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[1]:

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-

[1] 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 наблюденными) 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 (proxy) к нужным параметрам (например, квалификация работника сама по себе может не быть наблюдаема, однако в качестве апроксимации 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 исследуется 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 статисических 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, Стьюарт 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):

$$\mathsf{E}[y|x] = f(x). \tag{2.1}$$

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

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.2}$$

где y_i — зависимая переменная, x_i — вектор объясняющих переменных, $x_i \in \mathbb{R}^p$, β — вектор параметров соответствующей размерности, ε_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:

$$\mathbf{y} = \mathbf{X}^T\beta + \varepsilon, \tag{2.3}$$

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

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

Чаще всего полагается, что $x_{i1} = 1$, тогда коэффициент β_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:

$$\mathsf{E}\varepsilon_i = 0 \tag{2.5}$$

$$\mathsf{E}\varepsilon_i^2 = \sigma^2 \tag{2.6}$$

$$\mathsf{E}\varepsilon_i\varepsilon_j = 0 \; \forall i \neq j \tag{2.7}$$

$$\operatorname{rk}\mathbf{X} = p < n \tag{2.8}$$

$$X_j \quad \text{детерминированы}. \tag{2.9}$$

Часто бывает полезным предположение о явной форме ошибок:

$$\varepsilon_i \sim N(0, \sigma^2) \tag{2.10}$$

2.2.2 Метод наименьших квадратов

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

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

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

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

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

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[2], которая равна

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

[2] 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 σ^2:

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

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

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

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

где C — матрица $q \times p$ полного ранга по строкам ($\mathrm{rk}\, C = q < p$), а r — вектор $q \times 1$. Иными словами, гипотеза H_0 накладывает на коэффициенты q ограничений. Примером такой гипотезы может служить $H_0 : \beta_2 = \ldots = \beta_p = 0$, 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:

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

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[3]

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

имеющая при H_0 распределение Стьюдента с $n-p$ степенями свободы, где оценка дисперсии $\widehat{\mathrm{Var}}(\hat{\beta}_k)$ — 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):

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

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

$$\mathsf{P}\left[|\hat{\beta}_k| > |\hat{\beta}_k \text{ наблюденное}| \middle| H_0 \right]. \tag{2.20}$$

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.

[3] 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 матожиданиями при условии x. При этом сама задача должна быть переформулирована в терминах случайной выборки, и необходимость в условии (2.7) отпадает по определению последней [4]. Необходимо так-

[4] 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 $\mathbf{X}^T\mathbf{X}$:

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}^T\mathbf{X} = M > \mathbf{0}_{p\times p} \tag{2.21}$$

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

$$\mathsf{E}[\varepsilon|x] \neq 0 \tag{2.22}$$

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. парагаф 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 ε, 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 χ^2 ((Neyman, 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 χ^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:

$$y_i = x_i^{*T} + \varepsilon_i \tag{2.23}$$

$$x_i = x_i^* + \delta_i \tag{2.24}$$

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 δ_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

$$\Omega = \operatorname{Var} \varepsilon \tag{2.25}$$

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

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

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

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

а дисперсия "наивной" оценки МНК —

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

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 Ω делаются разнообразные предположения о параметрической зависимости Ω от некоторого малого числа параметров θ известного вида: $\Omega = \Omega(\theta)$, где вектор параметров θ 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 θ (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 $\widehat{\Omega(\theta)}$), самой регрессионной модели. Для уточнения оценок процедуру "оценивание $\theta \to$ 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):

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

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

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

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

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)):

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

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 является 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 [5] (за исключением досадных оплошностей типа включения в набор регрессоров всех 0/1-переменных, порождаемых одним и тем же фактором, например, индикаторов u мужского, u 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 категорийными 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 категорийными 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).

[5] 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

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

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:

$$\mathrm{Var}\,\beta_j = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1)\,\mathrm{Var}\,X_j^2} \tag{2.33}$$

и показывает, во сколько раз дисперсия оценки больше "идеальной", если бы мультиколлинеарности не было[6]. 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-

[6] Стандартная ошибка оценки, очевидно, увеличивается в $\sqrt{\mathrm{VIF}}$ раз. Эта величина имеет смысл диагностический, а не практический: нельзя делить на VIF для того, чтобы получить "правильную" дисперсию!

ми[7]: 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 $\operatorname{Var} X_j^2$ 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 среднеквадратическое отклонение, или среднеквадрати-

[7] 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:

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

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 νI_p, где I_p — единичная матрица размера p. Тогда оценка будет иметь вид:

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

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 ν дисперсия оценок уменьшается, хотя увеличивается их смещение. Можно показать, что существует ν такое, что среднеквадратическая ошибка из (2.34) смещенной оценки ниже, чем у несмещенной оценки МНК, т.е. можно подобрать ν 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 ε) 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 принципиального 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 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 наблюденных 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 $y_i \to \infty$, то и $\hat{\beta}_{\text{МНК}} \to \infty$.

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

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

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

Еще одна частно используемая спецификация — функция Хьюбера (Huber)

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

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

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

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

Здесь c — также параметр робастности. При $c \to \infty$ 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

[8] 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):

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

где $\dot{y} = \left(\prod_{i=1}^n y_i\right)^{1/n}$ — среднее геометрическое y_i. Оценку необходимой степени преобразования λ можно произвести методом максимального правдоподобия [9]. 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

[9] 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

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

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 синтакс 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 $R^2 \to 1$.
- 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^2_{adj}, в которой более тонко учитывается число степеней свободы модели:

$$R^2_{adj} = 1 - \frac{\mathbf{e}^T\mathbf{e}/n - p}{\mathbf{y}^T\mathbf{y}/n - 1}, \tag{2.41}$$

где e — вектор регрессионных остатков, а 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 [10]. Исторически первым, а потому наиболее популярным информационным критерием является *критерий Акайке* (AIC, Akaike information criteria):

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

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:

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

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

[10] Формально, информационные критерии являются более точными оценками ожидаемой информации модели, или математического ожидания функции правдоподобия, чем само максимальное значение функии 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^2_{adj}, 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 Бозтогана, находится на web-страничке автора и называется `icomp`[11].

2.4.2 Нелинейность

Другим возможным нарушением классической модели регрессии может быть случай, когда функция регрессии $\mathsf{E}[y|x]$ 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

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

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.

[11] 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 ε — "хорошие" (центрированные, независимые, с конечной дисперсией) ошибки. В подобных случаях преобразованием 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

$$y_i = f(x_i, \beta) + \varepsilon_i, \tag{2.45}$$

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 регресии $f(\cdot)$ 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-

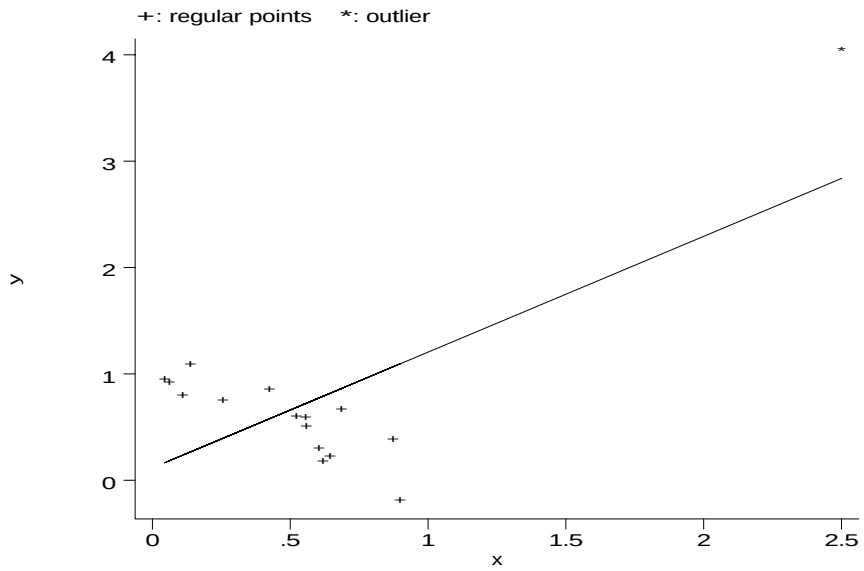


Рис. 2.1: Линия регрессии оттянута на себя выбросом. Истинная линия регрессии: $y = 1 - x + \varepsilon$

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 ε_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 [12]. Рассмотрим

[12] Данная тема, пожалуй, наименее типична для стандартных курсов по эконометрике, хотя стати-

прогнозные значения зависимой переменной:

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

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 наблюденного 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*[13]:

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

где $s_e^{(i)}$ — оценка стандартного отклонения остатков при исключении i-го наблюдения, а появление коэффициента $\sqrt{1-h_i}$ связано с тем, что $\operatorname{Var} e_i|H_0 = (1-h_i)\sigma^2$. 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).

[13] 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 γ 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):

$$D_i = \frac{e_i^2}{p}\frac{h_i}{1-h_i} \tag{2.48}$$

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 $DFBETA_{k,i}$:

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

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 строем 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:

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

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

Стьюдентизированные 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 ... [14]

[14] 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.

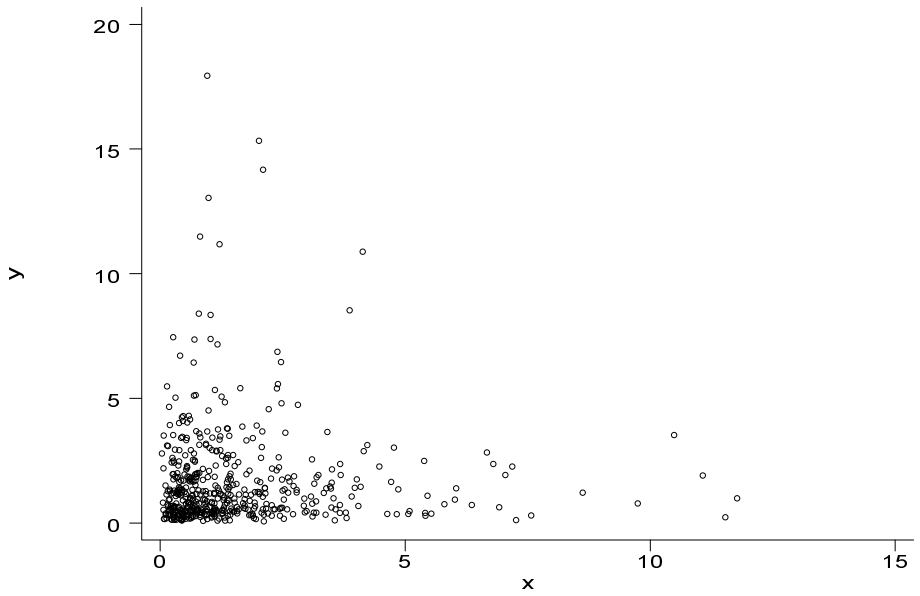


Рис. 2.2: Частные распределения обеих переменных асимметричны; график заполнен в основном около нуля и возле осей; необходимо преобразование к нормальности?

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

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

и

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

где верхний индекс $(-k)$ означает отутствие 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*:

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

Stata Соответсвующие 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,

$$\mathsf{P}\left(\cup_k \bar{A}_k\right) \leq \sum_k \mathsf{P}\left(\bar{A}_k\right) \tag{2.54}$$

а следовательно,

$$\mathsf{P}\left(\cap_k A_k\right) \geq 1 - \sum_k \mathsf{P}\left(\bar{A}_k\right) \tag{2.55}$$

В левой части (2.55) фигурирует вероятность принять нулевую гипотезу *во всех* регрессиях. Соответственно, если требуется, чтобы *совокупный* уровень значимости составлял α, 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 $\mathsf{P}(\bar{A}_k)$ не превосходил α/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 результы 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 $\mathsf{P}(X_j$ пропущено|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 $\mathsf{P}(X_j$ пропущено|прочие X) не зависит от X_j (но могут зависеть от других X). Оказывается, что в этих случаях механизм пропусков *несущественнен* (ignorable), и к данным применимы вариации метода максимального правдоподобия. Наконец, если $\mathsf{P}(X_j$ пропущено|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 существеннен.

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 линеной модели, обыгрывает идею восстановления данных по условному распределению: если условием является категорийная переменная (возможно, многомерная), то пропущенные данные можно подставить из числа наблюденных в той же группе (или, в некотором более общем виде, подставить значение, наблюденное 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 наблюденные величины, а Y_{miss} — пропущенные, которые исследователь мог бы наблюдать, если бы данные были полными.

Для стандартных моделей функция правдподобия 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 $R_{ij} = I(y_{ij}$ наблюдается) со своей функцией распределения $g(R|Y,\psi)$ [15], можно получить общую функцию правдоподобия в виде

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

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

[15] Очевидно, R наблюдается всегда.

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

Название теста	Принцип	"Плохие" признаки	Stata
Робастность, выбросы			
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	`summarize`; `sktest`; `graph переменная, norm`; `kdensity`; `qnorm`
D-статистика Кука, $DFFITS$, $DFBETA$	Идентификация выделяющихся наблюдений	Точки с высоким значением статистик влияния	`regress` → `predict, cooksd`; `predict, dfit`; `predict, dfbeta`
Визуальный анализ	Графики частных регрессий и остатков-прогнозов	Отдельно отстоящие точки	`avplot`; `rvfplot`
Стохастичность регрессоров			
Тест Хаусмана	Сравнение эффективной (при H_0), но несостоятельной (при H_a) модели с состоятельной (при обеих гипотезах), но менее эффективной (при H_0)	$\chi^2 \to \infty$	`hausman`

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
---------+------------------------------           F(  3,    70) =   23.29
   Model |   317252881     3   105750960           Prob > F      =  0.0000
Residual |   317812515    70  4540178.78           R-squared     =  0.4996
---------+------------------------------           Adj R-squared =  0.4781
   Total |   635065396    73  8699525.97           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
------------------------------------------------------------------------------
```

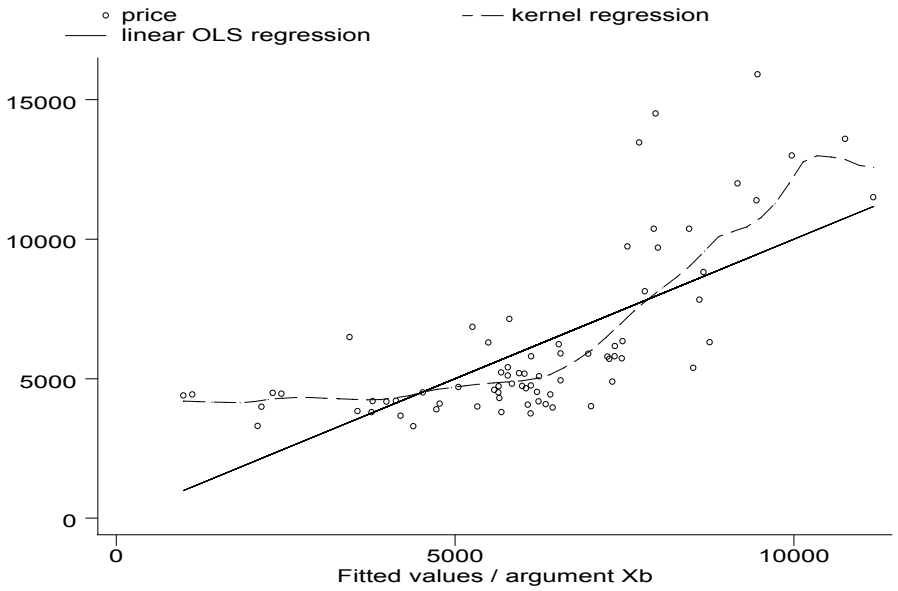
Здесь в левом верхнем углу — таблица дисперсионного анализа (с указанием суммы квадратов и доли дисперсии y, объясненных моделью, суммы квадратов остатков и их дисперсии, общая сумма квадратов и дисперсия y), справа вверху — прочая информация, связанная с регрессией (количество наблюдений, общая F-статистика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

H_0: все коэффициенты равны нулю, кроме константы; статистики R^2 и R^2_{adj} 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 апроксимация функции регрессии не является адекватной, что и подтверждается тестом Рамсея на нелинейность (2.44).

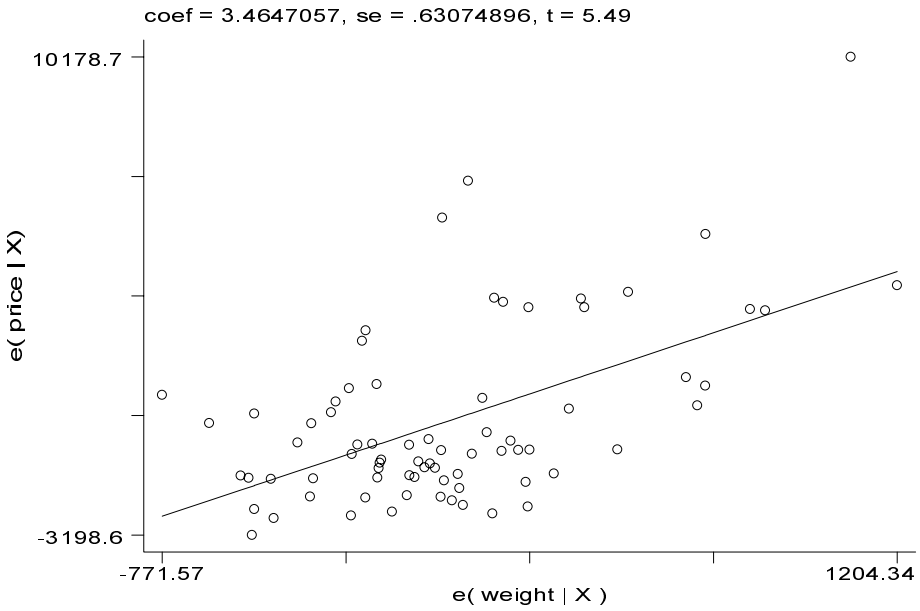
Рис. 2.3: Регрессия в пространстве прогнозных значений: прямая, полученная по МНК, и непараметрическая оценка кривой регрессии. Видно значительное расхождение.



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` , `cooksd` и `hat` [16].

[16] Подчеркивания показывают минимально возможные сокращения; см. раздел 3.1

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

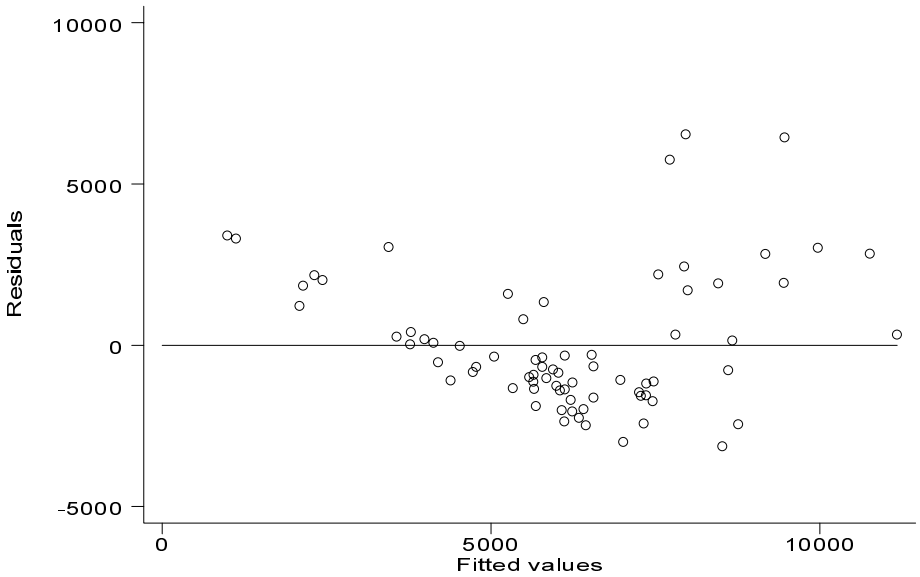
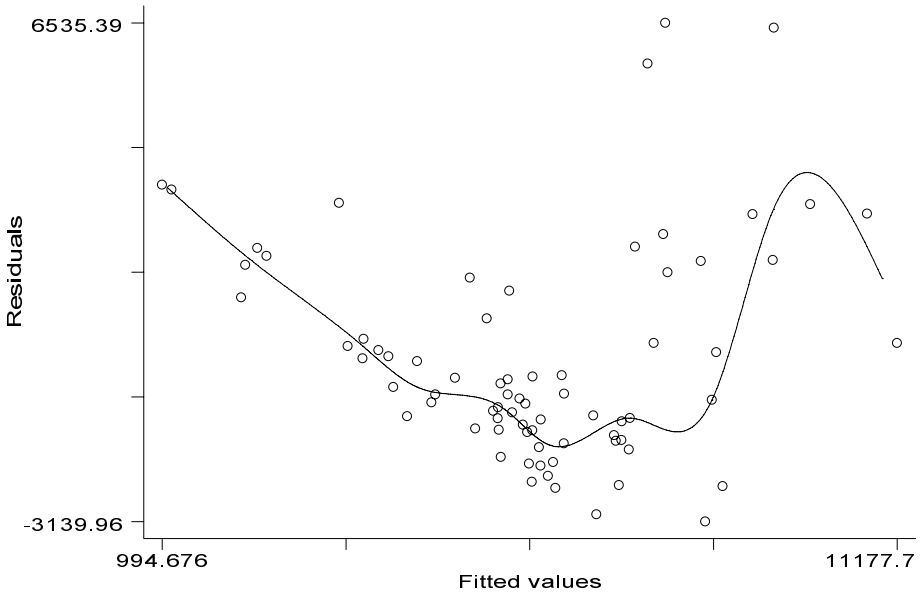
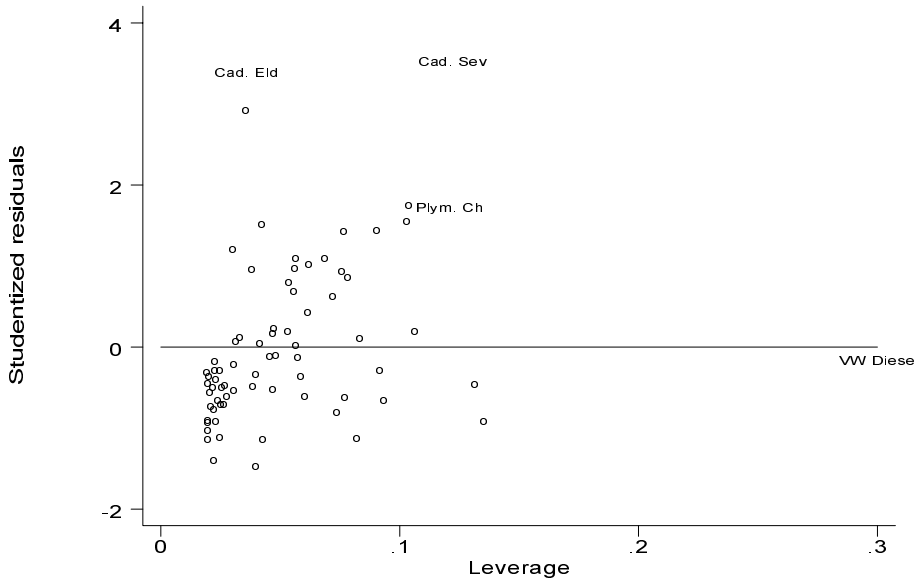


Рис. 2.6: Диаграмма рассеяния остатков (`rvfplot, c(s) bands(10) d(50)`).



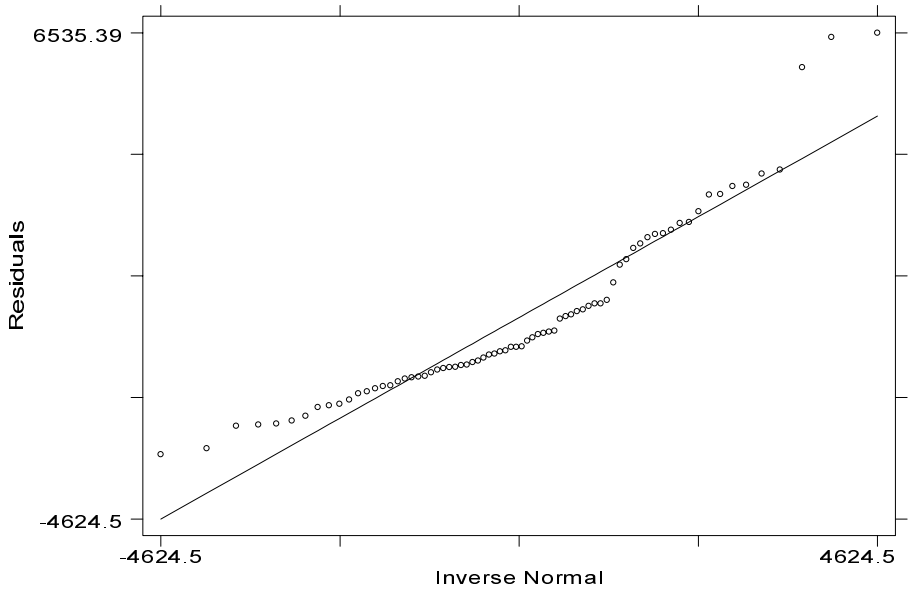
На рис. 2.7 приведен график, связывающий относительное влияние каждого наблюдения (leverage) и величину стьюдентизированного остатка. Произведение этих величин составляет расстояние Кука D. Более подробное объяснение см. в разделе 2.4.3. Наблюдения, которые могут оказывать существенное влияние на коэффициенты, 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 ... , cooksd` 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 χ^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:

$$y_{it} = x_{it}^T \beta + u_i + \varepsilon_{it} \tag{2.57}$$

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 временно́й 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:

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

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 стратифицкации.

Stata Пакет Stata обладает весьма обширным набором средств, позволяющих учитывать стратификационный характер выборок — это около двух десятков комманд 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,

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

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:

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

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: $sup_{x\in(-\infty,+\infty)}|F_{logit}(x) - F_{N(0,1)}(x)| < 0.02$, 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) [17]. 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 (Gomperz, соответственно, гомпит/gompit-модель):

$$F(z) = 1 - \exp[-\exp(z)] \tag{2.61}$$

[17] 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:

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

что может быть очень удачно переписано как

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

Тогда общая функция правдоподобия имеет вид:

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

Задача максимизации этой функции по β 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 коэффицентов 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 χ^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 интепретированы как предельные эффекты (т.е. изменения зависимой переменной при измененении 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.

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

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)):

$$\sum_{i=1}^{N} |y_i - x_i\beta| \to \min \tag{2.66}$$

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 наблюденные 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:

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

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:

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

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):

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

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

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

$$\int K(u)du = 1 \tag{2.72}$$

Здесь (2.70) — непараметрическая (ядерная) оценка плотности в данной точке (называемая также *оценкой Розенблата-Парзена*), (2.71) — ядро масштаба h_n (ширина которого может зависить 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:

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

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

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

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

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

Здесь I(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[18]. 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.

[18] При $h \to \infty$, $f(x) \to \bar{y}$.