выборочные статистики, в т.ч. оценки МНК, оказываются смещенными (Greene 1997); модель Хекмана предлагает способ устранения этого смещения. Именно за эту работу профессор Чикагского университета Джеймс Хекман был удостоен Нобелевской премии по экономике 2000 г.

**Stata** Соответствующие регрессии в пакете Stata вызываются командами `probit`, `logit` и `cloglog`.

Оценивание коэффициентов в данных моделях производится по методу максимального правдоподобия. Если наблюдения независимы, то функция правдоподобия для отдельных наблюдений имеет вид:

$$L(y_i, x_i, \beta, F) = \begin{cases} F(x_i^T \beta), & y_i = 1 \\ 1 - F(x_i^T \beta), & y_i = 0 \end{cases} \quad (2.62)$$

что может быть очень удачно переписано как

$$L(y_i, x_i, \beta, F) = F(x_i^T \beta)^{y_i} (1 - F(x_i^T \beta))^{1-y_i} \quad (2.63)$$

Тогда общая функция правдоподобия имеет вид:

$$\ln L(y, \mathbf{X}, \beta, F) = \sum_{i=1}^n \{y_i \ln F(x_i^T \beta) + (1 - y_i) \ln(1 - F(x_i^T \beta))\} \quad (2.64)$$

Задача максимизации этой функции по  $\beta$  решается численными методами.

**Stata** Одним из очень существенных достоинств пакета Stata является доступ программистов к алгоритму численного решения задач максимизации функции правдоподобия пользователя (Gould, Sribney 1999). Оценивание по методу максимального правдоподобия осуществляется командами набора `ml`.

К оценкам коэффициентов пробит- и логит-регрессий относятся все комментарии о методе максимального правдоподобия (Кендалл, Стьюарт 1973). В определенном классе оценок оценки максимального правдоподобия являются асимптотически эффективными, однако они очень чувствительны к нарушениям формы распределения. Тесты на значения коэффициентов или их линейных комбинаций (в т.ч. на значимость регрессии в целом) осуществляются с помощью статистики отношения правдоподобия или ее асимптотических аналогов — теста Вальда (Wald test) и множителей Лагранжа (LM test, Lagrange multiplier test, score test). Все эти тесты имеют асимптотическое распределение  $\chi^2$  с числом степеней свободы, равном числу накладываемых ограничений (Айвазян, Мхитарян 1998), (Greene 1997).

Определенное неудобство логит- и пробит-моделей (как, впрочем, и всех нелинейных моделей) заключается в том, что оценки коэффициентов, в отличие от линейной



регрессии, не могут быть интерпретированы как предельные эффекты (т.е. изменения зависимой переменной при изменении независимой, в том числе бинарной, на единицу), поскольку предельные эффекты в нелинейных моделях зависят от точки, в которой берется такое приращение. Для того, чтобы получить хоть какое-то представление о предельных эффектах, можно рассчитать предельные эффекты для выборочного среднего по всем независимым переменным, или рассчитать предельные эффекты во всех точках и усреднить.

**Stata**

В шестой версии функцию расчета предельных эффектов для пробит-модели выполняет команда `dprobit`, которая оценивает пробит-модель точно так же, как `probit`, но вместо коэффициентов выводит предельные эффекты для выборочных средних всех регрессоров. В седьмой версии пакета Stata появилась очень удобная команда `mfx`, которая рассчитывает эти самые предельные эффекты для произвольной оцененной модели.

## 2.6.4 Квантильные регрессии

Иногда предметом интереса исследователя могут быть не средние значения зависимой переменной при фиксированных объясняющих, а определенные квантили распределения. В исследованиях финансового риска интерес могут представлять, к примеру, 5% или 10% точки, и т.д. Кроме того, знание набора (условных) квантилей позволит понять, меняется ли форма распределения в зависимости от объясняющих переменных.

$$P[y < m|x] = p \tag{2.65}$$

Примером квантильной регрессии является упоминавшаяся ранее в контексте проблем робастности *условная медиана* при  $p = 0.5$ .

**Stata**

Квантильные регрессии реализованы в пакете Stata командой `qreg`. Опция `qreg ... , quantile()` этой команды позволяет явно указать, квантиль какого уровня  $p$  следует исследовать.

Можно показать, что медианная регрессия является решением задачи минимизации

суммы абсолютных отклонений (ср. (2.11)):

$$\sum_{i=1}^N |y_i - x_i\beta| \rightarrow \min \quad (2.66)$$

Данная задача решается симплекс-методом или другими методами линейного программирования.

### 2.6.5 Непараметрические регрессии

Методы непараметрической регрессии являются формализацией интуитивного понятия сглаживания “на глаз”. Если мы будем проводить на глаз кривую на двумерном графике рассеяния, чтобы описать примерный вид зависимости  $E[y|x]$ , мы будем учитывать, где лежат наблюдаемые значения  $y$  вблизи интересующей нас точки  $x$ , повторяя характерные пики и впадины кривой регрессии (см., например, рис. 2.3).

Непараметрическая оценка кривой регрессии имеет вид:

$$\hat{m}(x) = n^{-1} \sum_{i=1}^n W_{ni}(x)y_i, \quad (2.67)$$

где  $W_{ni}$  — веса сглаживания, которые могут зависеть от всего вектора  $x$ . В такой постановке задачу сглаживания можно интерпретировать как задачу нахождения оценки локально взвешенных наименьших квадратов:

$$n^{-1} \sum_{i=1}^n W_{ni}(x)(y_i - \hat{m}(x_i))^2 \rightarrow \min_{m(x)} \quad (2.68)$$

**Stata**

Один из методов, явно использующий многократно прогоняемые регрессии для локального сглаживания — `lowess` (locally weighted smoothing) Fox (1997). Его реализация в пакете Stata осуществлена командой `ksm` с опцией `ksm ... , lowess`.

В эконометрической литературе варианты непараметрической регрессии известны под названиями локальной регрессии (local regression) и “катящейся” регрессии (rolling regression). В них используется та же самая идея локального взвешивания.

Формализация близости заключается во введении “ядра сглаживания” с определенной “шириной окна”. Точки, не попадающие в ядро, будут иметь нулевой вес; таким образом, внимание процедуры сглаживания будет сосредоточено вблизи требуемой точки.

Понятие ядра и его применение в непараметрической регрессии формализуется следующим образом Хардле (1993):

$$W_{ni}(x) = K_{h_n}(x - x_i) / \hat{f}_{h_n}(x) \quad (2.69)$$

$$\hat{f}_{h_n}(x) = n^{-1} \sum_{i=1}^n K_{h_n}(x - x_i) \quad (2.70)$$

$$K_{h_n}(u) = h_n^{-1} K(u/h_n) \quad (2.71)$$

$$\int K(u) du = 1 \quad (2.72)$$

Здесь (2.70) — непараметрическая (ядерная) оценка плотности в данной точке (называемая также *оценкой Розенבלата-Парзена*), (2.71) — ядро масштаба  $h_n$  (ширина которого может зависеть от числа наблюдений). Нормализация (2.70) гарантирует, что сумма весов равна единице. Полученная таким образом *ядерная оценка* функции регрессии носит название *оценки Надарая-Ватсона*.

Есть ряд наиболее популярных ядерных функций:

$$\text{ядро Епанечникова: } K(u) = 0.75(1 - u^2)I(|u| \leq 1) \quad (2.73)$$

$$\text{квартическое ядро: } K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1) \quad (2.74)$$

$$\text{равномерное ядро: } K(u) = \frac{1}{2} I(|u| \leq 1) \quad (2.75)$$

$$\text{треугольное ядро: } K(u) = (1 - |u|) I(|u| \leq 1) \quad (2.76)$$

$$\text{нормальное (гауссово) квазиядро: } K(u) = \frac{1}{\sqrt{2\pi}} \exp[-u^2/2] \quad (2.77)$$

Здесь  $I(\text{условие})$  — индикаторная функция, принимающая значение 1, если условие выполняется, и 0, в противном случае.

Если по отношению к параметрическим моделям всегда могут возникнуть вопросы: “Почему именно такая спецификация модели? Почему именно такая форма ошибок?”, то естественные вопросы к непараметрическим моделям — “Почему именно такая форма ядра? Почему именно такая ширина окна?”. Есть результаты, показывающие, что ядерная оценка будет асимптотически состоятельна независимо от выбора ядра, однако ядро Епанечникова обладает определенными оптимальными свойствами в смысле среднеквадратической ошибки. Что же касается выбора ширины окна  $h_n$ , то выбор слишком малого значения будет означать, что оценка кривой регрессии пройдет через все точки выборки, тогда как слишком большое значение сгладит истинную кривую слишком

сильно<sup>18</sup>. Со статистической точки зрения, задача заключается в том, чтобы соблюсти компромисс между дисперсией точечной оценки и ее смещением. Асимптотически максимальная скорость сходимости среднеквадратической ошибки прогноза составляет в одномерном случае  $n^{-4/9}$  (т. е. медленнее, чем в параметрических задачах), а ширина окна при этом пропорциональна  $n^{-1/9}$ .

**Stata**

Непараметрическая регрессия выполняется командой `kernreg`, входящей в состав дополнения STB-30. Данная команда позволяет указать тип ядра (Епанечникова по умолчанию, равномерное, нормальное, квартическое, триквартическое, треугольное, косинусоидальное), ширину окна, а также точки, в которых будет произведена оценка. Непараметрическая оценка плотности осуществляется встроенной командой `kdensity`, которая изначально существовала как команда STB, а потом стала частью официального дистрибутива Stata.

Наиболее существенным недостатком непараметрической регрессии является ее одномерность. Обобщение на случай многомерного вектора объясняющих переменных, безусловно, возможно — достаточно использовать многомерные плотности, или произведения одномерных ядер — однако число соседей убывает с ростом размерности очень быстро (эффект, известный под названием “проклятие высокой размерности”, *dimensionality curse*), и окно приходится распространять чуть ли не на всю выборку. Кроме того, в многомерных задачах меняется и скорость сходимости, причем, конечно же, в сторону ухудшения.

**Stata**

Во всяком случае, упомянутая выше реализация алгоритма непараметрической регрессии рассчитана на единственный регрессор.

Я бы порекомендовал дополнять параметрические оценки регрессии непараметрическими в целях проверки точности подгонки. Сведенные на одном графике диаграмма рассеяния, предсказанные значения и непараметрическая оценка позволят выявить основные дефекты регрессии: неучтенную нелинейность, гетероскедастичность и т. п., как это сделано на рис. 2.3.

---

<sup>18</sup> При  $h \rightarrow \infty$ ,  $f(x) \rightarrow \bar{y}$ .

## Глава 3

# Краткое описание пакета Stata

Программа Stata StataCorp. (1999, 2001)— 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 80-х гг. В январе 1999 г. была выпущена шестая версия, в декабре 2000 г. — седьмая. Основными достоинствами Stata являются:

- большой спектр реализованных статистических методов (хотя и есть методы, не реализованные практически никак, например, дискриминантный анализ, кластерный анализ, обобщенный метод моментов, ряд других);
- возможности гибкой пакетной обработки данных (т. е. программирования всей последовательности команд, начиная от загрузки данных в память и вплоть до всех деталей анализа). Возможности интерактивного режима работы полностью идентичны 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 интернетовский список рассылки); огромный архив пользовательских программ в открытом доступе;
- возможность максимизации функций правдоподобия, задаваемых пользователем;

- наличие совместимых по функциональным возможностям и форматам данных реализаций для большинства популярных платформ (Windows, Macintosh, UNIX).

По поводу графических средств мнения пользователей разнятся: с одной стороны, они вполне достаточны для текущего графического анализа данных и подготовки научных публикаций (все рисунки в этой книге выполнены в Stata и импортированы в L<sup>A</sup>T<sub>E</sub>X), с другой, несравнимы с графическими возможностями специализированных пакетов типа Harvard Graphics или презентационных программ типа PowerPoint.

Ниже будет приведена сводка наиболее важных команд пакета. Эта сводка вряд ли сможет заменить изучение этих (и, естественно, других) команд по руководствам пользователя или хотя бы по встроенной подсказке Stata (например, не все детали синтаксиса и не все опции могут быть упомянуты в данном кратком введении). Скорее, она поможет найти, какими командами и для чего следует воспользоваться; более полное и точное описание этих команд можно найти во встроенной помощи Stata и в руководствах. Многие команды будут упомянуты лишь на уровне названия (что, впрочем, достаточно для поиска по встроенной подсказке Stata). *Читателю настоятельно советуется овладеть и пользоваться встроенной помощью Stata по командам и деталям внутреннего устройства пакета.*

### 3.1 Договориться: обозначения

Мы будем пользоваться следующими обозначениями, выдержанными в стиле руководств Stata. Так, `command` — команда, которую можно набирать целиком, а можно сократить до первых трех букв (например, `regress` можно написать как `reg`, а можно как `regress`). [в квадратных скобках] будут указаны необязательные фрагменты команды — 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. Через вертикальную черту будут перечисляться возможные варианты: [вариант 1|вариант 2]. Таким образом, запись `describe [переменные | using имя файла]` может разворачиваться в следующие варианты:

```
d
describe
```

```
describe x1 x2 x3
d using source
desc using source.dta
```

Эта команда выдает краткое описание файла данных в памяти Stata или на диске.

Ссылки на руководства также оформляются в стиле Stata: [R] **команда** означает, что эту команду можно найти в четырехтомном справочнике команд (Reference); [U] **3 A brief description of Stata** — это ссылка на Руководство пользователя, а именно на главу 3 в книге User's Guide (для Stata 6) — описание Stata в руководстве пользователя (то, что можно почитать о Stata вместо этого параграфа); [G] **twoway** — описание двумерных графиков в руководстве по графике.

## 3.2 Открыть: установка и запуск Stata

Обычно Stata устанавливается в каталог `c:/stata`, если при установке не было явно указано иное. Исполняемый файл называется `wstata.exe` (Stata for Windows).

Команда `verinst` проверяет корректность установки пакета.

Сам этот исполняемый модуль выполняет сравнительно небольшое число (около 200) базовых процедур. Подавляющее большинство собственно статистических задач выполняется внешними программами с расширением `.ado`, находящимися в каталоге `c:/stata/ado` и его подкаталогах. Эти `ado`-файлы с некоторой степенью условности можно разделить на *базовые* (около 900), отлаженные разработчиком и входящие в комплект поставки Stata, (хотя и в них иногда находят ошибки, и тогда Stata делает официальные обновления `ado`-файлов); *официально распространяемые*, входящие в состав официальных дополнений к Stata — Stata Technical Bulletin, сокращенно STB, которые рассылаются подписчикам и распространяются бесплатно через Internet; и, наконец, *пользовательские*.

При запуске Stata устанавливает ряд внутренних параметров, таких, как объем выделяемой памяти, и некоторые другие (о них можно узнать в [R] **limits** или в подсказке `help limits`). Практически наверняка вам придется менять следующие установки:

```
set memory объем памяти[k|m]
```

Объем памяти, выделяемой операционной системой для Stata. Чтобы отвести 10 мегабайт, надо напечатать: `set memory 10m`. Можно выделить память при запуске па-

раметром командной строки: `wstata /k 10240`. Если количество переменных ограничено 2047, то количество наблюдений может быть ограничено только возможностью выделения памяти операционной системой. При выделении количества памяти, приближающейся к физическому объему ОЗУ (или тем более превышающего этот объем), Stata начинает пользоваться виртуальной памятью (постоянно перезагружаемой с жесткого диска), и работа может замедляться в сотни раз.

`set matsize число`

Максимальный размер матрицы, которую Stata сможет обработать. По умолчанию устанавливается 10. Максимальный размер — 800. Этот параметр влияет на размерность статистических моделей, которые Stata будет в состоянии оценить.

Stata может быть запущена в *пакетном* режиме, в котором она обрабатывает заданную в качестве входного параметра программу<sup>1</sup>, а по завершении выполнения этой программы — передает управление операционной системе (или, попросту говоря, самоликвидируется). Такой вариант запуска задается (в Windows) как `wstata /b do имя файла с программой`.

Выход из Stata осуществляется командой `exit`. Если при этом данные не были сохранены, Stata об этом напомнит.

См. также: [U] 5 Starting and stopping Stata, [U] 6 Troubleshooting starting and stopping Stata

### 3.3 Придти, увидеть, посчитать: интерфейс Stata

Внешний вид Stata (рис. 3.3) несколько отличен от того, что обычно можно увидеть в других статистических пакетах. Внешний аскетизм интерфейса унаследован от идеологии командной строки UNIX, и пользователю Windows требуется некоторое привыкание.

Stata использует в работе несколько окон: окно ввода команд (Stata Command), окно вывода результатов (Stata Results), окно истории, или предыдущих команд (Review), окно переменных (Variables), окно поиска и помощи (Help), графический экран (Graph), окно файла-протокола, или log-файла (Log; в 7-й версии его функцию выполняет окно Viewer). Можно также вызвать окна просмотра данных (Stata Browser) или

---

<sup>1</sup>О программах см. ниже параграф 3.13.





редактирования данных (Stata Editor), а также редактор программ (Stata Do-file Editor). Переключаться между окнами можно, тыкаясь мышкой в любое место на нужном окне, либо через меню **W**indows.

При вводе команд в окне Stata Command можно пользоваться стандартными средствами редактирования в Windows (выделения, стирания, вставки в буфер и из буфера). Можно вызывать предыдущие команды, нажимая PgUp и PgDn, и редактировать их (что очень полезно, если при вводе команды были допущены мелкие опечатки, или если надо добавить какие-то опции или условия к предыдущей команде). Можно менять кое-какие установки Stata в меню **P**refs, например, сохранить текущие установки окон (размеры, положение, шрифты).

См. также: [GSW] , т.е. руководство Getting Started for Windows.

### 3.4 Обобщить: как выглядят команды Stata

Команды Stata, как правило, имеют следующий вид:

**команда** [*список переменных*] [*if условие*] [*in диапазон*] [*using имя файла*] [[*веса*]], [*опции*]

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). Условия *if* и *in* выделяют те наблюдения, для которых необходимо провести анализ (см. ниже параграф 3.6). Если команда предполагает работу с файлами (чтение, объединение и т.п.), то имя файла, с которым необходимо провести указанные действия, передается в конструкции **using**. Если разным наблюдениям необходимо придать разные веса, то для этого используется конструкция типа [**weight=выражение**] (см. **help weights**; квадратные скобки обязательны). Наконец, 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.

Есть несколько исключений из вышеупомянутого синтаксиса, в т.ч. команды, выполняющие повторные действия — см. ниже параграф 3.11.

См. также: [U] **14 Language syntax**

## 3.5 Узнать: помощь

В Windows-версии Stata для поиска нужной информации проще всего воспользоваться меню **Help**, в котором имеются подменю **Search** (поиск по ключевым словам, например, *Durbin Watson statistic*) и **Stata Command** (файл помощи по конкретной команде Stata). Впрочем, практически все то же самое можно сделать с клавиатуры командами **search**, **help** и **whelp**. Содержимое встроенной подсказки полностью дублируется в открытом доступе на сайте Stata: <http://www.stata.com/info/capabilities/>.

Встроенная помощь Stata устроена гипертекстовым образом: если подвести мышку к фрагменту текста, выделенному зеленым цветом, то курсор превратится в ладошку, а если нажать при этом на левую кнопку мыши<sup>2</sup>, то будет выведен соответствующий фрагмент подсказки Stata. Если зеленым цветом помечена ссылка в Internet, то Stata запустит внешний браузер (MS Internet Explorer, Netscape Navigator). В Stata 7 эти действия можно выполнять и с результатами, выводимыми в окно Results.

Полный список стандартных команд, входящих в состав начальной установки Stata, можно найти в меню **Help/Contents** (или по команде **help contents**). Эти команды сгруппированы по тематическим разделам: общее представление о пакете, синтаксис команд, работа с данными, графика, статистические средства, матричные команды, программирование, особенности работы в среде Windows.

Все файлы помощи представляют собой специальным образом отформатированные текстовые файлы с расширением **.hlp**<sup>3</sup>.

В Stata имеются собственные обучающие средства — мини-уроки (являющиеся, с технической точки зрения, специальным видом программ), доступ к которым обеспечивается командой **tutorial**. Они дают краткое введение в пакет, в графические и табличные средства Stata, знакомят с данными, поставляемыми вместе со Stata, и способами перевода текстовых файлов в формат Stata, а также освещают ряд основных статистических команд.

См. также: [U] **8 Stata's on-line help and search facilities**, [U] **9 Stata's on-line tutorials and sample datasets**.

---

<sup>2</sup>Для левшей эта кнопка, возможно, будет правой — имеется в виду та кнопка, на которой лежит указательный палец.

<sup>3</sup> В ОС Windows также имеется формат гипертекстовой помощи, несовместимый с форматом Stata, поэтому кликание на файлах помощи Stata из Проводника (Explorer) Windows ни к чему не приведет.

## 3.6 Ограничить: условные модификаторы

Многие команды Stata позволяют ограничить свое действие на определенные наблюдения. Делается это с помощью *условных модификаторов* [`if условие`] [`in диапазон`]. Условие, задаемое под `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), `!=` или `~ =` ("не равно"); логические операции `&` ("и"), `|` ("или"), `!` или `~` ("не"), указание на текущее наблюдение `_n` и на последнее `_N`, обычные операции и функции, а также скобки для указания приоритета. `in` указывает диапазон наблюдений вида *начало/конец*, где в качестве конца диапазона может быть использовано последнее наблюдение, обозначаемое латинской "эл" (1) или как `-1`.

Пропущенное значение (см. ниже описание команды `generate`, с. 79) тоже может фигурировать как аргумент логической команды. Так, `count if x!=.` выдаст количество наблюдений, для которых известно значение переменной `x`.

## 3.7 Загрузить, сохранить, объединить: работа с файлами

Естественно, для того, чтобы данные анализировать, их надо как минимум загрузить в память. Stata обладает достаточно гибкими средствами ввода данных из текстовых файлов (команды `infile`; `infix`; `insheet`; см. также `help dictionary` и [U] 24 **Commands to input data**), однако файлы других форматов (Excel, SAS, SPSS, Statistica и т.п.) необходимо предварительно сохранить в виде текста (с разделением данных запятыми, табуляциями, или в фиксированном формате), либо воспользоваться внешними средствами для конвертации данных. В комплект поставки Professional Stata входит чрезвычайно полезная Windows-утилита StatTransfer (<http://www.stattransfer.com>), позволяющая преобразовывать данные между двумя десятками различных форматов. Другая похожая по функциональным возможностям программа — DBMS/COPY.

Работу с уже имеющимися файлами данных формата Stata можно осуществлять из меню File, а можно и с клавиатуры.

```
use имя файла, [clear]
```

Загрузить в память указанный файл. Опция `use ... , clear` показывает, что при этом нужно уничтожить все данные, находящиеся в памяти. Если размер оперативной памяти (точнее, размер свободной памяти, остающейся после Windows и прочих приложений) не позволяет втиснуть в нее необходимый файл, то можно воспользоваться вариантом `use переменные using имя файла [if условие] [in диапазон]`, выбрав модификаторами только те переменные и/или только те наблюдения, которые нужны для работы. С помощью этого трюка можно проводить “черновой” анализ для задач большого объема, т.е. отработать последовательность команд на некоторой подвыборке, сохранить алгоритм работы с данными в виде do-файла (см. параграф 3.13), а потом оставить на выходные машину считать все то же самое по полной выборке.

`save имя файла , [replace old]`

Сохранить данные из памяти на диск под указанным именем. Опция `replace` указывает, что файл надо переписать, если он существует. Если нет — не беда, он будет создан. Опция `old` нужна для сохранения данных из-под Stata 6 в формате Stata 4-5 (т.е. для обмена данными с обладателями Stata более ранних версий). В Stata 7 опция `old` позволяет записать данные в формате Stata 6.

`merge список ключевых переменных using имя файла , [nokeep ]`

Добавить данные из указанного файла к данным, находящимся в памяти. Необходимо для пополнения данных “вширь”, т.е. для добавления *переменных*. Необходимо, чтобы в обоих файлах (которые на жаргоне Stata называются master data и using data) имелись *ключевые переменные*, т.е. переменные, однозначно идентифицирующие наблюдения, а также чтобы файлы были отсортированы по этим переменным, см. [R] `sort` и ниже команду `sort`. Некоторые из этих ограничений преодолеваются командой `mmerge` Wessie (1999), которую необходимо устанавливать дополнительно (см. раздел 3.17). Опция `nokeep` указывает, что не надо добавлять наблюдения, которые встречаются только в using data.

`append using имя файла`

Добавить данные из указанного файла *в длину*, т.е. добавить новые наблюдения.

См. также: [U] **25 Commands for combining data**

## 3.8 Добавить, выбросить, переименовать: работа с данными

В Stata имеется несколько типов данных. Первый уровень разделения — это данные строковые и числовые. Числовые делятся в свою очередь на целые и действительные, а внутри каждого класса есть различия в точности представления; см. [U] **data types**, **help datatypes**.

**generate** [*тип*] *имя переменной* = *выражение* [**if** *условие*] [**in** *диапазон*]

Создать новую переменную, возможно, указанного типа, и присвоить ей значение выражения. Имя переменной в шестой версии Stata может быть длиной до восьми символов, а в седьмой — 32, включать в себя буквы (верхний и нижний регистр различаются), цифры или знак подчеркивания, и должно начинаться с буквы. В *выражение* могут входить числа, переменные, фигурировать арифметические операции, функции (математические, статистические, строковые и пр.), логические условия (которые вычисляются как 1 — истина и 0 — ложь), а также обозначение для пропущенного значения (missing value) — точка (.). Stata устроена так, что пропущенное значение равняется плюс бесконечности. Все операции с пропущенным значением будут давать пропущенное значение (кроме логических операций сравнения). Команда **generate nonmissx=x** создаст новую переменную *nonmiss* типа **byte** (т.е. наименьшего возможного размера), которая будет равна 1, если переменная *x* имеет конечное значение, и 0, если *x* не определена. Об условиях и диапазонах говорится ниже, в разделе 3.6. См. также [U] **14 Language syntax**, [U] **15 Data**, [16] **Functions and expressions**.

**egen** [*тип*] *имя переменной* = *egen-функция(выражение)* [**if** *условие*] [**in** *диапазон*], [**by**(*идентификатор группы*)]

Более мощная функция для создания новых переменных, позволяющая рассчитывать средние, медианы, минимумы, максимумы, суммы значений и т. п. по всей выборке или по группам, задаваемым переменными-идентификторами. Подробный список поддерживаемых функций и статистик имеется в [R] **egen** или **help egen**.

**xi** *специальные выражения*

**xi:** *команда Stata со специальными выражениями*

Позволяет создать набор бинарных (0/1) переменных из категориальной, или выполнить указанную команду, включив в список переменных создаваемый на ходу набор бинарных переменных. Одна из категорий берется как базовая, и для нее бинарная пе-

ременная не создается, т.е. корректно отрабатывается проблема статистической связи между получаемыми бинарными переменными. Специальные выражения имеют вид *i.категорийная переменная*.

**recode**

Изменяет значения переменной. Актуально для перекодировки значений категорийной переменной или для соединения нескольких категорий в одну.

**replace** *имя переменной=выражение* [**if** *условие*] [**in** *диапазон*]

Заменить значения уже существующей переменной.

**rename** *имя переменной новое имя*

Переименовать переменную.

**drop if** *условие* | **in** *диапазон*

Удалить наблюдения, удовлетворяющие указанному условию.

**drop** *переменные*

Удалить указанные переменные.

**list** [*переменные*] [**if** *условие*] [**in** *диапазон*]

Вывести значения указанных переменных (если не указано ничего, то всех) для наблюдений, удовлетворяющих указанному условию (если никаких условий не указано, то вывести все наблюдения).

**edit** [*переменные*] [**if** *условие*] [**in** *диапазон*]

Вручную редактировать указанные переменные для указанных наблюдений. Stata предоставляет для этой цели что-то вроде примитивных электронных таблиц. Использовать подобный режим для внесения изменений в данные не рекомендуется в целях обеспечения воспроизводимости результатов.

**browse** [*переменные*] [**if** *условие*] [**in** *диапазон*]

Просмотреть значения указанных переменных для указанных наблюдений. То же, что и **edit**, только изменять ничего нельзя.

**aorder**

Отсортировать переменные по алфавиту.

**sort** *переменные*

**gsort** +|-*переменная* ...

Отсортировать данные по указанным переменным.

**compress** [*переменные*]

Привести переменные (если не указано, какие, то все) к минимально возможному типу, снижая тем самым объем памяти, необходимый для их хранения.

## reshape

Достаточно продвинутая команда, необходимая для изменения представления группированных данных — например, панельных. Она переводит данные из “длинного” (long) формата, в котором на каждый объект панели имеется несколько наблюдений, соответствующих разным моментам времени (много наблюдений, откладываемых “в длину”— мало переменных, откладываемых “в ширину”), в “широкий” формат (мало наблюдений — много переменных), в котором наблюдения соответствуют объектам, а данные записаны в виде переменных, названия которых заканчиваются на “дату”. Так, файл с переменными `income96`, `income97`, `income98` — это данные в широком формате, а файл с переменными `income`, `year`, где `year` принимает значения 96, 97, 98 — это данные в длинном формате. Панельные команды Stata, имеющие префикс `xt`, работают с данными в “длинном” формате.

**describe** [*переменные*] [*using имя файла*], [**short**]

Вывести описание данных и переменных: формат, метки и т. п. Эта команда показывает также количество наблюдений и переменных, изменялись ли данные с момента последнего сохранения, по каким переменным отсортированы наблюдения. Можно указать файл, находящийся на жестком диске.

## label

Приписать метки к данным или переменным. `label variable имя переменной "текст"` создает метку переменной, которая выводится командой `describe` и видна в окне переменных. Можно также задать метку для файла данных (информация о файле данных хранится в сопровождающем его объекте `_dta`, и соответствующая команда будет иметь вид `label data`). Эта метка будет выводиться при исполнении `use` и `describe`. Можно также задать метки для отдельных значений дискретной переменной через `label define` и `label values`. Признаком хорошего стиля работы с данными является придание меток создаваемым переменным: после любой команды `generate` или `egen` должно идти `label variable`.

**notes** [`_dta` | *переменная*] : " *текст*"

Еще один вариант создания примечаний о файле или переменных. Если командой `label` всем данным в целом или отдельной переменной можно приписать только одну метку ограниченной длины, то `notes` позволяет приписать к каждой переменной или к `_dta` произвольное число меток произвольной длины. Примечания удобны для внесения комментариев типа: “Разобраться с этой переменной”; “Данные за 1994 г. сверены”; “Файл получен программой `households.do`” и т.п.



`lookfor` *текст*

Ищет указанный текст в названиях и метках переменных.

`clear`

Очистить память, выгрузив все данные, метки, программы, макросы.

## 3.9 Оценить: основные статистические средства

`summarize` *переменные* [*if условие*] [*in диапазон*], [`detail` ]

Сводка описательных статистик, таких, как количество наблюдений, среднее, стандартное отклонение, максимум, минимум. Опция `detail` позволяет вывести также характерные квантили, несколько самых больших и самых маленьких значений и коэффициенты асимметрии и эксцесса. Прочие команды, описывающие данные в компактном виде — `lv`; `codebook` и `inspect`. Для дискретных переменных, принимающих небольшое число значений, будут полезны команды табуляции `tabulate` или `table` — см. ниже.

`correlate` *переменные* [*if условие*] [*in диапазон*], [`covariance` ]

Выводит матрицу корреляций между переменными в указанном диапазоне. Опция `covariance` указывает, что надо вывести ковариационную матрицу. Матрицы вычисляются по тем наблюдениям, для которых имеются значения *всех* указанных переменных.

`pwcorr` *переменные* [*if условие*] [*in диапазон*], `sig` `obs`

Выводит матрицу *попарных* корреляций, т. е. корреляций, рассчитанных по наблюдениям, в которых значения соответствующих переменных попарно не пропущены. Опция `sig` выводит уровень значимости корреляции (в предположении совместной нормальности), а `obs` — количество наблюдений.

`tabulate` *переменные* и `table` *переменные*

Построение различных таблиц, содержащих агрегированную информацию по переменным. Поддерживаются метки переменных и отдельных значений. Введение в эти команды дается уроком `tutorial tables`. См. также [U] **28 Commands for dealing with categorical variables**

`regress` *зависимая переменная объясняющие переменные* [*if условие*] [*in диапазон*], `robust`  
`noconst` `cluster`( *групповая переменная* )

Оценивание линейной регрессии зависимой переменной на объясняющие. Выводятся основные результаты оценивания: количество наблюдений, таблица дисперсионного анализа, статистики  $F, R^2, R^2_{adj}$ , а также таблица оценок коэффициентов, стандартных отклонений оценок,  $t$ -статистик и доверительных интервалов (см. с. 55 с примером регрессии). Опция **robust** задает оценку ковариационной матрицы оценок коэффициентов в форме Вайта (2.29), учитывающей гетероскедастичность. Опция **cluster** указывает, что ковариационная матрица должна учитывать группировку наблюдений (как в стратифицированных обследованиях). Опция **noconst** указывает, что в модель, оцениваемую Stata, не следует включать константу (как это делается по умолчанию). После команды **regress** можно получать прогнозные значения, остатки и строить диагностические переменные командой **predict** или проводить диагностику регрессии, не прогоняя регрессию заново. Введение в эту команду предоставляется уроком **tutorial regress**.

Команды оценивания статистических моделей в Stata имеют много общего. В частности, после всех таких команд можно отдавать команду **predict**, которая будет строить значения тех или иных выражений, связанных с результатами оценивания; получать матрицы самих оценок параметров (матрица-столбец  $e(b)$ ) и их ковариационную матрицу ( $e(V)$ ); строить тесты на линейные (**test**) и нелинейные (**testnl**, с использованием дельта-метода для получения ковариационной матрицы нелинейных функций оценок) комбинации параметров, и т.д. Можно вывести результаты оценивания, не показанные в основном блоке вывода, командой **estimates list**. Отдельные коэффициенты можно получать в виде **\_b[имя переменной]**, а их стандартные ошибки — **\_se[имя переменной]**. Специфика команд, оценивающих параметрические модели, описана в разделах **help est** и **help postest** встроенной подсказки Stata.

В пакете Stata имеется широчайший спектр статистических команд, важных для эконометрического анализа<sup>4</sup>:

- регрессия с инструментальными переменными **ivreg**, робастная регрессия **rreg**, одновременные уравнения **reg3**, нелинейный МНК **nl**;
- модели временных рядов (**help time**): модели авторегрессии со скользящим средним **arima**; автокорреляции **ac** и частные автокорреляции **pac**; модели с условной гетероскедастичностью **arch**; регрессия с ковариационной матрицей Ньюи-

---

<sup>4</sup> Подробности см. **help название команды**.

Веста (2.31) `newey`; проверка гипотез о единичных корнях временного ряда `dfuller`; `pperron`;

- обобщенные линейные модели (`glm`);
- средства дисперсионного анализа (`anova`; `oneway`; `loneway`),
- средства факторного анализа и анализа главных компонент (`factor`);
- средства анализа таблиц сопряженности (более подробные опции команд `table`; `tabulate`; `epitab`);
- средства анализа панельных моделей (команды с префиксом `xt`, например, `xtreg`, `re` и `xtreg, fe` — регрессии со случайным и фиксированным панельным эффектом; `xtgls` — регрессии с коррелированными остатками; `xtlogit` и `xtprobit` — панельные регрессии с бинарной зависимой переменной. Подробности см. `help xt`, а также [U] **29.13 Panel-data models**);
- средства анализа данных типа длительностей, или времени жизни, или времени отказа (`survival time`; команды с префиксом `st`; см. `help st`, а также [U] **29.14 Survival-time (failure time) models**);
- средства анализа стратифицированных обследований (`survey`; команды с префиксом `svy`; см. `help svy`, а также [U] **30 Overview of survey estimation**);
- средства анализа моделей с бинарной зависимой переменной (`logit`; `logistic`; `lfit`; `probit`; `dprobit` — предельные эффекты в пробит-модели);
- тесты на равенство средних (`ttest`), дисперсий (`sdtest`) и медиан (ранговые и знаковые тесты `signrank`; `signtest`; `ranksum`; `kwallis`);
- ранговые корреляции (`spearman`; `ktau`);
- возможность максимизации функций правдоподобия, запрограммированных пользователем (`ml`);
- в Stata 7 — исчерпывающий набор средств кластерного анализа;
- и многое, многое другое.

Полная стандартная поставка пакета Stata насчитывает около 500 команд для конечного пользователя (плюс большое число внутренних или программистских модулей). Примерно столько же содержится в официальных дополнениях (STB), и еще около полутысячи команд (по состоянию на конец 2000 г.) находится в интернетовском архиве SSC-IDEAS (см. раздел 3.16).

## 3.10 Посчитать: функции

В пакете Stata реализовано довольно большое число различных функций: математических (логарифмы, тригонометрические функции, модуль, корень и т. п.); статистические (плотности и функции распределения; в седьмой версии они приведены в некоторый порядок по сравнению с предыдущими), генератор псевдослучайных чисел (с периодом  $\approx 2^{126}$ ,  $2^{32}$  различными значениями и с возможностью инициализации пользователем для воспроизводимости вычислительных экспериментов), строковые функции, функции для работ с датами, функции от матриц (определитель, обратная), константа  $\pi$  (`_pi`) и ряд других. Полный список можно получить через `help functions` или [U] **16.3 Functions**, [R] **functions**. О возможностях написания пользовательских функций см. раздел 3.17.

## 3.11 Повторить: циклы

Stata обладает довольно своеобразными средствами повтора тех или иных команд для разных групп наблюдений, разных переменных и прочих случаев. Обычно этих средств хватает для выполнения требуемых задач, но иногда приходится прибегать к более изощренным трюкам.

*by идентификатор(ы) групп* : команда Stata

Эта команда повторяет указанную команду Stata отдельно для каждого набора наблюдений с одинаковыми значениями групповых переменных. Иными словами, Stata разбивает все данные на отдельные группы согласно групповым переменным (идентификаторам), и выполняет указанную команду для каждой из групп по отдельности. При этом указатель последнего наблюдения `_N` указывает на последнее наблюдение в группе. Необходимо, чтобы данные были отсортированы по этим групповым перемен-

ным, в противном случае Stata выдаст сообщение об ошибке.

`for min списка список [: for min списка список ... ]`: команда Stata с символом X [Y] [  
\ команда Stata с символом X [Y] ... ]

Поддерживаемые типы списков: список чисел (numlist), список переменных (varlist), произвольный список (anylist).

Числа от 1 до 10 можно задать следующими способами: 1(1)10, или 1 2 to 10, или 1/10.

В списке переменных можно использовать переменные, стоящие друг за другом, через тире. Можно использовать \* как заменитель любого символа: `u*` означает все переменные, начинающиеся на "u".

Подробнее о списках: [U] 14 **Language syntax**, `help numlist`, `help varlist`.

Команда `for` осуществляет цикл в обычном алгоритмическом понимании этого слова. Она перебирает элементы списка и подставляет их вместо X в исполняемой команде (командах). Если задано больше одного `for` через двоеточие, то Stata выполнит указанные команды для всех сочетаний X из первого списка × Y из второго, и т.д.

Stata честно пытается информировать пользователя о каждом отдельном значении групповых переменных `by` или параметра `for`, для которого выполняется очередное действие, и если список составляет несколько сотен, то весь процесс может оказаться несколько утомителен, да и вывод на экран иногда является самым медленным элементом вычислительного конвейера Stata. Чтобы команды ничего не выводили на экран, перед `for` и `by` можно задать, как и перед любой из команд Stata, префикс quietly, например: `qui for var x1-x5: g lX=log(X) \ lab var lX "log of X"`

В седьмой версии пакета возможности циклов несколько расширены командами `forvalues` и `foreach`.

## 3.12 Запомнить: результаты работы

Естественно, результаты работы по статистическому анализу данных не должны погибать вместе с концом сеанса Stata. Можно копировать эти результаты непосредственно из окна результатов Stata и через буфер обмена переносить в прочие приложения, однако есть более естественный способ.

`log using имя файла, [ append | replace ]`

`log on | off | close`

Эта команда записывает все, что Stata выводит в окно результатов, в указанный файл (добавляя либо перезаписывая этот файл, в соответствии с опциями `append` либо `replace`, если такой файл существует). `log off` временно прекращает запись в файл, `log on` возобновляет запись в файл, `log close` прекращает запись и закрывает файл. Команды, связанные с log-файлом, продублированы на панели инструментов Stata кнопкой со светофором. Log-файлы лучше всего печатать непосредственно из Stata, поскольку Stata умеет автоматически приукрашивать текст (выделяя полужирным шрифтом команды, проставляя даты и т.п.).

В Stata 7 есть два вида log-файлов: командный (в который пишутся только команды, отдаваемые пользователем, что дает возможность быстро конвертировать результаты работы в программу) и полный (в который пишутся как команды, так и результаты их исполнения). Запись команд в командный log-файл задается конструкцией `cmdlog using имя файла`. Есть также недокументированные способы записи log-файлов в форматах HTML и `texman` — `log html имя файла` и `log texman имя файла`.

Есть еще один вариант сохранения статистических результатов исследований — прекрасная пользовательская команда `outreg` Gallup (2001), которая записывает результаты регрессий в отдельный текстовый файл в соответствии с принятыми в статистической и эконометрической литературе обозначениями: столбцы коэффициентов со стандартными ошибками в скобках, число наблюдений, статистика  $R^2$  и прочие статистики. Этот модуль требует, впрочем, отдельной установки, см. [R] `stb`, `help stb`. Самую свежую версию можно найти на сайте архива SSC-IDEAS, см. параграф 3.16.

Наконец, список нескольких последних команд можно получить командой `#review [количество команд]` .

См. также: [U] **Printing and preserving output**.

### 3.13 Запустить: do-файлы

Произвольную последовательность команд Stata можно записать в отдельный файл — не более одной команды в строке — и выполнить всю последовательность одной командой. Традиционно файлы, в которых записаны подготовленные таким образом программы, носят расширение `.do`, а команда, выполняющая эти do-файлы, так и называется:

`do имя файла аргументы , [nostop ]`

Stata прекращает исполнение do-файла, когда натывается на ошибку. Можно этого избежать, установив опцию `nostop`.

Если не требуется вывод на экран, то вместо `do` можно запустить программу командой `run`. Впрочем, в случае аварийного останова Stata все равно выдаст сообщение об ошибке, вполне справедливо полагая, что пользователь должен об этом знать.

В текст do-файла можно вставлять комментарии, оформляемые в стиле языка программирования C, т. е. `/*` открывает комментарий, а `*/` — закрывает. Кроме того, строка, начинающаяся со звездочки `*`, также считается комментарием и полностью игнорируется. Эта строка, тем не менее, является командой, в том смысле, что Stata выводит ее в окно вывода и в log-файл. Можно таким образом вводить комментарии и при интерактивной работе.

Когда возможностей `for` не хватает, можно попробовать написать отдельный do-файл для выполнения требуемых действий и передавать ему `X` (или каким-то образом преобразованное выражение с `X`) в качестве одного из аргументов.

Можно дать несколько советов по созданию do-файлов <sup>5</sup>.

- Для того, чтобы гарантировать воспроизводимость всех результатов, необходимо оформлять все полезные действия, вплоть до изменения значения одной переменной в одном наблюдении, как строки do-файла. Автору этих строк неоднократно приходилось выяснять вместе с коллегами, почему у них получаются разные результаты при использовании вроде бы одних и тех же методов обработки и анализа и вроде бы одних и тех же файлов данных, и именно для исключения подобных ситуаций и разработаны эти советы.
- На каждый отдельный исследовательский проект надо заводить отдельный каталог, а исходные данные сохранять неизменными и соответствующими исходным статистическим первоисточникам (справочникам, известным базам данных и т.п.), вынося все необходимые поправки и изменения в do-файлы. После существенных изменений в данных (таких, как команды `reshape`, `merge` или создания большого количества новых переменных — не забывайте придавать им мет-

---

<sup>5</sup>С разрешения Stata Corp., по материалам Net Course 151 по программированию в пакете Stata.

ки!) стоит сохранить полученные (промежуточные) результаты. Название файла должно отражать его происхождение или содержание, либо же должно содержать признаки того, что данные вторичны (например, можно начинать названия несущественных файлов с тильды или подчеркивания). Более подробную информацию о происхождении файлов данных можно записывать в эти файлы командами `label data` и `notes`.

- Стоит каждый do-файл начинать "с нуля", а log-файл, отслеживающий происходящее, должен открываться в том же do-файле. Отслеживать результаты работы будет проще всего, если имя log-файла совпадает с именем do-файла (или, если do-файл должен использовать какие-то параметры, эти параметры также должны фигурировать в названии log-файла, благо Stata поддерживает длинные имена Windows).

Примерная "рыба" do-файла, который называется, скажем, `income98.do`, будет такова:

```
clear
version 6
set memory 10m
log using income98, replace
use income98
* еще какие-то действия
...
log close
exit
```

Stata Corporation предлагает превосходные Internet-курсы по программированию в пакете Stata. Автор этого пособия участвовал в таких курсах и считает, что они заметно помогли ему в освоении возможностей пакета.

См. также: [U] **19 Do-files**



## 3.14 Нарисовать: графика

Мир графических средств пакета Stata начинается командой `graph`, у которой имеется добрая сотня разнообразных опций на разнообразные случаи жизни. Наиболее часто используемые графики реализованы в виде отдельных команд.

`graph` *переменные*, [*опции*]

Команда `graph` одна, но вариантов воплощения у нее очень много. Краткий рассказ об этих возможностях дается уроком `tutorial graphics`.

Если команда `graph` содержит одну переменную, то эта команда интерпретируется как задание построить гистограмму. По умолчанию Stata разбивает диапазон изменения переменной на пять интервалов (bins), что, как правило, недостаточно информативно, поэтому имеет смысл увеличить число интервалов опцией `graph ... , bin(50)`. Можно наложить поверх гистограммы плотность нормального распределения с аналогичным средним и дисперсией для визуального контроля нормальности с помощью опции `graph ... , norm`. Еще несколько разновидностей графиков, описывающих одну переменную, даются опциями `graph ... , box` (график box-whisker, отражающий основные квантили распределения<sup>6</sup>) | `star` (роза ветров) | `bar` (столбцовая диаграмма) | `pie` (круговая диаграмма). Более подробную помощь можно найти по ключевым словам `grhist` и `graph`.

Диаграмма рассеяния выводится командой `graph`, но с двумя аргументами: `graph` “ось y” “ось x”. Из основных опций (перечисляемых через запятую в командной строке), которые имеет смысл указывать для диаграммы рассеяния, стоит упомянуть:

- `symbol` — символ, которым будет помечаться выводимое наблюдение; `symbol(.)` выведет маленькую точку, `symbol(o)` — маленький кружок, `symbol([переменная])` — значение указанной переменной; `symbol([_n])` — номер наблюдения.
- `connect` — соединение точек; `connect(.)` означает, что точки соединять *не* надо, `connect(1)` — что точки надо соединить тонкой линией; `connect(s)` — провести сплайн через соседние точки. Сплайн является одним из видов *непараметриче-*

---

<sup>6</sup>На таком графике ящик (box) ограничен верхним и нижним квартилями, средняя линия ящика проводится на уровне медианы, а усы (whiskers) — это удвоенные разности между медианой и квартилями.

ской регрессии (см. параграф 2.6.5). В седьмой версии пакета можно задать стили линий, указывая их в квадратных скобках после символа, задающего соединение точек: `connect(1[-])` — пунктирная линия, `connect(1[_])` — длинная пунктирная линия, `connect(1[.])` — короткая пунктирная линия. Эти стили можно сочетать — `connect(1[-.])` выдаст штрих-пунктирную линию.

- `sort` — перед соединением точек, задаваемой опцией `connect`, отсортировать наблюдения по переменной на оси  $x$  (во избежание заполнения экрана паутинообразной ломаной).
- `bands` — количество соседних точек, используемых для вычисления сплайна. Чем ниже число, задаваемое этой опцией, тем более гладкой будет кривая непараметрической регрессии.
- `density` — количество точек на графике сплайна. Чем больше это число, тем более гладким будет изображение сплайна. Гладкость самого сплайна регулируется опцией `bands`.
- `xlab` и `ylab` — числовые метки на осях.
- `xtick` и `ytick` — “зарубки” на осях.
- `xline` и `yline` — вертикальные и горизонтальные линии на графике.
- `xscale` и `yscale` — диапазон осей.
- `title` — заглавие графика. В данном контексте Stata не понимает русский текст.

Эти и другие опции описываются в подсказке `grtwoway`.

Если в команде `graph` указать более двух переменных, то Stata построит графики зависимости всех переменных от последней, т.е. список переменных интерпретируется как  $y_1, \dots, y_{n-1}, x$ . Матрица попарных диаграмм рассеяния выводится с помощью опции `graph, matrix`.

Графики Stata можно сохранять в собственном формате `.gph`, указывая после любой графической команды опцию `graph ... , saving(имя файла)`. Эти сохраненные графики можно потом просмотреть заново командой `graph using имя файла(ов)`. Stata позволяет сочетать на одном рисунке несколько графиков — см. подсказку по команде

`help grother`. Кроме того, через меню **File** можно сохранять графику и в виде, понятном Windows-приложениям (в виде растровой графики `.bmp` или векторной `.wmf`), или переносить в другие приложения через буфер Windows.

Для встраивания графики Stata 6 в  $\LaTeX$  надо приложить определенные старания. Общее направление действий — сохранить в формате PostScript (`.ps`) → конвертировать в Encapsulated PostScript (`.eps`) или PDF → экспортировать в  $\LaTeX$  средствами пакета `graphicx`. В UNIX-овых версиях Stata есть утилиты `gphpen` и `gphprint`, которые позволяют сохранить графику в виде PostScript-файла. Пользователям Windows, видимо, необходимо установить PostScript-принтер и получать PostScript-файлы, посылая графику на печать и отмечая опцию "Печатать в файл". В качестве PS-принтера можно использовать и Acrobat Distiller из пакета Adobe Acrobat (*не* Acrobat Reader). Возможно, впрочем, что полученный таким образом файл будет содержать специфические команды принтера, поэтому его предварительно необходимо пропустить через конвертер, который оставил бы только "ортодоксальные" ключевые слова PostScript. Если используется пакет, создающий из  $\LaTeX$ овского файла PDF-файл (например, `pdflatex` пакета `fpTeX`), то можно из Stata отправлять графические файлы на печать в Acrobat Distiller и получать PDF-файлы с рисунками, которые потом, возможно, надо будет обрезать в Adobe Acrobat.

Все эти процессы упрощены в Stata 7 и делаются одной командой `translate`, которая конвертирует графику в форматы PostScript и Encapsulated PostScript.

См. также: [G]

### 3.15 Уточнить: команды для удовлетворения любопытства

В данном разделе будет рассказано о командах, показывающих определенные параметры состояния Stata, и о случаях, когда бывает полезна представляемая ими информация.

#### query

Выводит установки текущих параметров (в т. ч. размер матрицы, см. выше `set matsize`, уровень значимости по умолчанию статистических тестов `level`, в %, имя текущего log-файла, и т. п.). Эти установки можно менять с помощью команды `set`,

примеры которой приводились в разделе 3.2.

**about**

Выводит основные параметры Stata и компьютера, на котором работает пакет: версия программы, дата создания exe-файла, общий и доступный объем памяти.

**memory**

Выводит информацию о том, сколько памяти отведено для Stata и как она используется. Рекомендуется иметь памяти по меньшей мере на 15–20 % больше, чем требуется для данных, поскольку очень многие команды создают временные переменные, временные матрицы или используют память иным образом.

**adopath**

Выводит информацию о том, в каких каталогах Stata ищет ado-файлы с новыми программами (см. с. 72 об ado-файлах). Необходимо для установки новых компонент Stata (например, STB-дополнений при их ручном скачивании из Internet, см. параграф 3.17), а также при написании собственных программ в виде ado-файлов.

**which** *название команды*

Выводит информацию о том, в каком файле и в каком каталоге найден ado-файл, выполняющий требуемую команду, а также информацию о версии команды. Может оказаться полезным, если программа дорабатывается автором и необходимо отслеживать наличие последних версий, а также при появлении сообщений об ошибках для обращений в службу технической поддержки Stata или к автору программы.

## 3.16 Законnectиться: Internet-возможности Stata

Адрес Stata в Интернете — <http://www.stata.com/>. На этом корпоративном сайте размещаются новости (выход обновлений и новых версий, дополнений к Stata — STB, встреч пользовательских групп, объявления об Интернет-курсах по программированию и использованию пакета). Еще один очень полезный адрес — <http://ideas.uqam.ca/>. Здесь располагается поисковая система архива RePEc (Research Papers in Economics), умение пользоваться которой само по себе полезно для всякого экономиста. Одной из составных частей RePEc является архив программ SSC-IDEAS (Statistical Software Components), написанных пользователями Stata. В этом архиве содержится несколько сотен различных программных модулей, что вполне сопоставимо с количеством команд в

минимальном варианте установки. Из прочих ресурсов стоит упомянуть лист поддержки — [statalist@hsphsun2.harvard.edu](mailto:statalist@hsphsun2.harvard.edu)<sup>7</sup>, на котором можно получить квалифицированную помощь как от других пользователей Stata, так и от самих разработчиков, вплоть до президента корпорации Уильяма Гулда (William Gould). По его словам, оперативная и персональная поддержка пользователей — это один из важнейших приоритетов компании. От себя добавлю — это еще и одно из самых больших ее достоинств, особенно по сравнению с огромными монстрами типа SAS.

Начиная с шестой версии, Stata обладает рядом полезных возможностей, реализуемых через всемирную сеть Интернет. Это — обновление пакета, а также доступ к пользовательским программам.

#### **update**

Позволяет загрузить официальные обновления Stata через Интернет. Запрос **update query** показывает, что нужно обновить (статистические компоненты, находящиеся в ado-файлах, или исполняемый файл `wstata.exe`). Затем можно обновить необходимые фрагменты с помощью **update ado**, **update executable** или **update all**.

#### **net [from URL]**

Установка программ Stata через Internet. Эта команда ищет по указываемым Интернет-адресам (URL) — или, по умолчанию, на вышеуказанном сайте Stata — описания пакетов, которые может установить пользователь, скачивает необходимые файлы и устанавливает их на вашем компьютере.

#### **webseek** *ключевые слова*

Осуществляет поиск в Internet команд Stata, соответствующих указанным ключевым словами. **webseek** обращается на сервер Stata, на котором содержится информация о программах STB и других архивах программ Stata, по которым и осуществляется рекурсивный поиск. В седьмой версии команда **webseek** заменена на **net search**.

Помимо этих команд, работающих через Internet, Stata может выполнять многие действия, связанные с файлами, используя URL файлов вместо их имен. Так, вполне осмысленная команда

```
use http://www.stata.com/users/vwiggins/auto.dta
```

загрузит ценный файл `auto.dta`, на тот случай, если вы случайно стерли оригинал,

---

<sup>7</sup> Чтобы подписаться на этот лист, надо послать письмо на адрес [majordomo@hsphsun2.harvard.edu](mailto:majordomo@hsphsun2.harvard.edu) с текстом `subscribe stataлист`.

поставляющийся вместе с пакетом. Можно получать через Интернет текстовые файлы с данными и конвертировать их в файлы Stata командами `infile`, `infix`, `insheet`, и т.п.

Для корректной работы через прокси-сервер необходимо установить его параметры в меню `Prefs/General Preferences/Internet Prefs`.

См. также: [U] **32 Using Internet to keep up to date.**

### 3.17 Надстроить: расширение возможностей Stata

Stata — динамичный и открытый пакет. От одного до трех раз в месяц Stata выпускает обновления на уровне ado-файлов, доступные по команде `update`, и примерно раз в квартал выходят обновления исполняемого файла. Однако основная динамика происходит на листе `statalist` и на архиве программ SSC-IDEAS, где за день может появиться с десятков новых команд (написанных пользователями Stata для решения своих исследовательских задач, либо в качестве ответа на вопросы, задаваемые на `statalist`).

Stata публикует официальные дополнения к пакету под названием `Stata Technical Bulletin` (или, сокращенно, `STB`). Они выходят раз в два месяца, и состоят из программ (распространяемых бесплатно через Интернет) и небольшой брошюры, похожей по формату на руководства пользователя Stata. Для того, чтобы установить у себя программы из этих бюллетеней, надо отдать команды

```
net
net cd stb
```

или обратиться к меню `Help/STB and User-written Programs` для доступа к ado- и hlp-файлам на сервере Stata.

В предыдущих версиях Stata всех этих возможностей работы через Интернет не было, поэтому для установки STB-дополнений или программ из архива SSC-IDEAS было необходимо скачивать их вручную с Интернета, а потом либо копировать в каталог, зарегистрированный в `adopath` (см. с. 93), либо устанавливать средствами Stata — командой `install`. Пользователи 6-й или 7-й версии, у которых нет постоянного или хотя бы модемного соединения с Интернетом, будут вынуждены ходить с дискетами к знакомым, у которых доступ есть, скачивать необходимые команды на дискету, а потом устанавливать их командой `install from a:.`

Есть еще один технический момент, связанный с представлением текстовых файлов в Windows и UNIX. В этих двух операционных системах концы строк представляются по-разному, причем UNIX понимает тексты Windows, но не наоборот. В архиве SSC-IDEAS находятся программы, написанные в обоих форматах. При копировании командой `net Stata` корректно обрабатывает концы строк, однако при описанном выше “ручном” копировании возможны проблемы у пользователей, работающих в Windows. Симптомом того, что у вас возникла проблема, связанная с концами строк, является неработоспособность свежеставленных файлов — Stata возвращает ошибку с кодом 199 (`unrecognized command: xyz not defined by xyz.ado` — команда не распознана; программа `xyz` не определена в файле `xyz.ado`); при этом Stata находит файл помощи на новую команду, но в нем все оказывается перепутано. Эту проблему можно решить, открыв оба файла (`.ado` и `.hlp`) в текстовом редакторе и сохранив их обратно — есть вероятность, что концы строк при этом будут расставлены заново.

На определенном уровне владения пакетом оказывается удобным писать по разным случаям свои собственные программы (ado-файлы). Их можно публиковать их в Интернете для всеобщего доступа. Например, страничка автора этой книги, посвященная Stata, размещается по адресу: <http://www.komkon.org/~tacik/stata>. На ней находятся программы, уроки (tutorials) и PDF-файл с этой книгой.

Частным случаем пользовательских программ являются функции для команды `egen`. Они позволяют в какой-то степени обойти невозможность написания функций пользователя, применимых наравне со встроенными. Файлы, в которых содержатся такие функции, имеют префикс `_g` и должны быть написаны в соответствии с определенными требованиями на обработку входных аргументов.

## 3.18 Научиться на опыте: сообщения об ошибках

В соответствии с общепринятыми программистскими соглашениями, каждая команда и программа должна уметь сообщать о результатах своей работы. Чаще всего это делается в виде целочисленного кода завершения программы. Нулевое значение этого кода свидетельствует об отсутствии каких-либо ошибок и проблем при выполнении задания;

ненулевое, как правило, обозначает те или иные ошибки. Помимо кода завершения, многие программы Stata сохраняют те или иные результаты своей работы, которые можно получить, в зависимости от выполненной команды, через `estimates list` (вспомним обсуждение команды `regress` на с. 83) или `results list`. См. `help estimates` и `help results`.

В окне вывода Stata текст подсвечивается одним из пяти цветов: белым, желтым, зеленым, голубым или красным. Белым цветом показываються команды, отданные пользователем, а также некоторые специфические сообщения; голубым — команды, записанные в do-файле, а также запрос на продолжение вывода `--more--` (пауза в процессе обработки данных; для получения следующей строки вывода надо нажать **Enter**, следующей страницы — клавишу "пробел", как в программе `more` ОС UNIX); зеленым — информационный (постоянный) текст; желтым — рассчитываемые числовые значения (переменный текст); красным — сообщения об ошибках. Сообщения об ошибках сопровождаются кодом ошибки, по которым можно найти более подробную информацию в [R] **error messages** или через меню `Help/Search/rc код ошибки`. Чаще всего ошибки вызваны неправильным синтаксисом вводимых пользователем команд (использованием одинарного = в условиях `if`, ссылкой на несуществующую переменную из-за опечатки в названии переменной, ссылкой на несуществующую команду при опечатке в названии команды, попыткой создать вновь уже существующую переменную, и т.п.). Иногда, впрочем, ситуации могут быть более серьезными и свидетельствовать о статистических или вычислительных проблемах — например, когда не достигается сходимость итерационных процессов или не хватает наблюдений для оценивания модели — или проблемах компьютерных — нехватке памяти (сообщение `no room to add more variables`, см. выше `set memory`).

В пакете Stata 7 функции голубого цвета несколько изменены: он означает ссылку на файл встроенной подсказки, на URL в Интернете или просто на команду Stata. Можно навести мышку на фрагмент, показанный голубым цветом, и по нажатию левой кнопки мыши Stata покажет необходимый файл помощи, запустит браузер или выполнит необходимую команду. В частности, коды ошибок показываються голубым цветом, и при кликании на коде ошибки показывается файл подсказки, поясняющий, почему возникла данная ошибка.

См. также: [U] **11 Error messages and return codes**



## 3.19 Разобраться: прочее

В этом разделе приведены сведения, которые пригодятся уже при достаточно серьезном уровне владения пакетом и достаточно серьезных запросах к сложности программ.

### Матрицы

Пакет Stata не является матричным, как, например, GAUSS. В нем, однако, реализовано большинство популярных матричных задач и алгоритмов: основные алгебраические действия, обращение, разложение Холецкого, решение задачи на собственные значения, сингулярное разложение. Столбцы и строки матриц можно “называть по именам” (что вполне естественно, например, для ковариационных матриц, возникающих при оценивании параметров статистических моделей). Знакомство с матричными средствами Stata можно начать с `help matrix`.

См. также: [U] **17 Matrix expressions**

### Макросы

Наиболее близким к понятию макросов Stata является, пожалуй, понятие локальной переменной в программировании. Макросы — это строки, имеющие содержанием другие строки (в т.ч. числовые значения, записанные в экспоненциальном формате). С их помощью в программах Stata можно устраивать циклы, получать передаваемые подпрограмме значения, и т.п. Макросы делятся на локальные, которые будут забыты по окончании того процесса, который их создал<sup>8</sup>, и глобальные, доступные всем программам Stata. Среди глобальных макросов есть ряд зарезервированных, описывающих состояние Stata (версия, дата, время, режим работы, пути для поиска ado-файлов и т.п.). Ссылки на глобальные макросы Stata начинаются со знака доллара (\$). Так, уровень значимости, по умолчанию используемый для построения доверительных интервалов, обозначается как `$$_level` и равен по умолчанию 95 (в процентах).

См. также: [U] **21.3 Macros**

---

<sup>8</sup> Собственно, эти макросы и являются аналогами локальных переменных.

## Русификация

К сожалению российских пользователей, пакет Stata не русифицирован в том смысле, что у него отсутствуют русские описания. Теоретически и технически, русификация выводимых результатов и встроенной подсказки возможна, но объем работы измеряется, как мне кажется, несколькими человеко-годами, так что всерьез на это рассчитывать пока что не приходится.

Тем не менее, Stata может оперировать нелатинскими символами в качестве строк. Русские буквы можно использовать в качестве содержимого строковых переменных, для примечаний, меток переменных и данных, однако нельзя использовать в названиях переменных. Чтобы эти буквы отображались, надо в соответствующем окне (в первую очередь, в окне результатов) установить русские шрифты. Для этого надо ткнуть мышкой в пиктограмму окна в левом верхнем углу нужного окна (см. рис. 3.3) и установить какой-нибудь из русских шрифтов.

### 3.20 С чего начать?

Самое трудное — начать работу с пакетом в первые минуты, и это верно для любого программного средства. Один из важнейших навыков, которым необходимо овладеть с самого начала — это умение пользоваться встроенной подсказкой (см. раздел 3.5, а также подсказки по ключевым словам `help`, `winhelp`).

Другой хороший вариант самообучения и начала эффективной работы — воспользоваться встроенными мини-уроками `tutorials`. Достаточно набрать `tutorial` в командной строке Stata — и дальше Stata сама расскажет, какие мини-уроки у нее есть и как их вызвать. Первый мини-урок вызывается командой `tutorial intro`, и именно с него мы начинали наши практические занятия с пакетом Stata. Цель этих мини-уроков — не решить какую бы то ни было статистическую задачу, а показать, как работают те или иные команды в практической работе, поэтому при просмотре этих уроков надо обращать внимание не на то, что *выводит* Stata, а того, что в нее **вводится**.

Для данного курса прикладной эконометрики автором этого пособия была написана обучающая программа, демонстрирующая основные средства диагностики регрессий. Эта программа доступна со страницы <http://www.komkon.org/~tacik/stata> или, пользуясь интернет-возможностями Stata (см. раздел 3.16), из самого пакета:

```
. net from http://www.komkon.org/~tacik/stata  
. net get aboutreg
```

Темпы обучения, безусловно, индивидуальны, однако обычно уже нескольких часов достаточно для того, чтобы начать самому вводить команды и понимать, что они означают. Для профессионального овладения пакетом нужны, наверно, недели и месяцы постоянной работы с разными задачами и разными данными, отлаживание собственных программ и попытки разобраться в чужих, участие в интернет-курсах по пакету, предлагаемых разработчиками Stata Corp., участие в листе рассылки. Никакая книжка не может заменить самостоятельного активного освоения!

## Глава 4

# Мониторинг экономического положения и здоровья населения России

В этой главе будет кратко описана имеющаяся в открытом доступе (и потому популярная среди исследователей-экономистов) база данных RLMS (Russia Longitudinal Monitoring Survey, Мониторинг экономического положения и здоровья населения России) Mroz *et. al* (1999), Swafford (1996). Это — панельное обследование, проводимое совместно Университетом Северной Каролины (Чапел-Хилл), компанией “Парагон”, Институтом Социологии РАН, Институтом Питания РАН и, на отдельных этапах, другими организациями. В мае 2001 г. были опубликованы данные девятого раунда. Первые четыре раунда проводились в 1992–1993 гг., и на настоящий момент признаются организаторами обследования неудачными. В 1994 г. выборка была создана заново, и с тех пор обследования проводятся регулярно в конце осени (за исключением 1997 г., когда проект не был профинансирован). Файлы данных<sup>1</sup> выложены на ftp-сервер университета, координаты которого (как и многое другое о RLMS) можно найти по адресу <http://www.cpc.unc.edu/rlms/>.

RLMS является панельным обследованием, т.е. интервьюерами посещаются одни и те же семьи. Выборка RLMS изначально является выборкой домохозяйств, и поэтому

---

<sup>1</sup>в формате SAS Transport files. Для конвертации в другие форматы можно воспользоваться программой StatTransfer, входящей в комплект поставки Professional Stata.

Таблица 4.1: Выборка RLMS

Параметры выборки	Проект	Реализация			
		Раунд 5	Раунд 6	Раунд 7	Раунд 8
Объем выборки					
домохозяйств	4718	3973	3781	3750	3831
индивидуумов		11284	10648	10465	10677
Кол-во страт	38				

результаты RLMS должны в первую очередь относиться к генеральной совокупности домохозяйств. Впрочем, представительность выборки индивидуумов, как показывает сравнение ее основных социальных и демографических характеристик с результатами переписи 1989 г., также вполне удовлетворительна. Данные о выборке и участии домохозяйств в обследовании приводятся в таблице 4.1.

Выборка домохозяйств RLMS была сделана по схеме многоступенчатой стратификации<sup>2</sup>, т.е. последовательного случайного выбора. В выборку были включены саморепрезентативные страты, т.е. страты, выбираемые с вероятностью 1 в силу своей уникальности — Москва, Московская область, С.-Петербург. В качестве первичных единиц выборки (PSU) были использованы административные районы областей или крупных городов. Ряд местностей был исключен из-за труднодоступности, низкой плотности населения или ведения боевых действий; общая численность населения исключенных местностей составляет порядка 4.4% населения РФ. Из каждой страты выбирался один район (PSU). Вторичной единицей выборки (SSU, secondary sampling unit) являются участки переписи, избирательные участки или почтовые отделения (в порядке предпочтения)<sup>3</sup>. Наконец, на третьем уровне выбираются сами домохозяйства.

В силу описанной структуры выборки, RLMS нельзя использовать для анализа региональных данных. Точнее, RLMS не является представительным обследованием на региональном уровне. Так, из 89 субъектов Федерации, обследование затрагивает чуть более трех десятков, при этом в одних субъектах опрашивается только городское население, в других — только сельское. Безусловно, при наличии внешних данных о состоянии

<sup>2</sup> См. также с. 61.

<sup>3</sup> У саморепрезентативных страт эти единицы являются *первичными*, что необходимо учитывать при расчете поправок на стратифицированную структуру выборки.

региона — таких, как темпы инфляции или уровень безработицы — их вполне можно включать в регрессии там, где это уместно. Корректность использования данных является в некотором смысле направленной: использовать хорошие региональные данные в RLMS можно, а данные RLMS в региональных исследованиях — нельзя.

Интервьюерами заполняются три типа анкет: семейная, индивидуальная для взрослых и индивидуальная для детей. Семейную анкету заполняет член семьи, наиболее сведущий в ее ресурсных и финансовых потоках. Детские анкеты заполняются родителями. Кроме того, создается также файл данных, содержащих сведения об инфраструктуре поселения и ценах местной торговой сети (коммунальные данные). Эти данные распространяются отдельно от индивидуальных и семейных данных, и для их использования необходимо заполнить определенное соглашение с университетом.

Данные, полученные из заполненных анкет, представлены в Интернете. Кроме того, разработчики RLMS проводят минимальную чистку и сверку этих данных, результаты которой также имеются в открытом доступе. Файлам данных даются следующие имена:

- **r#hh\*** — исходные данные семейных анкет;
- **r#he\*** — переработанные данные семейных анкет;
- **r#in\*** — исходные индивидуальные данные;
- **r#\*** — прочие вторичные данные (потребление алкоголя, табака, калорийность питания и т.п.)

Здесь # обозначает номер раунда, а \* — произвольное окончание. Так, файл с исходными данными о доходах домохозяйств за седьмой раунд будет носить название **r7hhincm**. Всего таких файлов — около двух десятков за каждый раунд. Кроме самих данных, в Интернете имеются и pdf-файлы с бланками всех анкет (на английском языке).

Во всех файлах данных имеются идентификаторы семей и/или индивидуумов, которые можно использовать в команде **merge**. В пределах одного раунда такими идентификаторами являются переменные **site#** (номер местности), **censused#** (номер участка — участка всеобщей переписи, избирательного участка или зоны охвата почтового отделения в городах, деревни в сельской местности), **family#** (идентификационный номер семьи) и **person#** (номер индивида в пределах домохозяйства — в индивидуальных данных), где # — по-прежнему номер раунда. Возможны, впрочем, мелкие отклонения; так,

в данных 6-го раунда вместо переменных `site6`, `censUSD6`, `family6`, `person6` имеются переменные `site`, `census`, `family`, `person`, что создает определенные неудобства при попытках написать программы, универсальные для всех периодов. Для совмещения данных за разные раунды<sup>4</sup> следует пользоваться переменными `aid`, `bid`, `cid` и `did`, представляющими собой уникальные идентификаторы домохозяйств или индивидуумов за соответствующие раунды. К сожалению, и для индивидуумов, и для домохозяйств используются переменные с одним и тем же названием; переменные для домохозяйств при этом на две цифры короче.

Для определенных задач (например, анализа с учетом стратификации командами `svu*` или бутстрепа) могут потребоваться идентификаторы страт (и, соответственно, первичных единиц выборки). Эта информация содержится в переменных `psu` или `psu#`. Они в файлах RLMS встречаются редко, однако, поскольку структура выборки зафиксирована, они однозначно соответствуют переменным `site#` за разные раунды (одному `psu` соответствует несколько `site`).

Основными темами обследования RLMS являются здоровье и экономические характеристики населения. Наряду с указанными домохозяйствами номинальными экономическими показателями, во вторичных файлах RLMS приводятся также “реальные” (дефлированные) показатели<sup>5</sup>.

Все переменные во всех файлах имеют описания (во всяком случае, в исходных файлах, размещенных в Интернете). При конвертации программой StatTransfer эти описания сохраняются.

Начиная работать с данными RLMS (как и любой другой базы данных), помните о правилах “хорошего стиля”:

---

<sup>4</sup>При работе с домохозяйствами, участвовавших во всех раундах обследования, следует иметь в виду, что подобная панель может иметь представительность хуже, чем исходная выборка, если выбывание домохозяйств из обследования не случайно (а, например, коррелировано с их доходом или составом; см. Айвазян, Колеников (2000)). Авторы RLMS стараются отслеживать домохозяйства, переезжающие на новое место, для того, чтобы RLMS сохраняла представительность не только текущую (т.е., скажем, выборка девятого раунда за 2000 г. представительна для населения России за этот год), но и продольную (т.е. домохозяйства, представленные во всех раундах, формируют представительную выборку), что, конечно, гораздо сложнее.

<sup>5</sup>Для построения дефляторов используются данные “Обзора экономики России” (Russian Economic Trends); базовым периодом выступает 1992 г.

1. Необходимо хранить исходные файлы в сохранности, модифицируя их do-файлами и сохраняя, при необходимости, в виде отдельных новых файлов. Это полезно не только для восстановления ценных исходных файлов данных и результатов собственных исследований в случае сбоя, но и для возможности, хотя бы теоретической, воспроизведения ваших результатов другими исследователями.
2. Из числа прочих правил работы с данными, упоминаемыми в разных частях главы 3, стоит напомнить о необходимости описания данных (`label data`) и переменных (`label variable`) непосредственно после их создания, а также о возможностях внесения комментариев в файлы данных (`notes`). Эти функции пакета Stata начинают особенно цениться при обращении к файлам, созданным несколько недель (и тем более месяцев) тому назад...

В заключение упомянем, что, по данным Университета Северной Каролины, базой данных RLMS пользуются около трехсот научно-исследовательских организаций по всему миру.



## Глава 5

### Заключение

В данном пособии были рассмотрены основные аспекты прикладного эконометрического анализа. Безусловно, приведенный материал страдает схематичностью: практически каждая из рассмотренных проблем вполне может послужить темой для отдельной монографии. Автор скорее ставил целью не изложить детально всевозможные аспекты регрессионного анализа, а подсказать читателю, какие методы анализа данных вообще существуют и как можно выяснить, следует ли применять эти методы в данном конкретном случае, а также познакомить читателя с эконометрическими методами, встречающимися в современной литературе.

Для дальнейшего чтения могут быть порекомендованы, в первую очередь, книги Айвазян, Мхитарян (1998) и Greene (1997). Некоторые из более узких тем освещены в специальной литературе, а также в справочниках по эконометрике и статистике, ссылки на которые также приводятся в списке литературы. Число источников на русском языке, к сожалению, достаточно ограничено, в особенности в отношении пособий и монографий по эконометрике, с которой российские исследователи и студенты стали знакомиться только в последние годы.

Тем не менее, автор надеется, что это пособие поможет в прикладной работе экономистам-исследователям в анализе реальных данных и студентам в освоении предмета эконометрики.

## Глава 6

### Домашние задания

Неотъемлемой частью любого учебного курса являются домашние задания. Данный курс является сугубо практическим и прикладным, и домашние задания выстроены соответствующим образом.

Перед семинаром предлагается нулевое домашнее задание, предназначенное в основном для отбора слушателей для семинара. Оно предназначено для того, чтобы потенциальные слушатели могли реально соотнести свои возможности с уровнем сложности материала курса. Представление о том, как решать такие и подобные задачи, является отправной точкой для усвоения материала курса.

Для выполнения нулевого домашнего задания я настоятельно рекомендую ознакомиться с книжкой по эконометрике Катышева и Пересецкого (хотя бы в объеме первых трех глав — Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс. М., Дело, 1997) или с соответствующими главами книжки Айвазяна и Мхитаряна (Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М., ЮНИТИ, 1999), посвященными регрессионным моделям и методу наименьших квадратов.

Если вам кажется, что утверждение задачи некорректно или ошибочно, укажите, почему.

1. (Магнус, Катышев, Пересецкий, 1997) Что произойдет с МНК-оценками, если к одному из регрессоров добавить константу? Если к зависимой переменной добавить константу? Если заменить регрессоры и зависимую переменную на отклонения

от средних значений? Исследуйте, как изменятся оценки (если изменятся) и как изменится значимость регрессоров (если изменится).

2. В модели множественной регрессии наряду с регрессором  $x$  не имеет смысла использовать его степени  $x^2, x^3, \dots$ , так как эти степени являются зависимыми от регрессора  $x$  и, следовательно, не дают никакой дополнительной информации. Обоснуйте или опровергните.
3. У всякой регрессии сумма остатков равна нулю. Обоснуйте или приведите контр-пример.

Дальнейшие задачи связаны с пакетом Stata, занятиями курса или материалом пособия.

1. Сколько параметров должно быть у команды `regress` пакета Stata?
2. Как по распечатке регрессии понять, какие переменные статистически значимы, и значима ли вся регрессия в целом?
3. Известно, что мультиколлинеарность и гетероскедастичность увеличивают ошибки МНК-оценок коэффициентов. Если оба этих эффекта действуют одновременно, можно ли за счет борьбы с одним из них ослабить эффект другого?
4. Какие значения статистики  $R^2$  вы бы сочли хорошими, и почему: 0.7315, 0.0082, 0.1041, 0.9989, 0.9305, 0.5000?
5. Воспроизведите на данных `auto.dta` графики на рис. 2.3–2.8.
6. Рассчитайте по данным RLMS среднедушевые доходы и расходы домохозяйств. Совпадают ли эти цифры? Должны ли они совпадать?

По окончании курса слушателям предлагается выполнить небольшое исследование по мотивам RLMS с использованием пакета Stata, заключающееся в подборе данных, выборе и обосновании спецификаций регрессии, формулировке и проверке статистических гипотез, а также в диагностике полученных результатов.

**Задание.** По данным одного из раундов RLMS рассчитайте, как связаны между собой уровень образования и доходы. Что необходимо учитывать, если объединять данные

*за несколько раундов? Можно ли на основе полученных результатов утверждать, что наличие высшего образования повышает или понижает зарплату на столько-то рублей / столько-то процентов?*

Срок выполнения задания — две недели.

Задание, безусловно, представлено в максимально общем виде, в целях приближения обстановки к “боевой”: в условиях реального исследования будет необходимо точно так же выбирать переменные для анализа, вычищать данные, выбирать спецификацию модели, проводить диагностику регрессии и т.п.

# Литература

- Айвазян С. А., И. С. Енюков, Л. Д. Мешалкин. Прикладная статистика. Исследование зависимостей. М., “ФиС”, 1983.
- Айвазян С. А., С. О. Колеников. Бедность и дифференциация по расходам в России. Заключительный отчет для Российской программы экономических исследований, 2000.
- Айвазян С. А., В. С. Мхитарян. Прикладная статистика и основы эконометрики. М., ЮНИТИ, 1998.
- Демиденко Е. З. Линейная и нелинейная регрессия. М., “ФиС”, 1981.
- Закс. Теория статистических выводов. М., Мир, 1978.
- Кендалл М. Дж., А. Стюарт. Статистические выводы и связи. М., Наука, 1973.
- Магнус Я., П. К. Катышев, А. А. Пересецкий. Эконометрика. Начальный курс. М., “Дело”, 1997.
- Математическая энциклопедия. М., “Советская энциклопедия”, 1984.
- Себер Дж. Линейный регрессионный анализ. М., “Мир”, 1980.
- Справочник по прикладной статистике. П/р Э. Ллойда и У. Ледермана. Пер. с англ. п/р Ю. Н. Тюрина. М., “ФиС”, 1989.
- Тюрин, Ю. Н., А. А. Макаров. Статистический анализ данных на компьютере. М., Инфра-М, 1998.
- Хардле В. Прикладная непараметрическая регрессия. М., “Мир”, 1993.

- Хьюбер П. Робастность в статистике. М., “Мир”, 1984.
- Шеффе Г. Дисперсионный анализ. М., Наука, 1980.
- Эфрон Б. Нетрадиционные методы многомерного статистического анализа. М., “ФИС”, 1988.
- Handbook of statistics. Volume 11. Econometrics. G.S. Maddala, C.R. Rao, H.D. Vinod (eds.). North-Holland, 1993.
- Handbook of econometrics, vol. 1 (ed. Z. Griliches, M. Intrilligator, 1983), 2 (ed. Z. Griliches, M. Intrilligator, 1984), 3 (ed. Z. Griliches, M. Intrilligator, 1986), 4 (ed. R. Engle, D. McFadden, 1994). Elsevier.
- Baltagi, B. H. Econometric Analysis of Panel Data. John Wiley & Sons, 1995.
- Dempster, A. P., M. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society*, **B39**, 1–38 (1977).
- Draper, N., H. Smith. Applied regression analysis. 3rd edition. Wiley, 1998 (имеется русские переводы 1-го и 2-го изданий: Н. Дрейпер, Х. Смит. Прикладной регрессионный анализ.).
- Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, **7**, 1–26, 1979.
- Fox, J. Applied regression analysis, linear models, and related methods. SAGE, 1997.
- Gallup, J. outreg — Formatting regression output. *Stata Technical Bulletin*, **46** (1998), **48** (1999), **58** (2000), **59** (2001).
- Gould, W., W. Sribney. Maximum Likelihood Estimation with Stata. Stata Press, 1999.
- Greene, W. H. Econometric Analysis. 3rd edition. Prentice Hall, 1997.
- Hausman, J. Specification Tests in Econometrics. *Econometrica*, **46**, 1251–1271, 1978.
- Kolenikov, S. Review of Stata 7. *J. of Applied Econometrics*, forthcoming.
- Konishi, S., and G. Kitagawa. Generalized information criteria in model selection. *Biometrika*, **83** (4), 875–890, 1996.

- Little, R. J. A., and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley (1987).
- Maddala, G. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge Univ. Press, 1983.
- Maddala, G. *The Econometrics of Panel Data*. Brookfield, 1993.
- Mander, A., and D. Clayton. Hotdeck imputation. *Stata Technical Bulletin*, **51** (1999), **54** (2000).
- Matyas, L., ed. *Generalized method of moments estimation*. Cambridge University Press, 1999.
- Mroz, T., D. Mancini, B. Popkin. *Monitoring Economic Conditions in the Russian Federation*. The Russia Longitudinal Monitoring Survey 1992–98. Report submitted to the USAID. Carolina Population Center, University of North Carolina at Chapel Hill, 1999.
- Newey, W. K., K. D. West. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, **55**, 703–708, 1987.
- Neyman, J., and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20-A**: 175–247, 264–299 (1928).
- Rubin, D. B. Inference and missing data. *Biometrika*, **63**, 581–592 (1976).
- Rubin, D. B. Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse. *Imputation and Editing of Faulty or Missing Survey Data*. U.S. Department of Commerce, pp. 1–23 (1978).
- Smith, R., and K. Young. *Linear Regression*. Oxford University Press (2001).
- StataCorp. *Stata Statistical Software*. Release 6 (1999). Release 7 (2001).
- Swafford, M. *Sample of the Russian Federation*. Rounds V and VI of the Russian Longitudinal Monitoring Survey. Technical Report. Paragon Research International, 1996.
- Wessie, J. mmerge — Safe and easy matched merging. *Stata Technical Bulletin*, **53** (1999).