квантиль

международный эконометрический журнал на русском языке

№7
сентябрь 2009 г.

СОДЕРЖАНИЕ ВЫПУСКА

квантиль

№7, сентябрь 2009 г.

Сайт в Интернете: http://quantile.ru
Адрес электронной почты: quantile@quantile.ru
Доступ к журналу бесплатный и неограниченный

К СВЕДЕНИЮ АВТОРОВ

Рукописи для публикации в разделе «Статьи» принимаются в электронном виде по адресу *submit@quantile.ru*. Работы могут принадлежать любой прикладной сфере экономической науки. Главным требованием является интенсивное использование адекватных эконометрических методов. Рукопись должна быть написана на русском (для русскоязычных авторов) или на английском (для остальных авторов) языке в формате *Microsoft Word* или (предпочтительнее) *LaTeX*, и по объему не превышать 30 страниц формата A4 с двойным междустрочным интервалом. Работы подвергаются контролю качества членами редакционного совета и независимыми референтами. Перспективная работа может быть при необходимости возвращена автору на доработку. Редакция также приглашает к сотрудничеству экспертов по эконометрике, готовых внести вклад в методологические рубрики журнала.

При публикации статьи или методологического эссе в журнале «Квантиль» передача авторских прав не происходит ни полностью, ни частично.

Решения задач из рубрики «Задачи и решения», а также новые задачи можно присылать по адресу *ps@quantile.ru*.

Эконометрический ликбез: вопросы микроэконометрики

Разрывный дизайн*

Антон Ниворожкин†

Институт проблем занятости, Нюрнберг, Германия

В настоящем эссе представлен краткий обзор теории и эмпирического применения разрывного дизайна для оценивания эффекта воздействия программ. На примерах осуществленных исследований рассмотрены основные трудности в применении разрывного дизайна на практике и интерпретации его результатов.

1 Введение

Разрывный дизайн впервые был предложен в Thistlethwaite & Campbell (1960) как альтернатива рандомизированным экспериментам для оценивания эффекта воздействия программ. Авторы, анализируя влияние наличия у части студентов именных стипендий на результаты успеваемости, использовали тот факт, что присуждение этих наград было основано на наблюдаемых значениях специального экзаменационного теста. В их исследовании победителями становились все соискатели, получившие оценку выше минимальной. Идея Thistlethwaite & Campbell (1960) состояла в том, что исследователь может использовать информацию об установленном пороге (разрыве) в значениях теста (выше или ниже минимальной оценки теста), чтобы выявить эффект влияния именной стипендии на уровень успеваемости среди лиц, которые получили оценки маржинально выше или ниже пороговой оценки (точки разрыва). При соблюдении некоторых условий распределение в группы воздействия и контрольную может рассматриваться как полностью случайное в непосредственной близости от порогового значения.

При определенных условиях сравнение среднего эффекта воздействия для объектов, находящихся в непосредственной близости от порогового значения, может помочь в выявлении причинно-следственных связей, имеющих социально-экономическое содержание, важное для принятия адекватных управленческих решений.

Несмотря на то, что данная стратегия оценивания известна уже более 50 лет, практическую значимость в социально-экономических исследованиях она приобрела сравнительно недавно. Следуя за Thistlethwaite & Campbell (1960), Angrist & Lavy (1999) использовали правило порога, генерирующее разрыв в величине школьного класса, для оценки влияния количества учеников в классе на результаты успеваемости.

Вслед за ранними статьями в области оценивания эффектов влияния различных факторов на качество образования, литература по разрывному дизайну обратилась и к другим социально-экономическим проблемам, таким как эффекты социальных программ (Lemieux & Milligan, 2008), эффекты воздействия пособий по безработице или пособий по нетрудоспособности на предложение труда (Lalive, 2008; Chen & van der Klaauw, 2008), эффекты воздействия субсидий в медицинском секторе на улучшение здоровья (Card, Dobkin & Maestas,

*Данное эссе в значительной мере основывается на статье Lee & Lemieux (2009) и Imbens & Lemieux (2008). Цитировать как: Ниворожкин, Антон (2009). «Разрывный дизайн», Квантиль, №7, стр. 1–8. Citation: Nivorozhkin, Anton (2009). "Regression discontinuity design," Quantile, No.7, pp. 1–8.

†Адрес: Institute for Employment Research (IAB), Regensburger Straße 104, 90478, Nürnberg, Germany. Электронная почта: *anton.nivorozhkin@iab.de*

2008), а также эффекты воздействия профсоюзов на заработную плату и трудоустройство (DiNardo & Lee, 2004).

Одним из объяснений возрастающей популярности разрывного дизайна является понимание того, что разрывный дизайн требует менее строгих допущений по сравнению с другими квази-экспериментальными методами. Другой причиной является мнение, высказанное в Lee & Lemieux (2009) и Lee (2008), о том, что разрывный дизайн является не просто дополнительным методом для изучения причинно-следственных связей, но потенциально заслуживает большего доверия, чем другие типичные стратегии, основанные на естественном эксперименте, такие как метод «разность разностей» (см. Вулдридж, 2009) или инструментальные переменные (см. Эббес, 2007). Эта идея получила теоретическое обоснование у Lee (2008), который формально доказал, что нет необходимости в предположении, что разрывный дизайн исключает возможность индивидов влиять на вероятность попадания в группу воздействия или контрольную группу. Однако данный процесс все равно можно рассматривать как случайный как следствие того, что процесс попадания индивидов в некоторую малую область выше или ниже установленного порога не может полностью быть ими контролируемым.

Разрывный дизайн появился относительно недавно в арсенале эмпирических исследований, и число публикаций по его применению пока не столь велико, хотя и быстро растет. В российских публикация мы пока не нашли примеров его использования. Для заинтересованных читателей можно порекомендовать статьи, опубликованные на английском языке (van der Klaauw, 2008; Imbens & Lemieux, 2008; Lee & Lemieux 2009).

2 Определения

Для того чтобы ввести разрывный дизайн в общий контекст современной литературы по вопросам оценивания эффекта воздействия, необходимо знакомство с квази-экспериментальным подходом, предложенным в Rubin (1974) и представленным в Еникилопов (2009). Впервые данный подход в применении к разрывной регрессии был описан в Hahn, Todd & Van der Klaauw (2001).

Пусть для N объектов, случайно выбранных из генеральной совокупности, где каждый объект обозначен индексом i, $i = 1, \dots, N$, определено, что каждый из них подвержен одному из двух типов воздействия: $W_i = 0$, если объект i подвергнут контрольному воздействию, и $W_i = 1$, если объект i подвергнут активному воздействию.[1] Предполагается, что для каждого объекта i определена пара потенциальных исходов, $Y_i(0)$ и $Y_i(1)$, и эффект воздействия определен как $Y_i(1) - Y_i(0)$. Основной проблемой, с которой сталкивается исследователь, является то, что невозможно наблюдать одновременно оба исхода $Y_i(0)$ и $Y_i(1)$. Таким образом, для каждого объекта мы наблюдаем исход Y_i, выраженный как

$$Y_i = (1 - W_i)Y_i(0) + W_i Y_i(1) = \begin{cases} Y_i(0) & \text{при} \quad W_i = 0, \\ Y_i(1) & \text{при} \quad W_i = 1, \end{cases}$$

где $W_i \in \{0, 1\}$.

Основная идея разрывного дизайна состоит в том, что воздействие определяется полностью или частично значением переменной, определяющей право на участие в программе, X_i. Эта переменная может коррелировать с потенциальными исходами, но предполагается, что эта корреляция распределена равномерно с каждой стороны разрыва и, таким образом, разрыв в условном распределении исходов, индексированных значением переменной X_i, можно интерпретировать как причинно-следственный эффект в непосредственной близости от точки разрыва.

Как было отмечено, разрывный дизайн часто возникает как следствие административных решений, при которых число объектов, получающих воздействие, ограничено доступными

[1] Объектами могут быть как индивиды, так и фирмы или даже страны.

ресурсами, и четкие правила, а не субъективное решение чиновника, частично или полностью определяют право на получение воздействия.

3 Типы разрывов: четкий и нечеткий дизайн

Для дальнейшей дискуссии будет полезно разделить разрывный дизайн на два типа: четкий (*sharp*) и нечеткий (*fuzzy*) (Trochim, 1984; Hahn, Todd & Van der Klaauw, 2001).

3.1 Четкий дизайн

При четком дизайне распределение W_i является детерминированной функцией переменной, определяющей право на участие в программе X_i:

$$W_i = \mathbb{I}\{X_i \geq c\},$$

где c – пороговое значение. При значениях $X_i \geq c$ объекты приписываются к группе воздействия (и участие для индивидов в ней является обязательным), а при значениях $X_i < c$ наблюдения приписываются к группе контроля (и члены этой группы не имеют права на участие в группе воздействия ни при каких обстоятельствах). В случае четкого дизайна при расчете эффекта участия в программе разрыв рассматривается как условное математическое ожидание исхода воздействия таким образом, что

$$\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1)|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i(0)|X_i = x],$$

и может быть интерпретирован как причинно-следственный эффект в непосредственной близости от точки разрыва:

$$\tau_{SRD} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c].$$

Для того чтобы данный эффект являлся причинно-следственным, используется допущение о сглаживаемости.

Допущение 1. Непрерывность регрессионных функций:
Если $\mathbb{E}[Y_i(1)|X_i = x]$ и $\mathbb{E}[Y_i(0)|X_i = x]$ являются непрерывными по x, то

$$\tau_{SRD} = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x].$$

В эмпирической работе всегда есть необходимость экстраполяции, так как не существует наблюдений со значением $X_i = c$, для которых можно было бы наблюдать значение $Y_i(0)$. Таким образом, необходимо использовать наблюдения со значениями X_i, произвольно близкими к c.

Рассмотрим применение точного дизайна на графическом примере, предложенном в Thistlethwaite & Campbell (1960) и Lee (2009) (см. Рис. 1). Представим что для некого индивида i переменная X_i равна c, экзаменационный балл на право получения стипендии, равен пороговому значению, и, таким образом, данный индивид попадает в группу воздействия. Для того, чтобы получить оценку причинно-следственного эффекта воздействия в данном случае, необходимо знать, каков был бы исход Y в случае, если бы данный индивид воздействию не подвергался.

В случае, если остальные факторы существенно не менялись, то, вероятно, что B будет верной аппроксимацией значения Y в случае получения воздействия, и A в случае, если воздействие не было оказано. Таким образом, разница $B - A$ будет являться оценкой эффекта.

Рис. 1 является иллюстрацией того, что в случае применения разрывного дизайна необходимо использовать наблюдения, у которых значение переменной X_i близко к пороговому. Вместе с тем необходимо отметить, что в практической работе далеко не всегда утверждение

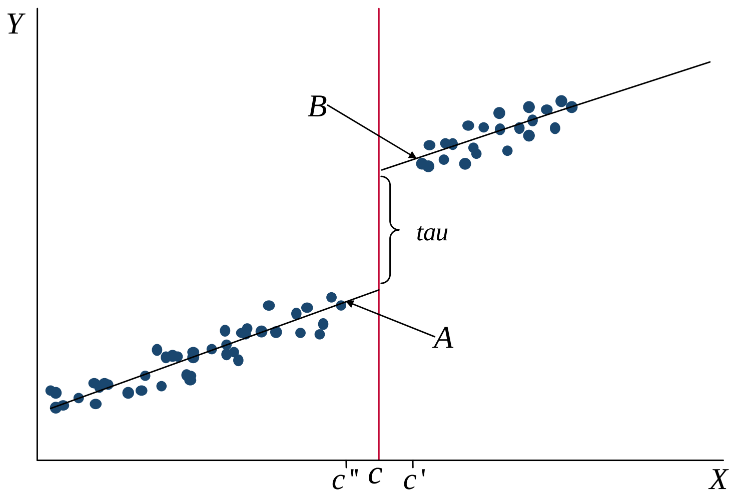


Рис. 1: Графический пример разрывного дизайна

«чем ближе к пороговому значению, тем лучше» оправданно. Чем более узок интервал, на котором оценивается эффект, тем меньше количество данных. К примеру, на Рис. 1 невозможно оценить эффект на интервале уже, чем c' и c'', так как такой интервал не содержит наблюдений.

В случае использования ограниченной выборки необходимо полагаться на аппроксимацию, использовать наблюдения на достаточно широком отрезке, таком как c' и c'', для того чтобы оценить эффект воздействия в точке c. В действительности, если взаимосвязь между Y и X линейная, то оценка параметра τ с помощью простой линейной регрессии в форме $Y = \alpha + \tau W + \beta X + \varepsilon$ будет являться наилучшей несмещенной оценкой.

Непременным условием для применения разрывного дизайна является то, что все остальные переменные, определяющие Y, должны быть гладкими функциями по X. Если одна или несколько переменных, определяющих Y, резко изменяются в точке c, то параметр τ будет смещенной оценкой воздействия. К примеру, если доходы родителей связаны с результатами сдачи теста и получением стипендии (поскольку дети из более состоятельных семей заканчивают школы с лучшим качеством подготовки), то проведенный анализ эффекта получения стипендии на академическую успеваемость будет, безусловно, смещен. Неизвестно, является ли успеваемость следствием получения стипендии или высоких доходов родителей, или же это – комбинация двух факторов.

3.2 Взаимосвязь разрывного дизайна и мэтчинга

Точный дизайн можно рассматривать как специальный случай отбора по наблюдаемым характеристикам, как поступили Heckman, Lalonde & Smith (1999) и Heckman & Robb (1985). Вместе с тем, разрывный дизайн существенно отличается от стандартных подходов к отбору по наблюдаемым характеристикам, таких как мэтчинг (см. Ениколопов, 2009). Допущение о несмешиваемости, являющееся нетестируемым в большинстве методов, тривиально выполняется в разрывном дизайне, так как при $X_i \geq c$ переменная W_i всегда будет равняться единице, а при $X_i < c$ переменная W_i будет всегда равна нулю. Таким образом, при условии наблюдения значений X_i вариации в W_i нет, и других ненаблюдаемых факторов не существует.

В то же время другое стандартное допущение о пересечении оказывается нарушенным, так как невозможно найти наблюдение с $W_i = 0$ или $W_i = 1$ для одного и того же значения переменной X_i. Это является причиной необходимости введения допущения о непрерывности в точке. И хотя нельзя наблюдать $W_i = 0$ и $W_i = 1$ для одного и того же значения переменной

X_i, можно наблюдать исходы для приблизительно одинаковых значений в непосредственной близости от порогового значения.

Приведем пример применения точного дизайна. Black (1999) предпринял попытку оценить готовность родителей платить за улучшение качества образования детей. В США качество школ зачастую тесно связанно с ценами на недвижимость, так как родители нередко предпочитают селиться в районах, где качество школ лучше. Стандартная регрессионная модель не может справиться с решением данной проблемы, поскольку велика вероятность того, что некоторые переменные, связанные с качеством школ в районе или с качеством жилья, будут опущены. Black (1999) для исследования отобрал только те домовладения, которые находились непосредственно на границе районов, соседствовали друг с другом и, следовательно, должны быть похожи. Единственным отличием являлось то, к какой школе был приписан дом. В терминах разрывной регрессии это представляется двумя школами по обе стороны границы (хорошая школа с правой стороны, $W_i = 1$, и плохая – с левой, $W_i = 0$) и ценами на недвижимость, плавно изменяющимися по мере отдаления от границы.

3.3 Нечеткий дизайн

В нечетком дизайне вероятность воздействия не меняется с нуля на единицу в точке разрыва. Вместо этого при нечетком дизайне вероятность изменения воздействия в точке разрыва всегда меньше единицы. Таким образом,

$$\lim_{x \downarrow c} \Pr\{W_i = 1 | X_i = x\} \neq -\lim_{x \uparrow c} \Pr\{W_i = 1 | X_i = x\}.$$

Данная ситуация возникает, когда стимул для участия в программе присутствует, но он недостаточно силен, чтобы все индивиды захотели участвовать в программе. В таком дизайне мы интерпретируем отношение изменения исхода к изменению в доле участников:

$$\tau_{FRD} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x]}.$$

Рассмотрим применение нечеткого дизайна на примере эффекта конкурса на получение стипендии для поступления в определенный университет, изученный van der Klaauw (2002). Определим X_i как оценку, данную университетом абитуриенту на основе его школьного аттестата и позволяющую университету разделить студентов согласно их успеваемости на L групп. Допустим, что в нашем случае существует только одна точка разрыва c и соответственно две группы. Получение оценки немного выше, чем c, позволяет абитуриенту войти в первую группу и существенно увеличить шансы на получение стипендии по сравнению со случаем, когда его оценка слегка ниже c, и он во второй группе.

В описанном случае связь между поступлением в университет и получением стипендии не является однозначной. С одной стороны, финансовая помощь, предложенная университетом, делает его более привлекательным для студента. С другой стороны, студент, который набрал достаточную сумму баллов для получения стипендии в одном университете, с большей вероятностью, чем не набравший баллов, получит помощь и в другом университете. В примере van der Klaauw (2002) решение университета о помощи не является детерминированной функцией оценок. Другие факторы, такие как грамотно составленное резюме и рекомендационные письма играют, несомненно, важную роль. Несмотря на это, автор нашел значительный разрыв в вероятности получения финансовой помощи в зависимости от оценок в аттестате, но данный разрыв не является четким, а скорее размытым, нечетким.

Перейдем к интерпретации τ_{FRD}. Эффект нечеткого дизайна может быть интерпретирован сквозь призму инструментальных переменных, когда эффект участия варьируется среди участников. Пусть $W_i(x)$ – потенциальный статус участника при условии наличия точки отсечения c для x в некоторой малой области вокруг x. $W_i(x)$ равен единице для всех i, которые могут получать или получают воздействие, если точка разрыва c равна x. Это, вообще говоря, означает, что существует возможность манипуляций с точкой c. Например, если X – возраст, по которому индивид отбирается для участия в программе, то можно представить ситуацию, при которой возраст изменится таким образом, что для индивида право на участие в программе изменится с c на $c + \varepsilon$. В этом случае необходимо принять допущение о монотонности.

Допущение 2: $W_i(x)$ – невозрастающая функция в точке x при условии, что $x = c$.

Определим статус соответствия (complier). Данная концепция схожа с концепцией инструментальных переменных, предложенных в Imbens & Angrist (1994). Наблюдение имеет статус соответствия, если

$$\lim_{x \downarrow X_i} W_i(x) = 0, \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

Соответствующий индивид – это тот, кто попадет под воздействие, если пороговое значение установлено на уровне X_i или ниже, и не попадет, если пороговое значение – выше, чем X_i. Таким образом,

$$\tau_{FRD} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x \text{ и наблюдение } i \text{ имеет статус соответствия}].$$

Эта формула позволяет оценить средний эффект участия для объектов с $X_i = c$ и только для наблюдений со статусом соответствия.

Вернемся к примеру о назначении финансовой помощи студентам. Допустим, что присуждение стипендии основано не только на академической успеваемости, но и на национальном признаке поступающего таким образом, что студенты национальностей, не представленных в университете, всегда получают стипендию, в то время как остальные студенты могут получить стипендию только при условии, если они набрали пороговое или выше число баллов. В случае, если национальный статус студентов является ненаблюдаемой характеристикой, дизайн исследования будет нечеткий. Эффект воздействия в свою очередь будет оценен только для подгруппы студентов, которые набрали балл, близкий к пороговому значению. В нашем случае это подгруппа студентов, не являющихся национальным меньшинством.

4 Ситуации, когда разрывный дизайн таковым не является

Критическим вопросом, на который необходимо ответить исследователю, является вопрос о том, насколько значимо индивиды могут влиять на параметры, определяющие право на участие в программе. Если программа очень привлекательная, а индивиды могут манипулировать параметрами таким образом, чтобы получить возможность участвовать в программе, то очень вероятно, что индивиды по обе стороны от порогового значения будут существенно различаться.

Приведем пример. Представим что существуют два типа студентов, А и Б, проходящих тест. В нашем примере студенты А обладают большими способностями для успешной сдачи теста (50 процентов правильных ответов на вопросы письменного теста), чем студены Б. Предположим также, что студенты Б не заинтересованы в материальной помощи университета. Более того, предположим, что ответы на 50 процентов теста являются тривиальными и ошибки в ответах на эти вопросы могут быть допущены лишь случайно и исправлены при перепроверке перед сдачей результатов. В данном случае, если студенты типа А перед сдачей экзамена проверят свои работы, они тем самым могут гарантированно успешно пройти тест. Таким образом, и А-, и Б-студенты могут сдать экзамен, в то время как студенты, которые не сдадут экзамен, будут всегда только типа Б. Сравнение группы студентов, которые маржинально не сдали экзамен, и студентов, которые маржинально экзамен сдали, не будет являться корректным, и, следовательно, в этом случае применять разрывную регрессию неправильно.

Внесем изменения в предыдущий пример. Пусть задания для экзамена не являются тривиальными, и нет гарантии того, что студент сдаст экзамен, несмотря на то, что он его перепроверит перед сдачей. В данном случае распределение студентов на маржинально сдавших и не сдавших экзамен является случайным.

Безусловно, студенты типа A могут лучше подготовиться к экзамену, потому что они знают, что их стипендия зависит от этого, но вместе с тем они не могут гарантированно успешно сдать экзамен. В данном случае можно сравнивать группы студентов, которые маржинально сдали и не сдали экзамен, и применение разрывной регрессии оправдано.

Приведенные примеры свидетельствуют о том, что у исследователя должно присутствовать адекватное понимание механизма и экономического смысла функционирования программы. Основная идея состоит в том, что необходимо понять, имеют ли объекты полный контроль над тем, в какую группу они входят (контрольную или воздействия), или же они могут лишь влиять на свое пороговое значение, но при этом распределение между контрольной группой и группой воздействия все равно случайно в непосредственной близости от порогового значения.

5 Интерпретация разрывной регрессии

Как было видно из приведенного анализа, четкий и нечеткий дизайн дают в лучшем случае оценку среднего эффекта воздействия для подгруппы из группы воздействия, такую, что $X_i = c$. Более того, нечеткий дизайн дает оценку среднего эффекта воздействия только для подгруппы индивидов, имеющих статус соответствия. Без дополнительных допущений (к примеру, допущения об однородности эффекта) экстраполяция результатов на другие подгруппы невозможна. Разрывный дизайн ни в каком случае не позволит получить оценку среднего эффекта для генеральной совокупности. В этом смысле разрывный дизайн имеет достаточно ограниченную интерпретацию. Несмотря на это, средние эффекты воздействия, которые можно получить, корректно применяя разрывный дизайн, могут быть полезны лицам, принимающим управленческие, политические решения, в случаях, когда возникает необходимость изменений пороговых значений, к примеру, возраста выхода на пенсию. Главным же преимуществом применения разрывного дизайна по сравнению с другими методами является то, что он не основывается на допущении о несмешиваемости и, следовательно, имеет более высокую степень достоверности.

Благодарности

Автор благодарит Людмилу и Евгения Ниворожкиных за комментарии.

Список литературы

Вулдридж, Дж. (2009). Оценивание методом «разность разностей». *Квантиль* 6, 15–23.

Ениколопов, Р. (2009). Оценивание эффекта воздействия. *Квантиль* 6, 3–14.

Эббес, П. (2007). Инструментальные переменные и эндогенность: Нетехнический обзор. *Квантиль* 2, 3–20.

Angrist, J.D. & V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114, 533–575.

Black, S. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics* 114, 577–599.

Card, D., C. Dobkin & N. Maestas (2008). The impact of nearly universal insurance coverage on health care utilization: Evidence from Medicare. *American Economic Review* 98, 2242–2258.

Chen, S. & W. van der Klaauw (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics* 142, 757–784.

DiNardo, J. & D.S. Lee (2004). Economic impacts of new unionisation on private sector employers: 1984–2001. *Quarterly Journal of Economics* 142, 1383–1441.

Hahn, J., P. Todd, & W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 201–209.

Heckman, J.J., R.J. Lalonde & J.A. Smith (1999). The economics and econometrics of active labor market programs. Глава в *Handbook of Labor Economics*, том 3A (под редакцией Ashenfelter, O. & D. Card). Elsevier Science.

Heckman, J. & R. Robb (1985). Alternative methods for evaluating the impact of interventions. Глава в *Longitudinal Analysis of Labor Market Data* (под редакцией Heckman, J.J. & B.S. Singer). Cambridge University Press.

Imbens, G. & J.D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica* 61, 467–476.

Imbens, G. & T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142, 615–635.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics* 142, 785–806.

Lee, D.S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142, 675–697.

Lee, D.S. & T. Lemieux (2009). Regression discontinuity designs in economics. NBER Working Paper 14723.

Lemieux, T. & K. Milligan (2008). Incentive effects of social assistance: A regression discontinuity approach. *Journal of Econometrics* 119, 807–829.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66, 688–701.

Thistlethwaite, D.L. & D.T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology* 51, 309–317.

Trochim, W.M.K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach.* Beverly Hills: Sage Publications.

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review* 43, 1249–1287.

van der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour* 22, 219–245.

Regression discontinuity design

Anton Nivorozhkin

Institute for Employment Research (IAB), Nürnberg, Germany

This essay presents a brief review of the theory and empirical implementation of the discontinuity design for estimation of treatment effects of programs. Using published studies as examples, we consider the main difficulties in applications of the discontinuity design in practice and interpretation of results.

Эконометрический ликбез: ограниченные зависимые переменные

Мультиномиальные модели дискретного выбора*

Золт Шандор†

Университет Гронингена, Гронинген, Нидерланды

В настоящем эссе описаны основные свойства хорошо известных мультиномиальных моделей дискретного выбора: стандартной логит-модели, смешанной логит-модели и пробит-модели.

1 Введение

В эссе описываются некоторые модели дискретного выбора, применяемые для оценивания параметров спроса на товары, приобретаемые в дискретных количествах. В этих моделях различные товары рынка считаются различными выборными альтернативами. В интересующих нас моделях предполагается, что в течение заданного периода времени потребитель, считающийся принимающие решения лицом, приобретает либо одну, либо ноль единиц товара. Набор всех альтернатив, из которых выбирает потребитель, считается конечным и содержащим все товары, доступные на рынке. Обозначим альтернативу как j, а набор всех альтернатив как $\{1, 2, ..., J\}$. В модели дискретного выбора потребитель i, выбирающий альтернативу j, приобретает случайную полезность u_{ij}, которая напрямую исследователем не наблюдается. Во многих ситуациях, однако, вероятность того, что выбрана альтернатива j, наблюдается с небольшой выборочной ошибкой в виде доли рынка. Следовательно, обычно предполагается, что i максимизирует свою полезность, выбирая j таким образом, что u_{ij} наибольшая из всех u_{ir}, $r = 1, ..., J$. Вероятность того, что i выбирает j, можно вычислить, если мы знаем распределения, лежащие в основе функции полезности. В настоящем эссе описаны основные свойства хорошо известных мультиномиальных моделей дискретного выбора: стандартной логит-модели, смешанной логит-модели и пробит-модели.

Изначальным источником логита и пробита было приложение вероятностных моделей к биологическим опытам бинарного выбора в первой половине прошлого века. Первые приложения в экономике использовали биномиальную пробит-модель (например, Farrell, 1954), которая, по-видимому, является победителем в споре между биологами по поводу способности логита и пробита моделировать бинарный выбор. Theil (1969) обобщил биномиальную логит-модель до мультиномиальной логит-модели, открыв дорогу дальнейшим улучшениям и приложениям. В начале 70-х McFadden с соавторами, которые изучали некоторые исследовательские проблемы транспортировки, обобщили логит-модель в нескольких направлениях и сделали ее научно признанной, предоставив теоретическую основу в теории полезности дискретного выбора (например, McFadden, 1973; McFadden & Reid, 1975; McFadden, 1977). Более детальные исторические данные содержатся в Cramer (1991, стр. 39–42) и Anderson, de Palma & Thisse (1992, глава 2).

*Перевод М. Кузина и С. Анатольева. Цитировать как: Шандор, Золт (2009). «Мультиномиальные модели дискретного выбора», Квантиль, №7, стр. 9–19. Citation: Sándor, Zsolt (2009). "Multinomial discrete choice models," Quantile, No.7, pp. 9–19.

†Адрес: Department of Economics and Econometrics, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands. Электронная почта: *Z.Sandor@rug.nl*

2 Стандартная логит-модель

В этом разделе мы кратко представляем стандартную логит-модель, обсуждаем ее главные теоретические преимущества и практические недостатки. Затем мы выведем некоторые основные формулы, связанные с асимптотическими свойствами оценки максимального правдоподобия модели.

2.1 Спецификация модели

В стандартной логит-модели полезность является линейной функцией свойств альтернативы:

$$u_{ij} = x'_{ij}\beta + \varepsilon_{ij}, \tag{1}$$

где x_{ij} – $K \times 1$ вектор, содержащий характеристики потребителя i и альтернативы j, β – $K \times 1$ вектор параметров, а переменные ε_{ij}, $j = 1, ..., J$, предполагаются случайными и имеющими независимые стандартные распределения экстремальных значений типа I, кумулятивная функция распределения которого равна

$$F\left(\varepsilon_{ij}\right) = \exp\left(-\exp\left(-\varepsilon_{ij}\right)\right), \tag{2}$$

а функция плотности равна

$$f\left(\varepsilon_{ij}\right) = \exp\left(-\varepsilon_{ij}\right) \cdot \exp\left(-\exp\left(-\varepsilon_{ij}\right)\right). \tag{3}$$

В силу принципа максимизации полезности, вероятность того, что i выбирает альтернативу j, есть

$$s_{ij} = \Pr\left\{u_{ij} \geq u_{ir},\ \text{для всех } r = 1, ..., J\right\}.$$

Используя выражение для полезности (1), эта вероятность равна

$$s_{ij} = \Pr\left\{\varepsilon_{ir} \leq \varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta,\ \text{для всех } r = 1, ..., J\right\}.$$

Вероятности такого типа обычно вычисляют как J-мерный интеграл, используя совместное распределение случайных элементов:

$$s_{ij} = \int \cdots \int_{\substack{\varepsilon_{ir} \leq \varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta \\ \text{для всех } r=1,...,J}} f\left(\varepsilon_{i1}, ..., \varepsilon_{iJ}\right) \mathrm{d}\varepsilon_{i1} ... \mathrm{d}\varepsilon_{iJ}.$$

Так как все ε независимы, имеем

$$s_{ij} = \int \cdots \int_{\substack{\varepsilon_{ir} \leq \varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta \\ \text{для всех } r=1,...,J}} f\left(\varepsilon_{i1}\right) \cdots f\left(\varepsilon_{iJ}\right) \mathrm{d}\varepsilon_{i1} ... \mathrm{d}\varepsilon_{iJ},$$

что равно

$$\int_{\mathbb{R}} f\left(\varepsilon_{ij}\right) \int \cdots \int_{\substack{\varepsilon_{ir} \leq \varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta \\ r \neq j}} \prod_{r \neq j} f\left(\varepsilon_{ir}\right) \mathrm{d}\varepsilon_{i1} ... \mathrm{d}\varepsilon_{iJ}.$$

Наиболее крупная, $(J-1)$-мерная часть этого интеграла равна произведению $J-1$ одномерных интегралов плотностей, так что мы можем использовать кумулятивные функции распределения:

$$s_{ij} = \int_{\mathbb{R}} f\left(\varepsilon_{ij}\right) \cdot \prod_{r \neq j} F\left(\varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta\right) \mathrm{d}\varepsilon_{ij}.$$

Используя выражения для этих функций (2) и (3) , получаем

$$s_{ij} = \int_{\mathbb{R}} \exp\left(-\varepsilon_{ij}\right) \cdot \exp\left(-\sum_{r=1}^{J} \exp\left(-\left(\varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta\right)\right)\right) \mathrm{d}\varepsilon_{ij}.$$

Этот интеграл имеет явный вид и может быть вычислен следующим способом:

$$s_{ij} = \frac{1}{\sum_{r=1}^{J} \exp\left(-(x'_{ij}\beta - x'_{ir}\beta)\right)} \exp\left(-\sum_{r=1}^{J} \exp\left(-\left(\varepsilon_{ij} + x'_{ij}\beta - x'_{ir}\beta\right)\right)\right)\Bigg|_{\varepsilon_{ij}=-\infty}^{\infty},$$

что в итоге дает

$$s_{ij} = \frac{\exp(x'_{ij}\beta)}{\sum_{r=1}^{J} \exp\left(x'_{ir}\beta\right)}. \tag{4}$$

Благодаря этой привлекательной явной форме вероятности в (4) стандартная логит-модель и популярна в тех дисциплинах, в которых используются модели выбора.

В следующих разделах мы применим стандартную логит-модель в ситуациях, в которых характеристики варьируются по продуктам и не варьируются по потребителям, то есть когда $x_{ij} = x_j$ для всех $j = 1, ..., J$. Благодаря этому вероятность того, что потребитель i выбирает продукт j, равна

$$s_j = \frac{\exp(x'_j\beta)}{\sum_{r=1}^{J} \exp\left(x'_r\beta\right)} \tag{5}$$

для всех потребителей. Следовательно, это вероятность того, что продукт j приобретается на рынке. Эта величина на рынке с большим числом потребителей равно доле рынка продукта j.

Хотя выражение для вероятности того, что выбрана такая-то альтернатива, выглядит просто и потому доступно для практиков в вычислительном смысле, стандартная логит-модель также имеет серьезный недостаток. Этот недостаток возникает из-за простоты модели для полезности, а именно, из-за того, что полезности всех потребителей зависят от скалярной функции характеристик, $x'_j\beta$, а не от всего вектора характеристик x_j, поэтому отношение вероятностей двух альтернатив j и q не зависит от наличия и свойств любых других альтернатив, так как

$$\frac{s_j}{s_q} = \frac{\exp(x'_j\beta)}{\exp\left(x'_q\beta\right)}. \tag{6}$$

Это свойство известно как независимость от посторонних альтернатив. Это непривлекательное свойство, так как если мы добавляем новую альтернативу к набору всех альтернатив, которые являются близкими заменителями для j, но не для q, мы ожидаем, что s_j снизится значительно сильнее, чем s_q. Поэтому отношение двух вероятностей также должно уменьшиться.

Мы проиллюстрируем эту проблему на простом примере. Предположим, что на рынке есть два товара со следующими свойствами $x_1 = (1, 1, 1)'$ и $x_2 = (.5, 1.5, 1)'$, и это единственные альтернативы для выбора. Пусть $\beta = (1, 1, 1)'$. Тогда $s_1 = s_2 = \frac{1}{2}$. Теперь введем третий продукт со свойствами $x_3 = (1, 1, 1)'$. В этой новой ситуации $s_1 = s_2 = s_3 = \frac{1}{3}$, то есть стандартная логит-модель предсказывает, что потребители будут замещать продукты 1 и 2 в одинаковой степени. Это неверно, так как приблизительно половина потребителей предпочитает продукт 1, который является близким заменителем товару 2, так что часть потребителей, предпочитающих 1, должны переключиться на 3, а те, кто предпочитает 2, должны остаться с 2.

2.2 Логарифмическое правдоподобие и свойства ее первой производной

Так как далее мы используем результаты, связанные с оцениванием стандартной логит-модели, здесь мы приведем некоторые из этих результатов. Оценивание параметров стандартной логит-модели основано на идее о том, что мы находимся в ситуации мультиномиального выбора, когда вероятность s_j того, что выбрана альтернатива j, известна. Тогда если мы наблюдаем большое число выборов, частота альтернативы j, то есть число раз, когда j выбрана, деленное на общее число выборов, должно равняться s_j. Обозначим частоту альтернативы j за f_j. Пусть n – общее число выборов, то есть число потребителей. Тогда $n \cdot f \equiv n \cdot (f_1, ..., f_J)'$ имеет мультиномиальное распределение с параметрами $s_1, ..., s_J$. Поэтому вероятность того, что альтернативы $1, ..., J$ встречаются с частотами $f_1, ..., f_J$, соответственно, на самом деле равна функции правдоподобия, соответствующей стандартной логит-модели:

$$L = C \cdot \left(s_1^{f_1} \cdots s_J^{f_J}\right)^n,$$

где C – коэффициент, соответствующий мультиномиальному распределению. Так как этот коэффициент не зависит от интересующих нас параметров, в дальнейшем мы игнорируем его, когда записываем логарифмическую функцию правдоподобия, которая равна

$$\ln L = n \cdot \sum_{j=1}^{J} f_j \ln s_j. \tag{7}$$

Оценивание параметров можно выполнить максимизацией логарифмической функции правдоподобия. Обозначим за S диагональную матрицу с диагональю $s = (s_1, ..., s_J)'$. Можно записать $\ln L = n \cdot f' \ln s$. Следовательно, производная первого порядка логарифмической функции правдоподобия может быть записана как

$$\frac{\partial \ln L}{\partial \beta} = n \cdot \left(f' \frac{\partial \ln s}{\partial \beta'}\right)' = n \cdot \left(\frac{\partial s}{\partial \beta'}\right)' S^{-1} f,$$

так как

$$\frac{\partial \ln s}{\partial \beta'} = S^{-1} \frac{\partial s}{\partial \beta'}.$$

Для того чтобы вывести $\dfrac{\partial s}{\partial \beta'}$, заметим, что

$$\frac{\partial s_j}{\partial \beta'} = x_j' s_j - s' X s_j,$$

где $X = (x_1 \ ... \ x_J)'$, что приводит к

$$\frac{\partial s}{\partial \beta'} = \left(S - ss'\right) X. \tag{8}$$

Эту формулу можно использовать для того чтобы получить первые производные логарифмической функции правдоподобия:

$$\frac{\partial \ln L}{\partial \beta} = n \cdot X' \left(S - ss'\right) S^{-1} f.$$

Теперь можно вычислить асимптотическую информационную матрицу:

$$\mathcal{I}\left(\beta | X\right) = \mathbb{E}\left[\frac{\partial \ln L}{\partial \beta} \frac{\partial \ln L}{\partial \beta'}\right] = n^2 \cdot X' \left(S - ss'\right) S^{-1} \mathbb{E}\left[f f'\right] S^{-1} \left(S - ss'\right) X.$$

Мультиномиальное распределение имеет свойство $\mathrm{E}[ff'] = \frac{1}{n}S + \left(1 - \frac{1}{n}\right) ss'$. Подставляя это выражение в предыдущую формулу, после некоторых вычислений получаем

$$\mathcal{I}\left(\beta|X\right) = n \cdot X' \left(S - ss'\right) X.$$

Эта формула использовалась напрямую для построения модели с улучшенной эффективностью для экспериментов выбора (см. например, Sandor & Wedel, 2001).

3 Смешанная логит-модель

Смешанная логит-модель не удовлетворяет условию независимости от посторонних альтернатив. Это происходит из-за того, что в спецификации полезности коэффициенты предполагаются случайными. Таким образом, потребители имеют различные коэффициенты в функциях полезности. Такая спецификация имеет следующее привлекательное свойство: она параметризует отношение потребителей к характеристикам альтернатив через вариацию случайных параметров. Как в стандартной логит-модели, первые экономические приложения смешанной логит-модели, по-видимому, встретились в сфере транспортных исследований (Boyd & Mellman, 1980; Cardell & Dunbar, 1980). Мы отсылаем читателей к более детальным обзорам смешанных логит-моделей McFadden & Train (2000) и Brownstone & Train (1999). В данном разделе мы представим модель и основные асимптотические свойства оценки максимального правдоподобия.

3.1 Спецификация модели

В структуре выбора, аналогичной случаю стандартной логит-модели, полезность потребителя i, выбирающего альтернативу j, определяется как

$$u_{ij} = x_j'\beta_i + \varepsilon_{ij}, \tag{9}$$

где единственное отличие от (1) заключается в том, что здесь β_i – вектор параметров, характерных для потребителя i. Мы предполагаем, что $\beta_i \sim N(\beta, \Sigma)$, $i = 1, ..., n$, независимые и одинаково распределенные случайные величины, где β – $K \times 1$ вектор параметров, а Σ – симметричная положительно определенная матрица параметров. Ошибки ε_{ij} предполагаются независимыми от β_i.

Для удобства вычислений запишем случайный вектор параметров в следующем виде:

$$\beta_i = \beta + \Lambda v_i, \tag{10}$$

где Λ – некая матрица, такая, что $\Sigma = \Lambda\Lambda'$, а v_i – $K \times 1$ случайный вектор, имеющий стандартное нормальное распределение.

Если бы мы знали v_i, мы могли бы использовать формулу (5), поскольку это была бы стандартная логит-модель. В этом случае вероятность того, что альтернатива j выбрана, была бы

$$\Pr\{j|v_i\} = \frac{\exp\left(x_j'\left(\beta + \Lambda v_i\right)\right)}{\sum_{r=1}^{J}\exp\left(x_r'\left(\beta + \Lambda v_i\right)\right)}. \tag{11}$$

Безусловная вероятность того, что альтернатива j выбрана, определяется формулой

$$s_j = \mathbb{E}\left[\Pr\{j|v_i\}\right] = \int_{\mathbb{R}^K} \frac{\exp\left(x_j'\left(\beta + \Lambda v_i\right)\right)}{\sum_{r=1}^{J}\exp\left(x_r'\left(\beta + \Lambda v_i\right)\right)}\phi\left(v_i\right)\mathrm{d}v_i, \tag{12}$$

где $\phi(v_i)$ – функция плотности v_i. Заметим, что поскольку (6) в данном случае не выполняется, то смешанная логит-модель не удовлетворяет свойству независимости от посторонних альтернатив.

Выражение для вероятности в (12) представляет собой интеграл, который в общем случае нельзя вычислить аналитически. Обычно он оценивается с помощью симуляций Монте-Карло или квази-Монте-Карло. Детальное обсуждение технологии такого оценивания выходит за рамки данного эссе, поэтому мы лишь упомянем здесь ее суть. Извлекаем несколько значений случайного вектора v_i, вычисляем функцию под интегралом при каждом его значении, и находим среднее по этим значениям. Полученное среднее и есть (квази-) оценка интеграла методом Монте-Карло.

Чтобы проиллюстрировать, как в смешанной логит-модели моделируется отношение потребителей к различным характеристикам, рассмотрим более простой вид ковариационной матрицы, а именно, когда она диагональна:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_K^2 \end{pmatrix}.$$

В этом случае можно взять

$$\Lambda = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_K \end{pmatrix},$$

так что можно записать полезность в (9) с использованием (10) как

$$u_{ij} = x_j'\beta + \sum_{k=1}^{K} \sigma_k v_{ik} x_{jk} + \varepsilon_{ij}.$$

Записав полезность в таком виде, мы можем проинтерпретировать эффект случайной части параметров. Каждый потребитель i сопоставляется с вектором v_i, элементы которого представляют собой оценку потребителем характеристик. Если предпочтения i таковы, что для него более ценны альтернативы с большими значениями какой-то характеристики, то у такого потребителя будет большое положительное значение v_i, соответствующее этой характеристике. Например, в случае автомобилей, если потребитель i предпочитает большие машины, и «размер» – это характеристика h в этой модели, то этот потребитель будет иметь большое положительное v_{ih}.

Теперь можно объяснить преимущества смешанной логит-модели перед обычной тем, что схемы замен, порождаемые смешанной логит-моделью, подвергаются меньшим ограничениям. Продолжая пример с машинами, если на рынке появляется маленькая машина, то потребитель, о котором шла речь выше, не будет сильно стремиться к переключению на новую маленькую машину. Происходит это из-за того, что для этого потребителя значение $\sigma_h v_{ih}$ несоразмерно больше, чем соответствующие величины для других характеристик, и поэтому член $\sigma_h v_{ih} x_{jh}$ более чувствителен к вариациям в x_{jh}, чем в случае с другими характеристиками. Поэтому полезность потребителя i высока для больших значений x_{jh} и низка для малых.

Описанное выше подразумевает, что чем больше изменения параметров отношения потребителей, σ_k, тем менее вероятно, что потребитель, предпочитающий определенную альтернативу, будет переключаться на несхожую альтернативу. Проиллюстрируем это примером из конца раздела 2.1. В дополнение к случаю стандартной логит-модели здесь мы возьмем $(\sigma_1, \sigma_2, \sigma_3) = \lambda(1,1,1)$. Когда только товары 1 и 2 находятся на рынке, их вероятности $s_1 = s_2 = \frac{1}{2}$, независимо от значения λ. Когда товар 3 появляется на рынке, s_2 становится

0.35 при $\lambda = 1$, 0.42 при $\lambda = 5$, 0.47 при $\lambda = 10$ и 0.49 при $\lambda = 30$. Таким образом, для больших вариаций нет замещения товаром 3 товара 2, в то время как для малых вариаций замещение очень похоже на то, что возникало в стандартной логит-модели.

3.2 Логарифмическое правдоподобие и свойства ее первой производной

В этом разделе мы выведем некоторые полезные формулы для смешанной логит-модели. Обозначим условную вероятность в (11) за $\pi_j(v)$. Тогда

$$s_j = \int_{\mathbb{R}^K} \pi_j(v)\,\phi(v)\,\mathrm{d}v.$$

Модель выбора в данном случае можно сформулировать похожим образом, как для стандартной логит-модели. Тогда функция логарифмического правдоподобия в этой модели равна

$$\ln L = n \cdot \sum_{j=1}^{J} f_j \ln s_j,$$

где f_j вновь обозначает наблюдаемую частоту товара j.

Чтобы сократить обозначения, мы пишем интегралы типа $\int_{\mathbb{R}^K}(\cdot)\phi(v)\mathrm{d}v$ как $\int(\cdot)\mathrm{d}\Phi$ (например, $\int_{\mathbb{R}^K}\pi_j(v)\,\phi(v)\mathrm{d}v \equiv \int \pi_j \mathrm{d}\Phi$). При оценивании методом максимального правдоподобия мы находим первые производные функции логарифмического правдоподобия. Так как нам нужны эти результаты для случая, когда ковариационная матрица случайных коэффициентов Σ диагональна, мы выводим формулы только для этого случая. Для удобства вычислений мы изменили обозначение случайной части коэффициентов с Λv на $V\sigma$, где V – диагональная матрица с диагональю v, а $\sigma = (\sigma_1, ..., \sigma_K)'$. Аналогично случаю стандартной логит-модели,

$$\frac{\partial L}{\partial \beta} = n \cdot \left(\frac{\partial s}{\partial \beta'}\right)' S^{-1} f \quad \text{и} \quad \frac{\partial L}{\partial \sigma} = n \cdot \left(\frac{\partial s}{\partial \sigma'}\right)' S^{-1} f. \tag{13}$$

Далее нам потребуются следующие формулы:

$$\frac{\partial s}{\partial \beta'} = \int \frac{\partial \pi}{\partial \beta'}\mathrm{d}\Phi \quad \text{и} \quad \frac{\partial s}{\partial \sigma'} = \int \frac{\partial \pi}{\partial \sigma'}\mathrm{d}\Phi,$$

где $\pi = (\pi_1, ..., \pi_J)'$. Аналогично, как для стандартной логит-модели, получаем:

$$\frac{\partial \pi}{\partial \beta'} = \left(\Pi - \pi\pi'\right) X \quad \text{и} \quad \frac{\partial \pi}{\partial \sigma'} = \left(\Pi - \pi\pi'\right) XV,$$

где Π – диагональная матрица с диагональю π. Отсюда

$$\frac{\partial s}{\partial \beta'} = \int \left(\Pi - \pi\pi'\right) X\mathrm{d}\Phi \quad \text{и} \quad \frac{\partial s}{\partial \sigma'} = \int \left(\Pi - \pi\pi'\right) XV\mathrm{d}\Phi. \tag{14}$$

Подставляя эти равенства в (13), получаем:

$$\frac{\partial L}{\partial \beta} = n \cdot \int X' \left(\Pi - \pi\pi'\right) \mathrm{d}\Phi \cdot S^{-1} f \tag{15}$$

и

$$\frac{\partial L}{\partial \sigma} = n \cdot \int VX' \left(\Pi - \pi\pi'\right) \mathrm{d}\Phi \cdot S^{-1} f. \tag{16}$$

Асимптотическая информационная матрица определяется как

$$\mathcal{I}(\beta, \sigma | X) = \begin{bmatrix} \mathbb{E}\left[\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \beta'}\right] & \mathbb{E}\left[\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \sigma'}\right] \\ \mathbb{E}\left[\frac{\partial L}{\partial \sigma}\frac{\partial L}{\partial \beta'}\right] & \mathbb{E}\left[\frac{\partial L}{\partial \sigma}\frac{\partial L}{\partial \sigma'}\right] \end{bmatrix}.$$

Для упрощения этого выражения используем тот факт, что по предположению f имеет мультиномиальное распределение с $\mathbb{E}[f] = s$. Тогда f обладает свойством $\mathbb{E}[ff'] = \frac{1}{n}S + \left(1 - \frac{1}{n}\right) ss'$. Используя этот факт и формулы (15) и (16), находим:

$$\mathcal{I}\left(\beta, \sigma | X\right) = n \cdot \begin{bmatrix} M'S^{-1}M & M'S^{-1}Q \\ Q'S^{-1}M & QS^{-1}Q \end{bmatrix}, \tag{17}$$

где

$$M = \int \left(\Pi - \pi\pi'\right) X \mathrm{d}\Phi \text{ и } Q = \int \left(\Pi - \pi\pi'\right) XV \mathrm{d}\Phi.$$

Формула (17) использовалась напрямую для построения экспериментов выбора для смешанной логит-модели (см. например, Sandor & Wedel, 2002).

4 Пробит

Пробит – это модель, похожая на смешанную логит-модель, в которой все случайные элементы, коэффициенты и ошибки по предположению имеют нормальное распределение. Одно из первых применений мультиномиальной пробит-модели в эконометрическом контексте встретилось в работе McFadden (1976). В этом разделе мы кратко представим модель и покажем способ оценивания вероятностей выбора, известный как GHK-симулятор (по первым буквам Geweke, Hajivassiliou и Keane, разработчиков симулятора для эконометрического контекста в начале 90-х). В отличие от стандартной и смешанной логит-моделей, для пробит-модели мы не приводим функцию логарифмического правдоподобия и свойства ее первых производных. Их краткое описание содержится в Wansbeek & Webel (2001).

4.1 Спецификация модели

Предположим, что полезности, соответствующие J продуктам для потребителя i, равны

$$u_{ij} = x_j'\beta_i + \varepsilon_{ij},$$

где, как и для смешанной логит-модели, $\beta_i \sim N(\beta, \Sigma)$, $i = 1, ..., n$, независимо и одинаково распределены, β – $K \times 1$ вектор параметров, а Σ – симметричная положительно определенная матрица параметров. Ошибка ε_{ij} здесь предполагается нормально распределенной без требования независимости от β_i. Благодаря привлекательному свойству нормального распределения, а именно, тому, что сумма двух нормально распределенных случайных величин нормально распределена, можно выбрать подходящие переменные μ и Ω таким образом, что вектор полезностей можно представить как

$$u_i = \mu + \Gamma e_i \quad \sim N\left(\mu, \Omega\right), \tag{18}$$

где $\mu : J \times 1$, $\Gamma : J \times J$, такая что $\Gamma\Gamma' = \Omega$, и $e_i \sim N\left(0, I_J\right)$. Если случайные члены e_i независимы, то можно опустить индекс i, и каждому потребителю сопоставить реализацию случайного вектора e. Отметим, что если в полезности в смешанной логит-модели (9) мы заменим предположение о распределении экстремальных значений остатков нормальным распределением, то получим частный случай (18). Причина, по которой последняя спецификация более общая, заключается в том, что в данном случае не предполагается независимости между остатками и случайными коэффициентами.

Из J продуктов выбран один с наибольшей полезностью. Вероятность того, что был выбран продукт j, равна

$$s_j = \Pr\left\{u_j \geq u_r, \forall r\right\} = \Pr\left\{V_j u \leq 0\right\} = \Pr\left\{V_j\mu + V_j\Gamma e \leq 0\right\}$$

с

$$V_j = \begin{bmatrix} 1 & 0 & \dots & -1 & \dots & 0 \\ 0 & 1 & \dots & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & -1 & \dots & 1 \end{bmatrix},$$

где столбец, состоящий из -1, стоит на j-ом месте. Обозначим $T \equiv V_j \Gamma$ и выберем Γ так, что матрица T нижняя треугольная. Это возможно сделать, так как $TT' = V_j \Omega V_j'$, а эта матрица положительно определена. Тогда

$$s_j = \Pr\{Te \le v\}, \text{ где } v \equiv -V_j \mu. \tag{19}$$

Если мы обозначим

$$T = \begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{12} & t_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ t_{J1} & t_{J2} & \dots & t_{JJ} \end{bmatrix},$$

то

$$s_j = \Pr\left\{e_1 \le \frac{v_1}{t_{11}}, e_2 \le \frac{v_2 - t_{21}e_1}{t_{22}}, \dots,\ e_J \le \frac{v_J - t_{J1}e_1 - \dots - t_{JJ-1}e_{J-1}}{t_{JJ}}\right\}.$$

Пусть

$$D = \left\{(e_1, ..., e_J) \in \mathbb{R}^J : e_1 \le \frac{v_1}{t_{11}}, ..., e_J \le \frac{v_J - t_{J1}e_1 - \dots - t_{JJ-1}e_{J-1}}{t_{JJ}}\right\}.$$

Тогда формула для вероятности того, что выбрана альтернатива j, выглядит как

$$s_j = \int_D \phi(e_1) \dots \phi(e_J)\, \mathrm{d}e_1 \dots \mathrm{d}e_J, \tag{20}$$

где ϕ – плотность стандартного нормального распределения.

4.2 Оценивание вероятностей

Чтобы получить более удобные в вычислительном смысле формулы вероятности, заметим, что e имеют усеченное нормальное распределение. Если $e \le b$ и $e \sim N(0,1)$, тогда можно получить e быстрее, извлекая u равномерно распределенным на $[0,1]$ и беря $e = \Phi^{-1}(u \cdot \Phi(b))$, где Φ – стандартная нормальная функция распределения. Оказывается, что основываясь на этой идее, можно трансформировать интеграл (20) в интеграл, аргументы которого принадлежат единичному гиперкубу.

Для того чтобы произвести трансформацию, можно воспользоваться идеей из предыдущего раздела в J-мерном случае. Это подразумевает трансформацию

$$\begin{aligned} e_1 &= \Phi^{-1}\left(u_1 \Phi\left(\frac{v_1}{t_{11}}\right)\right) \equiv \psi_1(u_1, ..., u_J), \\ e_2 &= \Phi^{-1}\left(u_2 \Phi\left(\frac{v_2 - t_{21}e_1(u_1)}{t_{22}}\right)\right) \equiv \psi_2(u_1, ..., u_J), \\ &\vdots \\ e_J &= \Phi^{-1}\left(u_J \Phi\left(\frac{v_J - t_{J1}e_1 - \dots - t_{JJ-1}e_{J-1}}{t_{JJ}}\right)\right) \equiv \psi_3(u_1, ..., u_J), \end{aligned} \tag{21}$$

где $u_1, ..., u_J \in [0,1]$, и в последней формуле e_j рассматривается как функция от $u_1, ..., u_j$ для $j = 1, ..., J-1$. Тогда (20) становится

$$\int_{[0,1]^J} \phi\left(\psi_1\left(u_1, ..., u_J\right)\right) ... \phi\left(\psi_J\left(u_1, ..., u_J\right)\right) |J| \, \mathrm{d}u_1 ... \mathrm{d}u_J,$$

где $|J|$ – якобиан, то есть

$$|J| = \begin{vmatrix} \frac{\partial \psi_1}{\partial u_1} & \cdots & \frac{\partial \psi_J}{\partial u_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_1}{\partial u_J} & \cdots & \frac{\partial \psi_J}{\partial u_J} \end{vmatrix} = \begin{vmatrix} \frac{1}{\phi(\psi_1(...))}\Phi\left(\frac{v1}{t_{11}}\right) & \cdots & \frac{\partial \psi_J}{\partial u_1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\phi(\psi_J(...))}\Phi\left(\frac{v_J - t_{J1}e_1 - ... - t_{JJ-1}e_{J-1}}{t_{JJ}}\right) \end{vmatrix},$$

что является определителем верхней треугольной матрицы, поэтому якобиан равен произведению диагональных элементов:

$$|J| = \frac{\Phi\left(\frac{v_1}{t_{11}}\right) \cdots \Phi\left(\frac{v_J - t_{J1}e_1 - ... - t_{JJ-1}e_{J-1}}{t_{JJ}}\right)}{\phi\left(\psi_1\left(...\right)\right) \cdots \phi\left(\psi_J\left(...\right)\right)}.$$

Используя полученные формулы, можно преобразовать (20) в

$$\int_{[0,1]^J} \Phi\left(\frac{v_1}{t_{11}}\right) \cdots \Phi\left(\frac{v_J - t_{J1}e_1 - ... - t_{JJ-1}e_{J-1}}{t_{JJ}}\right) \mathrm{d}u_1 ... \mathrm{d}u_J.$$

Тогда получаем формулу для вероятности

$$s_j = \Phi\left(\frac{v_1}{t_{11}}\right) \int_{[0,1]^{J-1}} \Phi\left(\frac{v_2 - t_{21}e_1}{t_{22}}\right) \cdots \Phi\left(\frac{v_J - t_{J1}e_1 - ... - t_{JJ-1}e_{J-1}}{t_{JJ}}\right) \mathrm{d}u_1 ... \mathrm{d}u_{J-1}.$$

Интеграл в этой формуле можно интерпретировать как ожидание от подынтегральной функции при равномерно распределенном случайном $(J-1)$-векторе. Этот интеграл можно оценить с помощью симуляций Монте-Карло, извлекая набор случайных векторов и вычисляя их среднее. Оценка, найденная таким способом, называется GHK-симулятором (Börsch-Supan & Hajivassiliou, 1993) или RIS-симулятором (от *recursive importance sampling*), основанным на усеченной нормальной плотности (Vijverberg, 1997). Vijverberg (1997) также обсуждает другие RIS-симуляторы. Ряд других типов симуляторов представлен в работе Hajivassiliou, McFadden & Ruud (1996).

Благодарности

Автор благодарит Тома Вансбеека за многочисленные полезные комментарии.

Список литературы

Anderson, S.P., A. de Palma & J-F. Thisse (1992). *Discrete Choice Theory of Product Differentiation.* Cambridge: MIT Press.

Börsch-Supan, A. & V.A. Hajivassiliou (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* 58, 347–368.

Boyd, H.J. & R.E. Mellman (1980). The effect of fuel economy standards on the U.S. automotive market: An hedonic demand analysis. *Transportation Research* 14, 367–378.

Brownstone, D. & K. Train (1999). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics* 89, 109–129.

Cardell, N.S. & F. Dunbar (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research* 14, 423–434.

Cramer, J.S. (1991). *The Logit Model: An Introduction for Economists.* London: Arnold.

Farrell, M.J. (1954). The demand for motorcars in the United States. *Journal of the Royal Statistical Society, A* 117, 171–200.

Hajivassiliou, V., D. McFadden & P. Ruud (1996). Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics* 72, 85–134.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. Глава в *Frontiers in Econometrics* под редакцией P. Zarembka. New York: Academic Press.

McFadden, D. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement* 5, 363–90.

McFadden, D. (1977). Econometric models of probabilistic choice. Стр. 171–260 в *Structural Analysis of Discrete Data with Econometric Applications* под редакцией C.F. Manski & D. McFadden. Cambridge: MIT Press.

McFadden, D. & F. Reid (1975). Aggregate travel demand forecasting from disaggregated behavioral models. *Transportation Research Record* 534, 24–37.

McFadden, D. & K. Train (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15, 447–470.

Sándor, Z. & M. Wedel (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research* 38, 430–444.

Sándor, Z. & M. Wedel (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science* 21, 55–75.

Theil, H. (1969). A multinomial extension of the linear logit model. *International Economic Review* 10, 251–259.

Vijverberg, W.P.M. (1997). Monte Carlo evaluation of multivariate normal probabilities. *Journal of Econometrics* 76, 281–307.

Wansbeek, T. & M. Wedel (2001). The structure of the multinomial probit Model. Working Paper, University of Groningen.

Multinomial discrete choice models

Zsolt Sándor

University of Groningen, Groningen, Netherlands

This essay briefly describes the main features of some well-known multinomial discrete choice models: the standard logit, the mixed logit, and the probit.

Заметки о моделях с самоотбором выборки*

Виктор Агиррегабирия†

Университет Торонто, Торонто, Канада

Проблемы, связанные с самоотбором выборки, часто встречаются при работе с микроэкономическими моделями и данными по индивидам, домашним хозяйствам или фирмам. За последние тридцать лет в этой области эконометрики были сделаны весьма значительные достижения. Предложены и применены на практике различные типы моделей. Разработаны новые методы оценивания и инференции, как параметрические, так и полупараметрические. Настоящее эссе является кратким введением в эту обширную литературу.

1 Введение

Рассмотрим модель регрессии:

$$Y^* = X^*\beta + \varepsilon, \tag{1}$$

где Y^* и ε – скалярные случайные величины, X^* – $1\times K$ вектор случайных величин, а β – $K\times 1$ вектор параметров. Случайная ошибка ε независима в среднем от X^*, и матрица $\mathbb{E}[X^{*\prime}X^*]$ имеет полный ранг. Тогда при наличии случайной выборки для (Y^*, X^*) МНК-оценка параметров состоятельна и асимптотически нормальна. Ключевая особенность моделей с самоотбором выборки состоит в том, что исследователь не наблюдает случайную выборку для (Y^*, X^*). Вместо нее имеется случайная выборка для пары переменных (Y, X), которые связаны с (Y^*, X^*), но отличаются от них. Переменные (Y^*, X^*) называют латентными. Задача заключается в состоятельном оценивании β по выборке для (Y, X).[1] В зависимости от соотношения между латентными и наблюдаемыми переменными выделяют различные классы моделей с самоотбором выборки. Далее запись $Y = \{X|Z > c\}$ означает, что Y – это случайная величина X при условии, что случайная величина Z больше константы c. Аналогично определяются выражения $Y = \{X|Z < c\}$ и $Y = \{X|b < Z < c\}$.

(a) *Модель усеченной регрессии.* Пусть c – известная константа. Если переменная Y усечена слева в точке c, то

$$(Y, X) = \{(Y^*, X^*)|Y^* > c\}. \tag{2}$$

Если переменная Y усечена справа в точке c, то

$$(Y, X) = \{(Y^*, X^*)|Y^* < c\}. \tag{3}$$

В обоих случаях случайная выборка для пары (Y, X) не является случайной выборкой ни для Y^*, ни для X^*.[2]

*Перевод Б. Гершмана. Цитировать как: Агиррегабирия, Виктор (2009). «Заметки о моделях с самоотбором выборки», Квантиль, №7, стр. 21–36. Citation: Aguirregabiria, Victor (2009). "Some notes on sample selection models," Quantile, No.7, pp. 21–36.

†Адрес: 150 St. George Street, Toronto, ON, M5S 3G7. Электронная почта: *victor.aguirregabiria@utoronto.ca*

[1] В некоторых приложениях интерес также представляет оценивание функции распределения случайной ошибки ε.

[2] Случайная выборка для пары (Y, X) приводит к случайной выборке для X^* только в случае, когда Y^* и X^* независимо распределены.

Пример 1. Рассмотрим уравнение для логарифма заработной платы, $W^* = X^*\beta + \varepsilon$, где W^* – логарифм заработной платы индивида, а X^* – вектор наблюдаемых характеристик его человеческого капитала. Предположим, что из-за конфиденциальности имеющаяся в распоряжении база данных не содержит информацию (ни о заработной плате, ни о прочих характеристиках) по индивидам с почасовой заработной платой более 800 долл. Тогда наблюдаются переменные (W, X), такие, что $(W, X) = \{(W^*, X^*)|W^* < \ln 800\}$. В этом случае говорят, что зависимая переменная усечена справа, и рассматривают модель усеченной регрессии, поскольку ни W^*, ни X^* не наблюдаются, когда заработная плата превышает 800 долл. в час.

Пусть f_{Y^*} и F_{Y^*} – функция плотности распределения (ФПР) и кумулятивная функция распределения (КФР) случайной величины Y^*, соответственно. Если переменная Y усечена слева в точке c, то ФПР Y имеет вид

$$f_Y(y) = \begin{cases} 0 & \text{при} \quad y \le c, \\ \dfrac{f_{Y^*}(y)}{1 - F_{Y^*}(c)} & \text{при} \quad y > c. \end{cases} \tag{4}$$

Если переменная Y усечена справа в точке c, то

$$f_Y(y) = \begin{cases} \dfrac{f_{Y^*}(y)}{F_{Y^*}(c)} & \text{при} \quad y < c, \\ 0 & \text{при} \quad y \ge c. \end{cases} \tag{5}$$

На рисунке 1 представлены ФПР нормальных случайных величин, усеченных слева и справа.

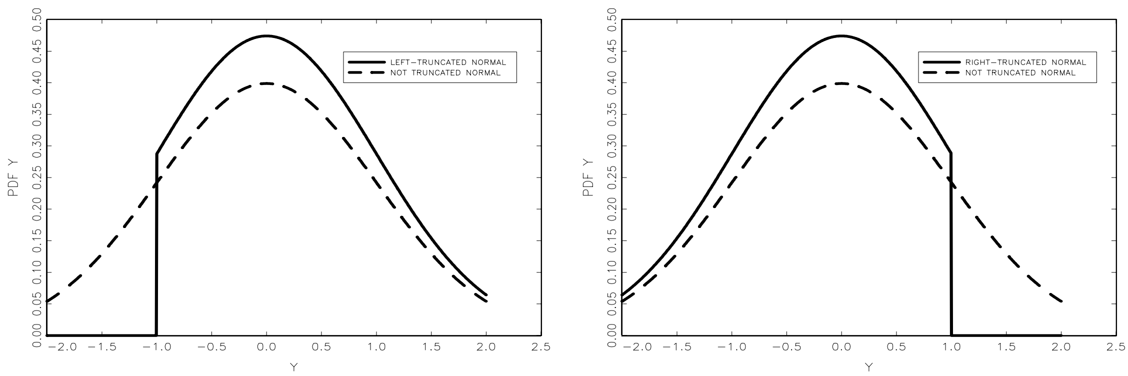


Рис. 1: Нормальные случайные величины, усеченные слева и справа, соответственно.

(b) *Модель цензурированной регрессии (или тобит-модель).* Основное отличие этой модели от модели усеченной регрессии состоит в том, что имеется случайная выборка для экзогенных регрессоров X^*. То есть случайные переменные X и X^* совпадают. Что касается зависимой переменной, то при цензурировании слева

$$Y = \max\{Y^*; c\} = \begin{cases} c & \text{при} \quad Y^* \le c, \\ Y^* & \text{при} \quad Y^* > c. \end{cases} \tag{6}$$

При цензурировании справа

$$Y = \min\{Y^*; c\} = \begin{cases} Y^* & \text{при} \quad Y^* < c, \\ c & \text{при} \quad Y^* \ge c. \end{cases} \tag{7}$$

Пример 2. Рассмотрим уравнение для логарифма заработной платы из примера 1. Теперь в распоряжении исследователя другой набор данных. Он включает информацию по всем индивидам независимо от уровня дохода. Доступна случайная выборка индивидов, содержащая информацию о переменных X. Но из-за конфиденциальности данные по наиболее высоким заработным платам недоступны. Если индивид имеет почасовую заработную плату менее 800 долл., наблюдается ее реальное значение. Но для индивидов, зарабатывающих более 800 долл. в час, данные о заработной плате отсутствуют. Следовательно, для каждого индивида в выборке наблюдается цензурированное справа значение логарифма заработной платы $W = \min\{W^*; \ln 800\}$. Зависимая переменная цензурирована справа, поэтому рассматривается модель цензурированной регрессии.

Пример 3. Рассмотрим следующую модель инвестиционных вложений фирмы в определенный вид оборудования, например, компьютеры. Пусть Q^* – «желаемый» объем инвестиций фирмы в соответствии с некоторой экономической моделью оптимальных инвестиций, например, объем инвестиций, максимизирующий прибыль без ограничения на неотрицательность Q^*, то есть $Q^* = \arg\max_q \Pi(q)$, где $\Pi(q)$ – (межвременная) функция прибыли. Предположим, что из этой модели следует следующее уравнение регрессионного типа: $Q^* = X\beta + \varepsilon$. Вектор X включает характеристики фирмы и рынка капитала, на котором действует фирма, такие как запас оборудования и цена нового капитала. β – вектор параметров, имеющих четкую экономическую интерпретацию в рамках модели. Имеется случайная выборка фирм, для которых наблюдается X и объем инвестиций Q. Глядя на эмпирическое распределение объема инвестиций Q, становится очевидным, что эта переменная всегда положительна с некоторой вероятностной массой в нуле. Эти свойства распределения нельзя объяснить предыдущей моделью регрессии, если не делать необоснованных предположений о распределении ε. Более того, рассмотренная теоретическая модель предполагает, что объем инвестиций Q^* может быть как положительным, так и отрицательным, а это противоречит наблюдаемым значениям Q. Рассмотрим тогда следующую модель для Q: $Q = \arg\max_q \Pi(q)$ при ограничении $q \geq 0$. Если функция прибыли $\Pi(q)$ строго вогнута, легко показать, что $Q = Q^*$ при $Q^* > 0$ и $Q = 0$ при $Q^* \leq 0$. То есть $Q = \max\{Q^*; 0\}$, где $Q^* = X\beta + \varepsilon$. С экономической точки зрения эту модель можно интерпретировать как модель необратимых инвестиций. С эконометрической точки зрения, это модель цензурированной регрессии.

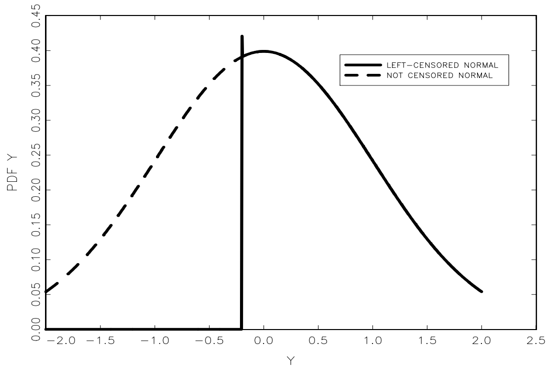
Примеры 2 и 3 представляют две различные модели цензурированной регрессии. Интересно отметить некоторые существенные различия между этими двумя примерами. Они основаны на весьма разных экономических и статистических предпосылках. В примере 2 цензурирование является свойством выборки. Заработная плата индивидов, превышающая 800 долл. в час, не является теоретическим объектом, а реально существует, хоть и не наблюдается в выборке. В примере 3 цензурирование – это предположение модели. Принимая во внимание определенные свойства распределения объема инвестиций, предполагается, что для этой переменной разумно рассматривать модель цензурированной регрессии. Переменная Q^* представляет собой теоретический объект, и для нее невозможно получить случайную выборку. Тем не менее параметры β могут иметь четкую экономическую интерпретацию в рамках этой модели и представляют интерес.

Пусть f_{Y^*} и F_{Y^*} – ФПР и КФР случайной величины Y^*, соответственно. Если переменная Y цензурирована слева в точке c, то ее ФПР имеет вид

$$f_Y(y) = \begin{cases} 0 & \text{при} \quad y < c, \\ F_{Y^*}(c) & \text{при} \quad y = c, \\ f_{Y^*}(y) & \text{при} \quad y > c. \end{cases} \tag{8}$$

Если переменная Y цензурирована справа в точке c, то

$$f_Y(y) = \begin{cases} f_{Y^*}(y) & \text{при} \quad y < c, \\ 1 - F_{Y^*}(c) & \text{при} \quad y = c, \\ 0, & \text{при} \quad y > c. \end{cases} \tag{9}$$



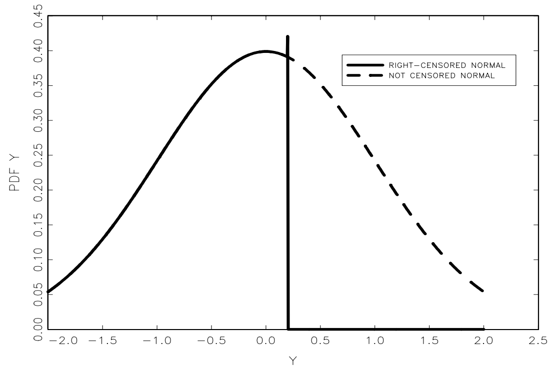


Рис. 2: Нормальные случайные величины, цензурированные слева и справа, соответственно.

На рисунке 2 представлены ФПР нормальных случайных величин, цензурированных слева и справа.

(c) *Модель с самоотбором выборки.* В простой модели с самоотбором выборки Y^* наблюдается только для индивидов, для которых определенная бинарная переменная, D, равна единице. Эта бинарная переменная не является независимой от Y^*.

$$Y = \{Y^* | D = 1\} \tag{10}$$

и ФПР$(Y^*|D = 1) \neq$ ФПР$(Y^*|D = 0)$. Заметим, что если D и Y^* независимы, то случайные переменные Y и Y^* совпадают и проблема самоотбора выборки отсутствует. Есть два вида моделей с самоотбором выборки: усеченного типа и цензурированного типа. В модели усеченного типа X^* также не наблюдается при $D = 0$. То есть в модели усеченного типа

$$(Y, X) = \{(Y^*, X^*) | D = 1\}. \tag{11}$$

В модели цензурированного типа имеется случайная выборка для X^* (то есть $X = X^*$). Тогда

$$(Y, X) = (\{Y^* | D = 1\}, X^*). \tag{12}$$

В такой модели цензурированного типа иногда удобно определить Y следующим образом: $Y = Y^*$ при $D = 1$, и $Y = 0$ при $D = 0$. Или в более компактной форме: $Y = DY^*$. Заметим, что модели усеченной и цензурированной регрессии являются частными случаями модели с самоотбором выборки. При $D = \mathbb{I}\{Y^* > c\}$ модель с самоотбором выборки становится моделью регрессии, усеченной/цензурированной слева, а при $D = \mathbb{I}\{Y^* < c\}$ получаем модель регрессии, усеченную/цензурированную справа.

Пример 4. Рассмотрим снова уравнение для логарифма заработной платы из примеров 1 и 2. Однако теперь интерес представляет не исключительно популяция работающих индивидов, а все индивиды, составляющие рабочую силу, занятые и нет. В таком случае W^* интерпретируется как латентная рыночная заработная плата индивида, и она существует независимо от того, работает индивид или нет. Имеется случайная выборка занятых и незанятых индивидов. Таким образом, есть случайная выборка для X^*, и рассматривается модель с самоотбором выборки цензурированного типа. Но рыночная заработная плата W^* наблюдается только для занятых индивидов. Пусть D – индикатор события «индивид работает». Тогда имеется случайная выборка для переменной W, где $W = \{W^*|D = 1\}$. Индикатор занятости D зависит от разных факторов, включая характеристики человеческого капитала, наблюдаемые и не наблюдаемые эконометристом. Следовательно, D и W^* не являются независимыми, и возникает проблема самоотбора выборки.

Спецификация модели с самоотбором выборки должна включать некоторые предположения о совместном распределении Y^* и D. Распространенная спецификация имеет вид

$$D = \mathbb{I}\{Z\gamma - u > 0\}, \tag{13}$$

где $\mathbb{I}\{\cdot\}$ – индикаторная функция, Z – вектор наблюдаемых переменных, γ – вектор параметров, и u не наблюдается. Переменные (X, Z) экзогенны, то есть независимы от случайных величин (u, ε). Условно на (X, Z) ненаблюдаемые величины u и ε не являются независимо распределенными.

(d) *Обобщенная модель с самоотбором выборки.* Рассмотрим следующую систему J линейных уравнений:

$$\begin{array}{rcl} Y_1^* & = & X^*\beta_1 + \varepsilon_1, \\ Y_2^* & = & X^*\beta_2 + \varepsilon_2, \\ \vdots & & \vdots \\ Y_J^* & = & X^*\beta_J + \varepsilon_J. \end{array} \tag{14}$$

Предположим, что наблюдается случайная выборка для X^*, то есть имеет место модель с самоотбором выборки цензурированного типа, где $X = X^*$.[3] Однако для каждого индивида в выборке не наблюдается весь набор J зависимых переменных $(Y_1^*, Y_2^*, ..., Y_J^*)$. Вместо этого, для каждого индивида наблюдается дискретная переменная $D \in \{1, 2, ..., J\}$ и зависимая переменная Y, такая, что

$$Y = \sum_{j=1}^{J} \mathbb{I}\{D = j\} Y_j^*. \tag{15}$$

Каждый индивид наблюдается только в одном *режиме*. Важно, что дискретная переменная D не является независимой от случайных ошибок ε_j в системе линейных уравнений.

Пример 5 (Модель Роя)[4]. Рассмотрим индивида, выбирающего между двумя возможными занятиями, 1 и 2. Предположим, что этот индивид выбирает работу, которая дает наибольший (за всю жизнь) заработок. При данных наблюдаемых и ненаблюдаемых характеристиках индивида доходы от двух занятий равны

$$\begin{array}{l} W_1^* = X\beta_1 + \varepsilon_1, \\ W_2^* = X\beta_2 + \varepsilon_2. \end{array} \tag{16}$$

Вектор X содержит наблюдаемые характеристики человеческого капитала, такие как образование и опыт работы. Векторы параметров β_1 и β_2 измеряют отдачу от характеристик человеческого капитала в занятиях 1 и 2, соответственно. ε_1 и ε_2 представляют собой отдачу от ненаблюдаемых (эконометристом, но не индивидом) характеристик человеческого капитала. Каждый индивид имеет только одно занятие. Пусть D – индикатор события «индивид выбирает занятие 1». Тогда наблюдаемый заработок индивида, W, можно представить в виде

$$W = DW_1^* + (1 - D)W_2^*. \tag{17}$$

При предположении, что индивиды максимизируют доход, получаем, что

$$D = \mathbb{I}\{W_1^* > W_2^*\} = \mathbb{I}\{X(\beta_1 - \beta_2) - (\varepsilon_2 - \varepsilon_1) > 0\}. \tag{18}$$

Ясно, что ненаблюдаемая величина $\varepsilon_2 - \varepsilon_1$ в уравнении для бинарной переменной выбора не является независимой от ненаблюдаемых величин в уравнениях для доходов, ε_1 и ε_2. В этих

[3] Можно также рассматривать версию этой модели, в которой переменная X^* усечена в некотором режиме $j \in \{1, 2, ..., J\}$.

[4] См. Roy (1951) и Heckman & Honoré (1990).

условиях требуется по случайной выборке индивидов с характеристиками X и заработными платами W оценить параметры β_1 и β_2.

Пример 6 (Эффекты воздействия). Задача состоит в оценке воздействия программы субсидирования инвестиций на объем капитальных инвестиций со стороны фирм. Пусть Q_1^* и Q_0^* – объемы инвестиций, осуществляемых фирмой при наличии воздействия (субсидии) и при его отсутствии, соответственно. Q_1^* и Q_0^* – латентные переменные. *Эффект воздействия* (ЭВ) для отдельной фирмы определяется как $TE = Q_1^* - Q_0^*$. Интерес представляет оценка *среднего эффекта воздействия* (СЭВ), который определяется как $ATE = \mathbb{E}[Q_1^* - Q_0^*]$. Также интерес может представлять условный средний эффект воздействия, $ATE(X) = \mathbb{E}[Q_1^* - Q_0^*|X]$, где X – вектор экзогенных характеристик фирмы. Имеется случайная выборка фирм. Каждая фирма наблюдается только один раз, либо при наличии воздействия ($D = 1$), либо при его отсутствии ($D = 0$). То есть при $D = 1$ наблюдается $Q = Q_1^*$, а при $D = 0$ наблюдается $Q = Q_0^*$. Обычно участие в программе субсидирования не является полностью случайным. Исследователь не располагает идеальными экспериментальными данными. Бинарная переменная воздействия D зависит от наблюдаемых характеристик Z и ненаблюдаемой величины u, которая может коррелировать с Q_0^* или/и Q_1^*. Необходимо, используя доступную выборку для (Q, D, X, Z), состоятельно оценить эффект программы предоставления субсидий на объем инвестиций, измеренный безусловным или условным средним эффектом воздействия.

Пример 7 (Модель инвестиций с издержками приспособления). Рассмотрим модель инвестиций в капитал, похожую на модель из примера 3. Теперь инвестиции не являются полностью необратимыми, то есть фирмы могут дезинвестировать или продавать подержанный капитал. Пусть K_t обозначает запас капитала фирмы, который продуктивен в момент времени t. Пусть $\Pi(K_t, K_{t-1})$ – (межвременная) функция прибыли. Прибыль зависит как от K_t, так и от K_{t-1} из-за наличия издержек приспособления. А именно, имеет место асимметрия между ценой нового капитала и ценой подержанного капитала, или, иными словами, между стоимостью капитала при $K_t > K_{t-1}$ и при $K_t < K_{t-1}$.

$$\Pi(K_t, K_{t-1}) = \begin{cases} \Pi^{(+)}(K_t, K_{t-1}) & \text{при} \quad K_t \geq K_{t-1}, \\ \Pi^{(-)}(K_t, K_{t-1}) & \text{при} \quad K_t \leq K_{t-1}. \end{cases} \tag{19}$$

Функции $\Pi^{(+)}$ и $\Pi^{(-)}$ непрерывны, дифференцируемы и строго вогнуты по K_t. Функция прибыли Π всюду непрерывна, но имеет излом (в котором она недифференцируема) в точке $K_t = K_{t-1}$. Определим $K_t^{(+)} \equiv \arg\max_k \Pi^{(+)}(k, K_{t-1})$ и $K_t^{(-)} \equiv \arg\max_k \Pi^{(-)}(k, K_{t-1})$. При указанных условиях легко показать, что $K_t^{(+)} < K_t^{(-)}$, и оптимальный уровень капитала в момент времени t равен

$$K_t = \begin{cases} K_t^{(+)} & \text{при} \quad K_{t-1} < K_t^{(+)}, \\ K_{t-1} & \text{при} \quad K_t^{(+)} \leq K_{t-1} \leq K_t^{(-)}, \\ K_t^{(-)} & \text{при} \quad K_{t-1} > K_t^{(+)}. \end{cases} \tag{20}$$

Модель дополняется спецификацией $K_t^{(+)}$ и $K_t^{(-)}$ в терминах наблюдаемых и ненаблюдаемых величин. Например, $K_t^{(+)} = \alpha^{(+)} + X_t\beta + \varepsilon_t$ и $K_t^{(-)} = \alpha^{(-)} + X_t\beta + \varepsilon_t$, где $\alpha^{(+)}$ и $\alpha^{(-)}$ – параметры, и $\alpha^{(+)} < \alpha^{(-)}$. По случайной выборке для (K_t, K_{t-1}, X_t) требуется оценить параметры $\alpha^{(+)}$, $\alpha^{(-)}$ и β.

2 Оценивание модели усеченной регрессии

2.1 Смещение МНК-оценки

Рассмотрим модель усеченной регрессии, задаваемую выражением $(Y, X) = \{(Y^*, X^*)|Y^* > c\}$, где $Y^* = X^*\beta + \varepsilon$. Поскольку константа c известна, без ограничения общности положим $c = 0$.[5] Предположим, что регрессия Y на X оценивается с помощью МНК. Рисунок 3 графически иллюстрирует смещение МНК-оценки. Истинный наклон линии регрессии равен 1.5, а его МНК-оценка равна 1.15 ($s.e. = 0.05$).[6]

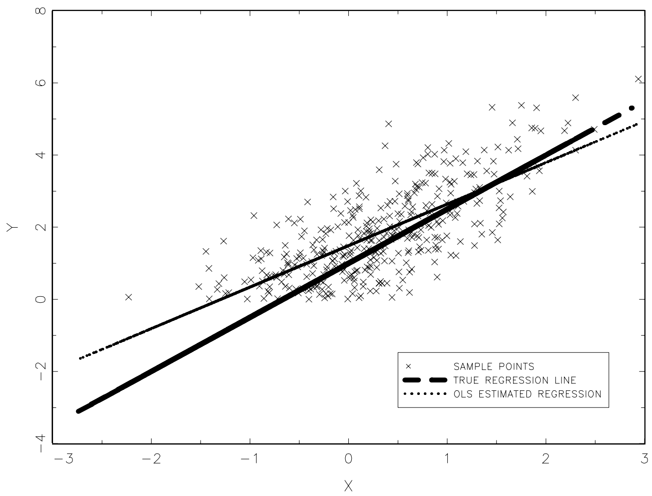


Рис. 3: Смещение МНК-оценки в модели усеченной регрессии.

Формально $Y = \{Y^*|Y^* > 0\} = X^*\beta + \varepsilon^{Trun}$, где $\varepsilon^{Trun} \equiv \{\varepsilon|Y^* > 0\}$. Следовательно,

$$\mathbb{E}[Y|X] = \mathbb{E}[X^*\beta + \varepsilon^{Trun}|X] = X\beta + \mathbb{E}[\varepsilon^{Trun}|X]. \tag{21}$$

Величина $\mathbb{E}[\varepsilon^{Trun}|X]$ отражает эффект *самоотбора выборки* в условном среднем Y при данном X. Заметим, что $\mathbb{E}[\varepsilon^{Trun}|X] = \mathbb{E}[\varepsilon|\varepsilon > -X\beta]$, что, вообще говоря, не равно нулю и зависит от X. Если ε не зависит от X, эффект самоотбора выборки зависит от X только через индекс $X\beta$. Тогда его можно представить в виде функции $s(X\beta)$. Легко показать, что смещение $s(X\beta)$ – убывающая функция от индекса $X\beta$. Чтобы убедиться в этом, заметим, что

$$\begin{cases} \text{при } X\beta \to +\infty & s(X\beta) \to \mathbb{E}[\varepsilon|\varepsilon > -\infty] = \mathbb{E}[\varepsilon] = 0, \\ \text{при } X\beta \to -\infty & s(X\beta) \to \mathbb{E}[\varepsilon|\varepsilon > +\infty] = +\infty. \end{cases} \tag{22}$$

Следовательно, $s(X\beta)$ отрицательно зависит от $X\beta$. В модели регрессии, усеченной справа, смещение вследствие самоотбора также убывает по $X\beta$. Принимая во внимание то, что $\mathbb{E}[Y|X] = X\beta + s(X\beta)$, можно записать следующее уравнение регрессии Y на X:

$$Y = X\beta + s(X\beta) + \tilde{\varepsilon}. \tag{23}$$

Случайная ошибка $\tilde{\varepsilon}$ равна $\varepsilon^{Trun} - s(X\beta)$, и, по построению, независима в среднем от X. Приведенное выражение показывает несостоятельность МНК-оценки, не учитывающей самоотбора выборки. Игнорирование самоотбора выборки ведет к тому, что случайная ошибка в регрессии равна $s(X\beta) + \tilde{\varepsilon}$ и отрицательно коррелирует с $X\beta$.

[5] Если c не равна нулю, всегда можно переопределить Y^* как исходную переменную Y^* за вычетом c.

[6] Процесс, порождающий данные, таков, что X^* и ε – независимые стандартные нормальные случайные величины, $Y^* = 1.0 + 1.5X^* + \varepsilon$, и усечение происходит слева в точке $y = 0$. Размер выборки $n = 500$.

2.2 Оценивание методом максимального правдоподобия

Для получения ММП-оценки параметра β необходимо ввести в модель дополнительную предпосылку о виде распределения ε. Типичным в этом классе моделей является предположение о том, что ε – IID с распределением $N(0,\sigma^2)$. Тогда логарифмическая функция правдоподобия для этой модели и выборки имеет вид

$$\mathrm{logL}(\beta,\sigma)=\sum_{i=1}^{n}\ln\mathbb{P}\{Y=y_i|X=x_i\},$$

и условные вероятности равны

$$\begin{aligned}\mathbb{P}\{Y=y_i|X=x_i\}&=\mathbb{P}\{Y^*=y_i|X=x_i;Y^*>0\}=\\&=\frac{\mathbb{P}\{\varepsilon=y_i-x_i\beta\}}{\mathbb{P}\{\varepsilon>-x_i\beta\}}=\frac{\dfrac{1}{\sigma}\phi\left(\dfrac{y_i-x_i\beta}{\sigma}\right)}{\Phi\left(\dfrac{x_i\beta}{\sigma}\right)},\end{aligned}\tag{24}$$

где $\phi(\cdot)$ и $\Phi(\cdot)$ – ФПР и КФР стандартного нормального распределения. Следовательно, логарифмическую функцию правдоподобия можно записать следующим образом:

$$\mathrm{logL}(\beta,\sigma)=-n\ln(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-x_i\beta)^2-\sum_{i=1}^{n}\ln\Phi\left(\frac{x_i\beta}{\sigma}\right).\tag{25}$$

Первая часть этого выражения – логарифмическая функция правдоподобия для классической модели линейной регрессии. Вторая часть учитывает усечение. Заметим, что, в отличие от модели бинарного выбора, логарифмическая функция правдоподобия в данном случае зависит не только от β/σ, но и от β и σ по отдельности, так что оба параметра можно идентифицировать.[7]

Логарифмическая функция правдоподобия $\mathrm{logL}(\beta,\sigma)$ не является глобально вогнутой по (β,σ). Это важный момент. Максимизация глобально вогнутых функций – очень простая задача, то есть можно использовать простые алгоритмы, такие как метод Ньютона или BHHH. Однако максимизация функций, не являющихся всюду вогнутыми, вычислительно более сложна, поскольку требует глобального поиска на всем пространстве параметров, чтобы гарантировать получение глобального, а не локального максимума. Тем не менее, для этой модели легко перепараметризовать логарифмическую функцию правдоподобия и получить глобально вогнутую функцию. Определим параметры $\theta=1/\sigma$ и $\gamma=\beta/\sigma$ и рассмотрим логарифмическую функцию правдоподобия в терминах этих параметров:

$$\mathrm{logL}(\gamma,\theta)=n\ln(\theta)-\frac{1}{2}\sum_{i=1}^{n}(\theta y_i-x_i\gamma)^2-\sum_{i=1}^{n}\ln\Phi\left(x_i\gamma\right).$$

Функция $\mathrm{logL}(\gamma,\theta)$ глобально вогнута по (γ,θ). Заметим, что существует взаимно однозначное соответствие между (γ,θ) и (β,σ). Следовательно, по свойству *инвариантности к перепараметризации*, ММП-оценки параметров (β,σ) имеют вид $\hat\sigma_{MLE}=1/\hat\theta$ и $\hat\beta_{MLE}=\hat\gamma/\hat\theta$. Дисперсионную матрицу можно получить, используя дельта-метод.

В контексте линейных регрессионных моделей МНК-оценка состоятельна, если регрессоры не коррелируют со случайной ошибкой. Состоятельность МНК-оценки робастна к гетероскедастичности, серийной корреляции и ненормальности ошибки. Гетероскедастичность является типичной характеристикой большинства кросс-секционных данных. Следовательно,

[7]Заметим также, что в этой модели можно получить остатки $\hat\varepsilon$, которые являются состоятельными оценками ε.

важный вопрос состоит в том, является ли полученная ММП-оценка робастной к гетероскедастичности ε. Остается ли ММП-оценка состоятельной, если ε гетероскедастична, а функция правдоподобия соответствует гомоскедастичной модели? Ответ: нет. Симуляции Монте–Карло показывают, что МНК-оценка может становиться серьезно смещенной. Эта проблема мотивирует изучение других оценок, робастных к гетероскедастичности и ненормальности случайной ошибки.

2.3 Симметрично урезанная МНК-оценка

Джеймс Пауэлл заложил основу полупараметрического оценивания моделей усеченной и цензурированной регрессии. В Powell (1984) предложены оценки наименьших абсолютных отклонений (НАО), которые робастны к гетероскедастичности и ненормальности ошибок. В Powell (1986) рассматривается другая робастная оценка, основанная на симметричном усечении (или цензурировании) хвостов распределения зависимой переменной. Остановимся на симметрично урезанной МНК-оценке (СУМНК-, или STLS-оценке).

Рассмотрим модель усеченной слева регрессии и определим следующую зависимую переменную:[8]

$$\tilde{Y} \equiv \{Y^*|0 < Y^* < 2X\beta\} = \{Y|Y < 2X\beta\}. \tag{26}$$

Переменная $\tilde{Y}$ усечена слева и справа. Заметим, что точки усечения переменной $\tilde{Y}$ (то есть 0 и $2X\beta$) находятся на одинаковом расстоянии от условного среднего $\mathbb{E}[Y^*|X] = X\beta$. При таком «симметричном усечении» получаем

$$\mathbb{E}[\tilde{Y}|X] = \mathbb{E}[X\beta + \varepsilon|X, 0 < X\beta + \varepsilon < 2X\beta] = X\beta + \mathbb{E}[\varepsilon|X, -X\beta < \varepsilon < X\beta]. \tag{27}$$

В линейной регрессии $\tilde{Y}$ на X выражение $\mathbb{E}[\varepsilon|X, -X\beta < \varepsilon < X\beta]$ представляет собой эффект *самоотбора выборки*. Ясно, что он равен нулю, если функция плотности распределения ε симметрична относительно нуля.

Таким образом, можно получить состоятельную оценку β с помощью МНК-оценивания регрессии $\tilde{Y}$ на X. Эта оценка робастна к гетероскедастичности ε. Более того, предположение о симметричности распределения ε является более общим, чем предположение о нормальности. Однако $\tilde{Y}$ не наблюдается. Чтобы получить случайную выборку для $\tilde{Y}$ нужно усечь наблюдаемую зависимую переменную Y справа в точке $2X\beta$. Но значение β неизвестно. Чтобы справиться с этой проблемой, рассмотрим следующий критерий:

$$Q(\beta) = \sum_{i=1}^{n} \mathbb{I}\{y_i < 2x_i\beta\}\,(y_i - x_i\beta)^2\,. \tag{28}$$

Эта функция представляет собой симметрично усеченную сумму квадратов остатков. Оценка СУМНК определяется как значение β, минимизирующее этот критерий. Эта оценка состоятельна и асимптотически нормальна. Асимптотическая дисперсионная матрица СУМНК-оценки имеет вид

$$\begin{aligned} \mathbb{V}[\hat{\beta}_{STLS}] &= C^{-1}DC^{-1}, \text{ где} \\ C &= \mathbb{E}[\mathbb{I}\{Y < 2X\beta\}XX'], \\ D &= \mathbb{E}[\mathbb{I}\{X\beta > 0\}\min\{\varepsilon^2; (X\beta)^2\}XX']. \end{aligned} \tag{29}$$

Заметим, что функция $Q(\beta)$ не является непрерывной и дифференцируемой относительно β во многих различных точках (стольких, сколько имеется точек в выборке). Следовательно, ее минимизация может быть затруднительной. Простой метод нахождения (локального) минимума состоит в следующем. Шаг 1: возьмем начальное значение β, скажем $\hat{\beta}^{(1)}$.

[8] Аналогично для модели усеченной справа регрессии можно определить $\tilde{Y} \equiv \{Y^*|2X\beta < Y^* < 0\} = \{Y|2X\beta < Y\}$.

Это может быть, например, МНК-оценка по всей выборке $\{y_i, x_i\}$. Шаг 2: получим усеченную переменную $\tilde{y}_i^{(1)} = \{y_i | y_i < 2x_i\hat{\beta}^{(1)}\}$. То есть удалим все наблюдения с $y_i > 2x_i\hat{\beta}^{(1)}$. Шаг 3: оценим с помощью МНК регрессию $\tilde{y}_i^{(1)}$ на x_i для получения нового значения β, $\hat{\beta}^{(2)}$. Итерируем шаги 2 и 3 до сходимости, то есть до тех пор, пока не выполнится условие $||\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}|| <$ *малое число*. По достижении сходимости эта процедура дает локальный минимум $Q(\beta)$. Для проверки на глобальность минимума необходимо выполнить глобальный поиск, повторяя эту процедуру для различных начальных значений β.

Данный метод прост и особенно полезен, когда имеется большая выборка и масштаб усечения незначительный. Для относительно малых выборок или при значительном усечении потеря эффективности, связанная с усечением, может быть очень серьезной, а оценки – неточными.

2.4 Тест Хаусмана на гетероскедастичность и ненормальность

Для осуществления теста Хаусмана нужна оценка, которая эффективна при H_0 и несостоятельна при H_1, и оценка, состоятельная как при H_0, так и при H_1. Следовательно, можно использовать ММП-оценку и оценку Пауэлла для построения теста на гетероскедастичность и ненормальность. Нулевая гипотеза состоит в том, что $\varepsilon_i \sim$ IID $N(0, \sigma^2)$, а тест-статистика имеет вид

$$\mathcal{H} = (\hat{\beta}_{STLS} - \hat{\beta}_{MLE})' \left(\mathbb{V}[\hat{\beta}_{STLS}] - \mathbb{V}[\hat{\beta}_{MLE}]\right)^{-1} (\hat{\beta}_{STLS} - \hat{\beta}_{MLE}), \tag{30}$$

и при H_0 имеет распределение хи-квадрат с k степенями свободы.

3 Модель цензурированной регрессии (тобит)

3.1 Смещение МНК-оценки

Рассмотрим модель цензурированной регрессии, такую, что имеется случайная выборка $X = X^*$ и $Y = \max\{Y^*, c\}$, где $Y^* = X\beta + \varepsilon$. Снова без ограничения общности можно положить $c = 0$. Предположим, что оценивается регрессия Y на X. Рисунок 4 графически иллюстрирует смещение МНК-оценки. Истинный наклон линии регрессии равен 1.5, а его МНК-оценка равна 1.10 ($s.e. = 0.04$).[9]

Формально $Y = \max\{X\beta + \varepsilon, 0\}$, или в форме регрессионного уравнения $Y = X\beta + \varepsilon^{Cens}$, где $\varepsilon^{Cens} \equiv \max\{\varepsilon, -X\beta\}$. Следовательно,

$$\mathbb{E}[Y|X] = X\beta + \mathbb{E}[\varepsilon^{Cens}|X] = X\beta + \mathbb{E}[\max\{\varepsilon, -X\beta\}|X]. \tag{31}$$

Величина $\mathbb{E}[\varepsilon^{Cens}|X]$ представляет собой эффект *самоотбора выборки* в условном среднем Y при данном X. Заметим, что $\mathbb{E}[\varepsilon^{Cens}|X] = \mathbb{E}[\max\{\varepsilon, -X\beta\}|X]$, что, вообще говоря, не равно нулю и зависит от X. Если ε не зависит от X, эффект самоотбора выборки зависит от X только через индекс $X\beta$, то есть $\mathbb{E}[\varepsilon^{Cens}|X] = s(X\beta)$, и $s(\cdot)$ – убывающая функция. Тогда, учитывая, что $\mathbb{E}[Y|X] = X\beta + s(X\beta)$, можно записать следующее регрессионное уравнение: $Y = X\beta + s(X\beta) + \tilde{\varepsilon}$, где ошибка $\tilde{\varepsilon} \equiv \varepsilon^{Cens} - s(X\beta)$ независима в среднем от X. МНК-оценка, игнорирующая эффект самоотбора выборки $s(X\beta)$, несостоятельна.

[9] Процесс, порождающий данные, таков, что X^* и ε – независимые стандартные нормальные случайные величины, $Y^* = 1.0 + 1.5X^* + \varepsilon$, а цензурирование слева происходит в точке $y = 0$. Размер выборки $n = 500$.

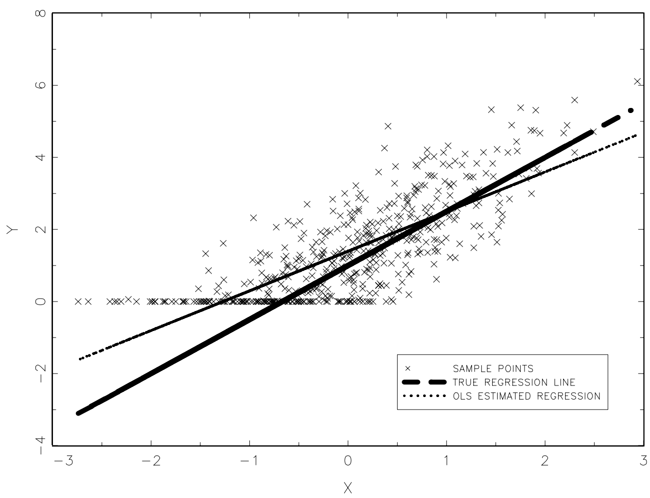


Рис. 4: Смещение МНК-оценки в модели цензурированной регрессии

3.2 Оценивание методом максимального правдоподобия

Логарифмическая функция правдоподобия для этой модели и выборки имеет вид

$$\text{logL}(\beta, \sigma) = \sum_{i=1}^{n} \ln \mathbb{P}\{Y = y_i | X = x_i\},$$

где условные вероятности равны

$$\mathbb{P}\{Y = y_i | X = x_i\} = \begin{cases} \mathbb{P}\{Y^* = y_i | X = x_i\} = f_\varepsilon(y_i - x_i\beta) & \text{при} \quad y_i > 0, \\ \mathbb{P}\{Y^* < 0 | X = x_i\} = F_\varepsilon(-x_i\beta) & \text{при} \quad y_i = 0. \end{cases} \tag{32}$$

При предположении о том, что $\varepsilon_i \sim \text{IID } N(0, \sigma^2)$, логарифмическая функция правдоподобия имеет вид

$$\text{logL}(\beta, \sigma) = -n_1 \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{y_i > 0} (y_i - x_i\beta)^2 + \sum_{y_i = 0} \ln \Phi\left(\frac{-x_i\beta}{\sigma}\right), \tag{33}$$

где n_1 – число наблюдений с $y_i > 0$. Все замечания по поводу модели усеченной регрессии, также справедливы в случае цензурирования.

3.3 Симметрично урезанная МНК-оценка

Рассмотрим модель цензурированной слева регрессии и определим следующую зависимую переменную:

$$\tilde{Y} = \begin{cases} 0 & \text{при} \quad Y^* \leq 0, \\ Y^* & \text{при} \quad 0 < Y^* < 2X\beta = \min\{Y; 2X\beta\}, \\ 2X\beta & \text{при} \quad Y^* \geq 2X\beta. \end{cases} \tag{34}$$

Переменная $\tilde{Y}$ цензурирована и слева, и справа. Ясно, что точки цензурирования $\tilde{Y}$ (то есть 0 и $2X\beta$) находятся на одинаковом расстоянии от $X\beta$, условного среднего Y^*. При таком симметричном цензурировании получаем, что $\tilde{Y} = \min\{Y; 2X\beta\}$=$\min\{\max\{X\beta + \varepsilon; 0\}; 2X\beta\}$.[10] Или, в виде наподобие регрессионного уравнения:

$$\tilde{Y} = X\beta + \mathbb{I}\{\varepsilon < -X\beta\}(-X\beta) + \mathbb{I}\{\varepsilon > X\beta\}(X\beta) + \mathbb{I}\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon. \tag{35}$$

[10]Заметим, что $\max\{X\beta+\varepsilon; 0\} = X\beta + \max\{\varepsilon; -X\beta\}$. Следовательно, $\min\{\max\{X\beta+\varepsilon; 0\}; 2X\beta\} = \min\{X\beta + \max\{\varepsilon; -X\beta\}; 2X\beta\} = X\beta + \min\{\max\{\varepsilon; -X\beta\}; X\beta\}$. Или, что эквивалентно, $X\beta + \mathbb{I}\{\varepsilon < -X\beta\}(-X\beta) + \mathbb{I}\{\varepsilon > X\beta\}(X\beta) + \mathbb{I}\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon$.

В линейной регрессии $\tilde{Y}$ на X, эффект самоотбора – это условное матожидание случайного члена $\mathbb{I}\{\varepsilon < -X\beta\}(-X\beta) + \mathbb{I}\{\varepsilon > X\beta\}(X\beta) + \mathbb{I}\{-X\beta \leq \varepsilon \leq -X\beta\}\varepsilon$. Как и в случае усечения, этот эффект равен нулю, если функция плотности распределения ε симметрична относительно нуля.

СУМНК-оценка модели цензурированной регрессии определяется как значение β, минимизирующее следующий критерий:

$$Q(\beta) = \sum_{i=1}^{n} \left(\min\{y_i; 2x_i\beta\} - x_i\beta\right)^2. \tag{36}$$

Эта функция представляет собой сумму квадратов остатков линейной регрессии $\tilde{Y}$ на X. Получаемая оценка состоятельна, асимптотически нормальна и робастна к ненормальности и гетероскедастичности ε.

4 Модели с самоотбором выборки

Рассмотрим модель с самоотбором выборки, в которой $Y = (1-D)Y_0^* + DY_1^*$, где

$$\begin{aligned} Y_0^* &= X\beta_0 + \varepsilon_0, \\ Y_1^* &= X\beta_1 + \varepsilon_1, \end{aligned} \tag{37}$$

и

$$D = \mathbb{I}\{Z\gamma - u > 0\}. \tag{38}$$

Ненаблюдаемые величины ε_0, ε_1 и u не являются независимо распределенными. Например, предположим, что D – индикатор события «индивид является членом профсоюза», Y_1^* – заработная плата индивида, если он входит в профсоюз, а Y_0^* – если не входит. Задача состоит в оценке параметров β_0 и β_1. Иногда особый интерес представляет средний эффект воздействия $ATE(X) = X(\beta_1 - \beta_0)$, то есть средняя отдача от членства в профсоюзе для индивида с характеристиками X.

4.1 Смещение МНК-оценки

Можно построить следующие два вида МНК-оценок векторов β_0 и β_1: (а) *совместную МНК-оценку*, когда с помощью МНК оценивается регрессия Y на X и DX, то есть $Y = X\beta_0 + DX(\beta_1 - \beta_0) + e$; (б) *отдельные МНК-оценки*, когда по отдельности оцениваются регрессия $Y = X\beta_0 + e_0$ для подвыборки наблюдений с $D = 0$ и регрессия $Y = X\beta_1 + e_1$ для подвыборки наблюдений с $D = 1$. Ясно, что при отсутствии ограничений на параметры β_0 и β_1 между уравнениями эти оценки совпадают, а следовательно, можно рассматривать лишь один способ оценивания, скажем (б).

По построению, случайная ошибка $e_j \equiv \{\varepsilon_j | D = j\}$. Следовательно,

$$\begin{aligned} \mathbb{E}[e_0|X] &= \mathbb{E}[\varepsilon_0|X, D=0] = \mathbb{E}[\varepsilon_0|X, u \geq Z\gamma], \\ \mathbb{E}[e_1|X] &= \mathbb{E}[\varepsilon_1|X, D=1] = \mathbb{E}[\varepsilon_1|X, u < Z\gamma]. \end{aligned} \tag{39}$$

Если ε и u не являются независимыми, и если только X и Z не являются независимыми (что крайне нереалистично при использовании неэкспериментальных данных), эти эффекты самоотбора коррелируют с X. Значит, ошибки e_0 и e_1 коррелируют с X, и соответствующие МНК-оценки для β_0 и β_1 несостоятельны.

Дадим интерпретацию этому смещению в контексте примера об отдаче от участия в профсоюзе. МНК-оценка $\beta_1 - \beta_0$ в регрессии $Y = X\beta_0 + DX(\beta_1 - \beta_0) + e$ представляет собой комбинацию двух эффектов: (1) реальной отдачи от участия в профсоюзе, $\beta_1 - \beta_0$ и (2) того факта, что работники, которые решают вступить в профсоюз, обычно также являются теми, для кого велик «эффект воздействия», или разница заработных плат $Y_1^* - Y_0^*$. Первый элемент – это тот причинно-следственный эффект, который требуется оценить. Второй элемент – это «ложный» эффект, не являющийся следствием участия в профсоюзе. Для иллюстрации предположим, что X – это просто константа. Предположим также, что участие в профсоюзе имеет два эффекта: оно увеличивает константу, то есть $\beta_1 > \beta_0$, и сокращает дисперсию заработных плат, то есть $\varepsilon_1 = \lambda\varepsilon_0$, где $\lambda < 1$. Допустим, что единственный фактор, влияющий на решение об участии в профсоюзе, – это разница заработных плат (модель Роя), так что $Z\gamma - u = Y_1^* - Y_0^* = (\beta_1 - \beta_0) + (\varepsilon_1 - \varepsilon_0) = (\beta_1 - \beta_0) - (1-\lambda)\varepsilon_0$. В этом примере ясно, что

$$\begin{aligned} \text{plim}\hat{\beta}_0^{OLS} &= \mathbb{E}[Y|D=0] = \beta_0 + \mathbb{E}\left[\varepsilon_0 \,\middle|\, \varepsilon_0 > \frac{\beta_1 - \beta_0}{1-\lambda}\right] > \beta_0, \\ \text{plim}\hat{\beta}_1^{OLS} &= \mathbb{E}[Y|D=1] = \beta_1 + \lambda\mathbb{E}\left[\varepsilon_0 \,\middle|\, \varepsilon_0 < \frac{\beta_1 - \beta_0}{1-\lambda}\right] < \beta_1. \end{aligned} \tag{40}$$

Следовательно, в данном примере оценка $\hat{\beta}_0^{OLS}$ переоценивает β_0, поскольку не участвующие в профсоюзах работники имеют большие значения ε_0 (то есть большую производительность), $\varepsilon_0 > (\beta_1 - \beta_0)/(1-\lambda)$. Также оценка $\hat{\beta}_1^{OLS}$ недооценивает β_1, поскольку участвующие в профсоюзах работники имеют более низкие значения ε_0, то есть $\varepsilon_0 < (\beta_1 - \beta_0)/(1-\lambda)$. В результате при изложенных предпосылках МНК-оценка $\beta_1 - \beta_0$ недооценивает истинную отдачу от участия в профсоюзе.

4.2 Оценивание методом максимального правдоподобия

Зависимые переменные модели – это Y и D, а экзогенные объясняющие переменные – X и Z. Логарифмическая функция правдоподобия для этой модели и выборки имеет вид

$$\text{logL}(\beta, \gamma, \Omega) = \sum_{i=1}^{n} \ln \mathbb{P}\{Y = y_i, D = d_i | X = x_i, Z = z_i\} \tag{41}$$

с вероятностями

$$\begin{aligned} \mathbb{P}\{Y = y_i, D = 0 | X = x_i, Z = z_i\} &= \mathbb{P}\{\varepsilon_0 = y_i - x_i\beta_0; u_i > z_i\gamma\} \\ &= \int_{z_i\gamma}^{+\infty} f_{\varepsilon_0,u}(y_i - x_i\beta_0, u)du \end{aligned} \tag{42}$$

и

$$\begin{aligned} \mathbb{P}\{Y = y_i, D = 1 | X = x_i, Z = z_i\} &= \mathbb{P}\{\varepsilon_1 = y_i - x_i\beta_1; u_i < z_i\gamma\} \\ &= \int_{-\infty}^{z_i\gamma} f_{\varepsilon_1,u}(y_i - x_i\beta_1, u)du, \end{aligned} \tag{43}$$

где $f_{\varepsilon_0,u}$ и $f_{\varepsilon_1,u}$ – функции плотности совместных распределений (ε_0, u) и (ε_1, u), соответственно.

При ММП-оценивании данной модели обычно предполагается, что тройка $(\varepsilon_0, \varepsilon_1, u)$ имеет совместное нормальное распределение. Дисперсия u нормируется к 1. Параметры, которые входят в данную функцию правдоподобия, – это β_0, β_1, γ, стандартные отклонения σ_0 и σ_1 и ковариации σ_{0u} и σ_{1u}. В общем, эта функция правдоподобия не является глобально вогнутой и может иметь несколько локальных максимумов. Более того, в отличие от моделей усеченной и цензурированной регрессии, не существует перепараметризации, при которой функция правдоподобия всюду вогнута. Следовательно, при реализации оптимизационного алгоритма необходимо выбирать различные начальные значения параметров и сравнивать получающиеся по достижении сходимости значения функции правдоподобия в надежде получить глобальный максимум.

4.3 Двухшаговый метод Хекмана

Хекман (Heckman, 1976, 1979) предложил альтернативный двухшаговый подход, который дает состоятельные оценки в модели с самоотбором выборки и очень легок в применении. Вычислительная простота этого двухшагового метода делает его очень привлекательным на практике. Однако есть по крайней мере одна другая причина, по которой двухшаговый метод Хекмана так популярен в практических приложениях. Как и в случаях моделей усеченной и цензурированной регрессии, ММП-оценка в общей модели с самоотбором выборки не является робастной к гетероскедастичности и ненормальности. Хотя двухшаговый подход Хекмана был предложен в контексте параметрической модели с нормальными и гомоскедастичными ошибками, одно из наиболее привлекательных свойств получаемой оценки состоит в том, что ее можно расширить на полупараметрический случай с ненормальными и гетероскедастичными ошибками.

Рассмотрим сначала эту оценку в контексте полностью параметрической модели с нормальными и гомоскедастичными ошибками. Заметим, что

$$\begin{aligned}\mathbb{E}[Y|X,Z,D=0] &= X\beta_0 + \mathbb{E}[\varepsilon_0|X,Z,D=0]\\ &= X\beta_0 + \frac{1}{1-F_u(Z\gamma)}\int_{Z\gamma}^{+\infty}\mathbb{E}[\varepsilon_0|u]f_u(u)du\end{aligned} \tag{44}$$

и

$$\begin{aligned}\mathbb{E}[Y|X,Z,D=1] &= X\beta_1 + \mathbb{E}[\varepsilon_1|X,Z,D=1]\\ &= X\beta_1 + \frac{1}{F_u(Z\gamma)}\int_{-\infty}^{Z\gamma}\mathbb{E}[\varepsilon_1|u]f_u(u)du.\end{aligned} \tag{45}$$

При нормальности $(u,\varepsilon_0,\varepsilon_1)$ эти выражения принимают вид

$$\begin{aligned}\mathbb{E}[Y|X,Z,D=0] &= X\beta_0 + \sigma_{0u}\lambda(-Z\gamma),\\ \mathbb{E}[Y|X,Z,D=1] &= X\beta_1 - \sigma_{1u}\lambda(Z\gamma),\end{aligned} \tag{46}$$

где функцию

$$\lambda(c) \equiv \frac{\phi(c)}{\Phi(c)}$$

называют обратным отношением Миллса, или лямбдой Хекмана.

Основываясь на этом результате, Хекман предложил следующую двухшаговую процедуру. Шаг 1: оценим γ методом максимального правдоподобия в пробит-модели $D=\mathbb{I}\{Z\gamma-u>0\}$. Получим $\{z_i\hat{\gamma}\}$ для каждого наблюдения в выборке и найдем оценки лямбд Хекмана, $\hat{\lambda}_{0i}=\phi(-z_i\hat{\gamma})/\Phi(-z_i\hat{\gamma})$ и $\hat{\lambda}_{1i}=\phi(z_i\hat{\gamma})/\Phi(z_i\hat{\gamma})$. Шаг 2: с помощью МНК оценим регрессию Y на X и $\hat{\lambda}_0$, используя подвыборку наблюдений с $D=0$, и регрессию Y на X и $\hat{\lambda}_1$, используя подвыборку наблюдений с $D=1$. Эта процедура дает состоятельные оценки β_0, β_1, σ_{0u} и σ_{1u}. Amemiya (1985, с. 370–371) приводит выражение для корректировки стандартных ошибок оценок параметров с учетом ошибки оценивания в переменных $\hat{\lambda}_0$ и $\hat{\lambda}_1$.

Каким образом данная процедура учитывает смещение вследствие самоотбора? Это достигается путем включения в регрессию (оценки) эффекта самоотбора $\hat{\lambda}$. Как по отдельности идентифицировать причинно-следственный эффект X на Y (через $X\beta_j$) и смещение вследствие самоотбора $\hat{\lambda}_j$? Или, другими словами, почему $\hat{\lambda}$ и X неколлинеарны? Есть две возможные причины. Во-первых, в Z могут быть переменные, которых нет в X (то есть выполняется условие невключения). В этом случае, если эти переменные имеют достаточную объясняющую силу в пробит-модели, $\hat{\lambda}_j$ имеет источник вариации, который *не зависит* от X. Во-вторых, $\hat{\lambda}$ – нелинейная функция от $Z\hat{\gamma}$. Даже если $Z\subseteq X$, переменная $\hat{\lambda}$ имеет выборочную вариацию, которая *линейно независима* от X. Первый источник идентификации называют *идентификацией через инструментальные переменные*, и он не зависит от предположений о функциональной форме, то есть идентификация возможна, даже если модель специфицирует непараметрическую взаимосвязь между Y_j^* и X. Второй источник идентификации называют *идентификацией через условие невключения* и для него критичны параметрические предположения, то есть линейность взаимосвязи между Y_j^* и X и нормальность ошибок.

Предшествующее обсуждение указывает на дополнительную причину, по которой желательно ослабить предположение о нормальности. Даже если интерес представляют линейные эффекты X на Y_j^*, лучше, если идентификация этих эффектов не полагается исключительно на предположение о линейности и параметрическое предположение о распределении ненаблюдаемых величин. Опишем расширение двухшаговой процедуры Хекмана, которое допускает произвольное распределение ненаблюдаемых величин. Рассмотрим модель с самоотбором выборки, в которой ненаблюдаемые величины $(\varepsilon_0, \varepsilon_1, u)$ независимы от (X, Z) и имеют произвольное распределение с носителем в евклидовом пространстве и заданной мерой Лебега. На самом деле, можно разрешить гетероскедастичность $(\varepsilon_0, \varepsilon_1, u)$, если дисперсии и ковариации этих величин зависят от (X, Z) только через индекс $Z\gamma$. Без дополнительных предположений из модели следует, что

$$\begin{aligned} &\mathbb{E}[Y|X, Z, D = 0] = X\beta_0 + s_0(Z\gamma), \\ &\mathbb{E}[Y|X, Z, D = 1] = X\beta_1 + s_1(Z\gamma), \end{aligned} \tag{47}$$

где теперь функциональная форма эффектов самоотбора $s_0(\cdot)$ и $s_1(\cdot)$ неизвестна. Тем не менее, известно, что они являются одноиндексными функциями: они зависят только от $Z\gamma$. При наличии оценки γ из модели бинарного выбора[11] можно с произвольной точностью аппроксимировать элементы $s_j(Z\hat{\gamma})$, используя полином порядка q от $Z\hat{\gamma}$. То есть на втором шаге можно с помощью МНК оценить следующие регрессии:

$$\begin{aligned} &\{Y|D = 0\} = X\beta_0 + \sum_{j=1}^{q} \rho_{0j}(Z\hat{\gamma})^j + e_0, \\ &\{Y|D = 1\} = X\beta_1 + \sum_{j=1}^{q} \rho_{1j}(Z\hat{\gamma})^j + e_1. \end{aligned} \tag{48}$$

Некоторые исследователи также предлагают использовать полином от оцененной лямбды Хекмана или от оценки вероятности из модели дискретного выбора (то есть оценки вероятности воздействия). Можно также использовать другие виды полупараметрических оценок для частично линейных моделей (см. Robinson, 1983, и Yatchew, 2003). Ясно, что идентификация β_0 и β_1 основана только на условии невключения. Из этого также следует, что для идентификации параметров при данном подходе индекс $Z\hat{\gamma}$ должен обладать достаточной выборочной вариацией, независимой от X. Кроме того, необходимо обосновать выполнение условия невключения, исходя из экономических соображений и понимания задачи.

Литература

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press: Cambridge, Massachusetts.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such model. *Annals of Economic and Social Measurement* 15, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153–161.

[11]Параметрическая спецификация модели дискретного выбора в данном случае неважна. В пробит-модели можно также использовать полином от z_i.

Heckman, J., & B. Honoré (1990). The empirical content of the Roy model. *Econometrica* 58, 1121–1149.

Powell, J. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303–325.

Powell, J. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* 54, 1435–1460.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers (New Series)* 3, 135–146.

Robinson, P. (1988). Root-N-consistent semiparametric regression. *Econometrica* 56, 931–954.

Yatchew, A. (2003). Semiparametric regression for the applied econometrician. Глава в сборнике *Themes in Modern Econometrics* под редакцией P.C.B. Phillips. Cambridge University Press.

Some notes on sample selection models

Victor Aguirregabiria

University of Toronto, Toronto, Canada

Sample selection problems are pervasive when working with micro economic models and datasets of individuals, households or firms. During the last three decades, there have been very significant developments in this area of econometrics. Different type of models have been proposed and used in empirical applications. And new estimation and inference methods, both parametric and semiparametric, have been developed. These notes provide a brief introduction to this large literature.

Эконометрический ликбез: непараметрические и полупараметрические методы

Непараметрическая регрессия*

Станислав Анатольев†

Российская экономическая школа, Москва, Россия

Настоящее эссе повествует о принципах и методологии непараметрического оценивания регрессии среднего. Акцент делается на методах ядерного сглаживания, но дается и обзор неядерных методов.

1 Введение

Пусть имеется случайная выборка $\{(x_i, y_i)\}_{i=1}^n$ из популяции пар (x, y). Нас интересует оценивание регрессии среднего $g(x) = \mathbb{E}[y|x]$ в предположении, что она существует для всех x носителя и является гладкой. Для этого чаще всего пользуются *параметрическими* методами, когда предполагается, что регрессионная функция имеет известную функциональную форму и конечное число неизвестных параметров. Оценивание этих параметров автоматически дает оценки для $g(x)$. Естественно, неверная спецификация функциональной формы может привести к серьезным искажениям при оценивании и инференции, причем часто непредсказуемым (см. Крил, 2008).

В настоящем эссе мы даем обзор *непараметрического* оценивания регрессии среднего, то есть такого, при котором избегают параметрических предположений о функциональной форме. Как и большая часть соответствующей литературы, мы делаем акцент на методах ядерного сглаживания. В отличие от остальной литературы, однако, мы рассматриваем с самого начала регрессию среднего, избегая предварительного разговора об оценивании плотности. Другие источники информации на данную тему включают обзоры Härdle & Linton (1994) и Расин (2008), а также монографии Härdle (1990), Pagan & Ullah (1999) и Li & Racine (2007).

2 Построение непараметрической оценки

Мы вначале предполагаем, что регрессор x – единственный. Впоследствии мы обсудим и случай многопеременной регрессии.

2.1 Дискретный регрессор

Прежде всего рассмотрим случай дискретного регрессора. Пусть носитель x сосредоточен в $a_{(1)}, \ldots, a_{(k)}$, где $a_{(1)} < \ldots < a_{(k)}$, и k конечно (если носитель – бесконечное, но счетное множество, мало что меняется в анализе). Зафиксируем $a_{(j)}$, $j = 1, \ldots, k$. Заметим, что

$$g(a_{(j)}) = \mathbb{E}\left[y|x = a_{(j)}\right] = \frac{\mathbb{E}\left[y\mathbb{I}\left\{x = a_{(j)}\right\}\right]}{\mathbb{E}\left[\mathbb{I}\left\{x = a_{(j)}\right\}\right]}$$

*Работа основана на лекциях, читаемых автором в РЭШ. Цитировать как: Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37–52. Citation: Anatolyev, Stanislav (2009) “Nonparametric regression,” Quantile, No.7, pp. 37–52.

†Адрес: 117418, г. Москва, Нахимовский проспект, 47, офис 1721(3). Электронная почта: *sanatoly@nes.ru*

из-за справедливости следующих равенств:

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}\left\{x=a_{(j)}\right\}\right] &= \mathbb{P}\{x_i=a_{(j)}\},\\ \mathbb{E}\left[y\mathbb{I}\left\{x=a_{(j)}\right\}\right] &= \mathbb{E}\left[y|x=a_{(j)}\right]\mathbb{P}\{x=a_{(j)}\}.\end{aligned}$$

Согласно принципу аналогий можно построить оценку $\hat{g}(a_{(j)})$ как

$$\hat{g}(a_{(j)})=\frac{\sum_{i=1}^{n}y_i\mathbb{I}\left\{x_i=a_{(j)}\right\}}{\sum_{i=1}^{n}\mathbb{I}\left\{x_i=a_{(j)}\right\}}, \qquad (1)$$

что можно интерпретировать как среднее по наблюдениям, попадающим в вертикальное сечение $x=a_{(j)}$.

2.2 Непрерывный регрессор

Если регрессор непрерывно распределен, описанный метод не работает, так как в произвольном сечении $x=a$ нечего усреднять (ибо туда попадет максимум одно наблюдение), хотя и $f(a)\neq 0$, где $f(a)$ – значение плотности регрессора $f(x)$ в a. Поэтому необходимо привлечь информацию откуда-то еще. Если регрессионная кривая непрерывна, наблюдения, попадающие в окрестность a, являются наиболее информативными о значении регрессии в a.

Выберем положительное h и назовем его *шириной окна* (хотя впоследствии необязательно будет задействовано окно в буквальном смысле). Обобщим формулу (1) следующим образом:

$$\hat{g}(a)=\frac{\sum_{i=1}^{n}y_i\mathbb{I}\left\{a-h\le x_i\le a+h\right\}}{\sum_{i=1}^{n}\mathbb{I}\left\{a-h\le x_i\le a+h\right\}}, \qquad (2)$$

т.е. мы усредняем y по наблюдениям, попадающим в окно $[a-h,a+h]$. При изменении a $\hat{g}(a)$ описывает оцененную регрессионную кривую. Отметим, что последняя имеет разрывы из-за попадания в окно новых наблюдений и выпадения из него старых.

Информация от наблюдений, попавших в окно $[a-h,a+h]$, используется одинаково. То есть те наблюдения, которые попали в окно, участвуют в оценивании с равным весом, в то время как наблюдения, не попавшие в окно, вообще в нем не участвуют. Разумной представляется идея сделать веса зависимыми от расстояния от x_i до a, а также, возможно, использовать информацию во всех наблюдениях. Введем с этой целью симметричную *ядерную функцию* $K(u)$, интегрирующуюся в единицу, т.е. $\int K(u)\,du=1$, где интеграл $\int$ берется по всей области определения, которой может быть отрезок, обычно $[-1,1]$, или вся числовая ось. Популярными ядрами являются:

$$\begin{aligned}\text{Равномерное:}\quad & K(u)=\frac{1}{2}\mathbb{I}\left\{|u|\le 1\right\},\\ \text{Треугольное:}\quad & K(u)=(1-|u|)\,\mathbb{I}\left\{|u|\le 1\right\},\\ \text{Епанечниково:}\quad & K(u)=\frac{3}{4}\left(1-u^2\right)\mathbb{I}\left\{|u|\le 1\right\},\\ \text{Гауссово (нормальное):}\quad & K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right).\end{aligned}$$

Область определения первых трех ядер – отрезок $[-1,1]$, в то время как последнее имеет бесконечный носитель. Следовательно, при использовании равномерного, треугольного или Епанечникова ядра оценка будет использовать информацию в ограниченном окне в окрестности a, а оценка, использующая гауссово ядро, будет использовать информацию из всех наблюдений. Заметим также, что в принципе ядро не обязано быть всюду неотрицательным.

Далее, введем обозначение

$$K_h(u)=\frac{1}{h}K\left(\frac{u}{h}\right).$$

Теперь обобщим формулу (2) и получим

$$\hat{g}(a) = \frac{\sum_{i=1}^{n} y_i K_h (x_i - a)}{\sum_{i=1}^{n} K_h (x_i - a)}. \tag{3}$$

Оценка (3) называется *оценкой Надарайа–Уотсона* для регрессии среднего, в честь Nadaraya (1965) и Watson (1964). Заметим, что нормализация делением на h в определении $K_h(u)$ не влияет на численное значение оценки и сделано лишь для удобства, в чем мы убедимся позднее.

Заметим, что при использовании трех последних ядер (треугольного, Епанечникова и гауссова), так же как и многих других, оцененная регрессионная кривая непрерывна, так как новые и старые наблюдения вводятся в формулу и выводятся из нее непрерывным образом по мере того как a меняется. При этом велика роль параметра ширины окна. Если h слишком велика, оценка задействует слишком много нерелевантной информации, что увеличивает смещение и приводит к явлению *сверхсглаживания*. Сверхсглаженная кривая слишком «линейная», в то время как ей положено быть более извилистой и более пристально отслеживать рисунок наблюдений. Если же h слишком мала, оценка задействует маловато точек, что увеличивает дисперсию и приводит к явлению *недосглаживания*. Недосглаженная кривая слишком извилистая, так как она слишком пристально отслеживает отдельные наблюдения.

3 Асимптотические свойства

В случае дискретного регрессора легко получить, используя ЗБЧ и ЦПТ, что

$$\sqrt{n}\left(\hat{g}(a_{(j)}) - g(a_{(j)})\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{V}\left[y|x = a_{(j)}\right]}{\mathbb{P}\{x = a_{(j)}\}}\right). \tag{4}$$

Интерпретация выражения для асимптотической дисперсии следующая. Качество оценивания положительно связано с частотой попадания точек в вертикальное сечение $x = a_{(j)}$ и отрицательно связано со степенью разброса точек вдоль него. Заметим, что скорость сходимости оценки – параметрическая, $\sqrt{n}$. Действительно, задачу можно трактовать как параметрическую, ибо вектор параметров $\left(a_{(1)}, \ldots, a_{(k)}\right)'$ конечномерен.

В случае непрерывно распределенного регрессора необходимо, чтобы ширина окна асимптотически падала до нуля, иначе смещение из-за нерелевантности используемой информации из соседних к a наблюдений испортит состоятельность оценки. Таким образом, необходимо установить правило $h \to 0$ по мере того как $n \to \infty$. С другой стороны, ширина окна не должна падать слишком быстро, иначе дисперсия из-за малого количества участвующих в оценивании точек не будет падать, что также испортит состоятельность оценки. Точнее, поскольку дисперсия обратно пропорциональна эффективному количеству участвующих наблюдений, которое в свою очередь прямо пропорционально nh, необходимо установить правило $nh \to \infty$ по мере того как $h \to 0$ и $n \to \infty$.

Обозначим

$$\sigma_K^2 = \int u^2 K(u)\, du$$

и

$$R_K = \int K(u)^2\, du.$$

Эти две константы зависят только от выбранного ядра. Подразумевается, что обе величины конечны. Далее, установим дополнительное требование к скорости падения ширины окна:

$$\lambda = \lim_{n\to\infty} \sqrt{nh^5},$$

предполагая $\lambda < \infty$. Заметим, что λ может равняться, а может и не равняться нулю. Мы также предполагаем непрерывность и ограниченность $g(x)$, $g'(x)$, $g''(x)$, $f(x)$ и $f'(x)$ всюду на области определения.

Рассмотрим разницу между оценкой и оцениваемой величиной:

$$\hat{g}(a) - g(a) = \frac{\hat{q}_1(a) + \hat{q}_2(a)}{\hat{f}(a)},$$

где

$$\begin{aligned}
\hat{q}_1(a) &= \frac{1}{n}\sum_{i=1}^{n} e_i K_h(x_i - a), \\
\hat{q}_2(a) &= \frac{1}{n}\sum_{i=1}^{n} (g(x_i) - g(a)) K_h(x_i - a), \\
\hat{f}(a) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - a),
\end{aligned}$$

и за e_i обозначены, как обычно, регрессионные ошибки в точках выборки: $e_i = y_i - g(x_i)$.

Начнем со знаменателя

$$\hat{f}(a) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - a).$$

Величина $\hat{f}(a)$, называемая *оценкой плотности Надарайа–Уотсона*, и правда оценивает плотность регрессора $f(x)$ в точке a, отсюда и обозначения. Действительно, рассмотрим

$$\begin{aligned}
\mathbb{E}[\hat{f}(a) - f(a)] &= \mathbb{E}[K_h(x-a)] - f(a) \\
&= \frac{1}{h}\int K\left(\frac{x-a}{h}\right) f(x)\,dx - f(a) \\
&= \int K(u) f(a+hu)\,du - f(a) \\
&= \int K(u) f(a+hu)\,du - f(a).
\end{aligned}$$

Разложим $f(a+hu)$ до первого порядка:

$$\begin{aligned}
\mathbb{E}[\hat{f}(a) - f(a)] &= \int K(u)\big(f(a) + f'(a)hu + o(h)\big)\,du - f(a) \\
&= f(a)\int K(u)\,du + f'(a)h\int uK(u)\,du + o(h) - f(a) \\
&= o(h),
\end{aligned}$$

так как ядро интегрируется в единицу и симметрично. Использую ту же технологию,

$$\begin{aligned}
\mathbb{V}[\hat{f}(a)] &= \frac{1}{n}\mathbb{V}[K_h(x-a)] \\
&= \frac{1}{n}\mathbb{E}\left[K_h(x-a)^2\right] - \frac{1}{n}\mathbb{E}[K_h(x-a)]^2 \\
&= \frac{1}{n}\int\left(\frac{1}{h}K\left(\frac{x-a}{h}\right)\right)^2 f(x)\,dx - \frac{1}{n}\left(\int\frac{1}{h}K\left(\frac{x-a}{h}\right)f(x)\,dx\right)^2 \\
&= \frac{1}{n}\frac{1}{h}\int K(u)^2 f(a+hu)\,du - \frac{1}{n}\left(\int K(u) f(a+hu)\,du\right)^2 \\
&= \frac{1}{nh}\int K(u)^2 (f(a) + o(1))\,du - \frac{1}{n}O(1) \\
&= O\left(\frac{1}{nh}\right).
\end{aligned}$$

Поскольку $h \to 0$ и $nh \to \infty$, имеем $\mathbb{E}[\hat{f}(a) - f(a)] \to 0$ и $\mathbb{V}[\hat{f}(a)] \to 0$, откуда следует, что действительно $\hat{f}(a) \xrightarrow{p} f(a)$.

Теперь рассмотрим первую часть числителя

$$\hat{q}_1(a) = \frac{1}{n}\sum_{i=1}^{n} e_i K_h(x_i - a).$$

Это среднее регрессионных ошибок, взвешенных ядром. Как и в параметрическом анализе, такое среднее должно дать асимптотическую нормальность. Разница в том, что дисперсия отдельного слагаемого здесь не постоянна, а зависит от асимптотически падающего h. Следовательно, необходимо рассчитать предел этой дисперсии. Представим $\sqrt{nh}\hat{q}_1(a)$ как

$$\sqrt{nh}\hat{q}_1(a) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{e_i}{\sqrt{h}}K\left(\frac{x_i - a}{h}\right).$$

Дисперсия отдельного слагаемого равна

$$\begin{aligned}
\mathbb{V}\left[\frac{e}{\sqrt{h}}K\left(\frac{x-a}{h}\right)\right] &= \mathbb{E}\left[\frac{\sigma^2(x)}{h}K\left(\frac{x-a}{h}\right)^2\right] \\
&= \frac{1}{h}\int \sigma^2(x) K\left(\frac{x-a}{h}\right)^2 f(x)\,dx \\
&= \int \sigma^2(a+hu) K(u)^2 f(a+hu)\,du \\
&= \int \sigma^2(a) K(u)^2 f(a)\,du + o(1) \\
&= \sigma^2(a) f(a) R_K + o(1).
\end{aligned}$$

По ЦПТ Линдберга–Леви, $\sqrt{nh}\hat{q}_1(a) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(a) f(a) R_K\right)$.

Наконец, рассмотрим вторую часть числителя

$$\hat{q}_2(a) = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - g(a)) K_h(x_i - a).$$

Это среднее взвешенных ядром отклонений значений регрессионной функции в точках наблюдений от ее значения в точке a, где она оценивается. Эти отклонения рождают смещение. Конечно же, $\hat{q}_2(a)$ обладает и дисперсией, но она мала по сравнению с дисперсией, рождаемой в $\hat{q}_1(a)$ регрессионными ошибками.

Рассмотрим математическое ожидание $\hat{q}_2(a)$:

$$\begin{aligned}
\mathbb{E}[\hat{q}_2(a)] &= \mathbb{E}\left[(g(x) - g(a)) K_h(x-a)\right] \\
&= \frac{1}{h}\int (g(x) - g(a)) K\left(\frac{x-a}{h}\right) f(x)\,dx \\
&= \int (g(a+hu) - g(a)) K(u) f(a+hu)\,du \\
&= \int \left(g'(a) hu + \frac{g''(a)}{2}(hu)^2 + o\left(h^2\right)\right) K(u) \left(f(a) + f'(a) hu + o(h)\right) du \\
&= g'(a) f(a) h \int uK(u)\,du \\
&\quad + \left(g'(a) f'(a) + \frac{g''(a)}{2} f(a)\right) h^2 \int u^2 K(u)\,du + o\left(h^2\right) \\
&= h^2 f(a) B(a) \sigma_K^2 + o\left(h^2\right),
\end{aligned}$$

где

$$B(a) = \frac{g'(a) f'(a)}{f(a)} + \frac{g''(a)}{2}.$$

Легко также получить, что

$$\mathbb{V}[\hat{q}_2(a)] = o\left(\frac{1}{nh}\right).$$

Эти два результата приводят к тому, что

$$\sqrt{nh}\hat{q}_2(a) \xrightarrow{p} \lambda f(a) B(a) \sigma_K^2.$$

Все выведенное выше в совокупности дает

$$\begin{aligned} \sqrt{nh}\left(\hat{g}(a) - g(a)\right) &= \frac{\sqrt{nh}\hat{q}_1(a) + \sqrt{nh}\hat{q}_2(a)}{\hat{f}(a)} \\ &\xrightarrow{d} \frac{\mathcal{N}\left(0, \sigma^2(a) f(a) R_K\right) + \lambda f(a) B(a) \sigma_K^2}{f(a)} \\ &\sim \mathcal{N}\left(\lambda B(a) \sigma_K^2, \frac{\sigma^2(a)}{f(a)} R_K\right). \end{aligned}$$

Отметим две интересные черты этого асимптотического результата. Во-первых, непараметрическая скорость сходимости $\sqrt{nh}$ меньше, чем параметрическая $\sqrt{n}$, поскольку асимптотически $h \to 0$. Этот факт отражает меньшую точность оценивания бесконечномерных объектов, чем точность оценивания конечномерных. Во-вторых, хотя асимптотическое распределение и нормальное, оно нецентрированное. Асимптотическое смещение отражает тот факт, что информация, используемая при оценивании, не является всецело релевантной.

Формула для асимптотической дисперсии похожа на свой аналог в случае дискретного регрессора: скедастическая функция в числителе и «вероятностная масса» в знаменателе. Асимтотическое смещение зависит от множества характеристик форм регрессионной и плотностной функций, зашитых в формуле для $B(a)$. Одна часть асимтотического смещения пропорциональна $g'(a)f'(a)$ и отражает смещение, возникающее, когда регрессионная кривая имеет наклон, а наблюдения падают несимметрично слева и справа от a, в результате чего точки слева и точки справа создают неодинаковое и взаимно не компенсирующееся смещение вверх и вниз. Вторая часть асимтотического смещения пропорциональна $g''(a)$ и отражает смещение, возникающее, когда регрессионная кривая локально нелинейна, в результате чего ординаты точек слева и точек справа от a несимметрично распределены выше и ниже $g(a)$, даже если их абсциссы распределены симметрично слева и справа от a. Отметим, что и дисперсия, и смещение в общем случае обратно пропорциональны плотности $f(a)$, что отражает тот факт, что точность оценивания, и в смысле дисперсии, и в смысле смещения, низка при оценивании $g(a)$ около тех границ носителя x, где плотность убывает в ноль.

Полученный асимптотический результат также означает, что оптимальная скорость падения ширины окна $h \propto n^{-1/5}$, так как в этом случае $\lambda > 0$, и асимптотические смещение и дисперсия уравновешены. С другой стороны, если положить $h = o\left(n^{-1/5}\right)$, можно добиться того, что λ будет равно нулю, и асимптотическое смещение исчезнет. Конечно, это удобно с точки зрения реализации, т.к. в таком случае нет необходимости оценивать компоненты $B(a)$, но подобные действия скрывают истинную картину, связанную со смещением оценки, и могут привести к плохому качеству асимптотического приближения.

Выведенный выше асимптотический результат означает, что приближенно

$$\hat{g}(a) \sim \mathcal{N}\left(g(a) + \frac{\sqrt{nh^5} B(a) \sigma_K^2}{\sqrt{nh}}, \frac{\sigma^2(a)}{f(a)} \frac{R_K}{nh}\right).$$

Как обычно, асимптотическое распределение можно использовать для тестирования статистических гипотез о $g(a)$ и построения доверительных интервалов для этой величины. Доверительный интервал, например, выглядит так:

$$\hat{g}(a) - h^2 \hat{B}\left(a\right) \sigma_K^2 \mp z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}^2\left(a\right)}{\hat{f}\left(a\right)} \frac{R_K}{nh}},$$

где $z_{1-\alpha/2}$ – $(1-\alpha/2)$-квантиль стандартного нормального распределения, а $\hat{f}\left(a\right)$, $\hat{\sigma}^2\left(a\right)$ и $\hat{B}\left(a\right)$ – непараметрические оценки соответствующих функций в точке a. Заметим, что если построить доверительные интервалы для $g(a)$ при всех значениях a (на решетке) с вероятностью покрытия $1-\alpha$, возникнет *поточечный доверительный коридор* для $g(x)$. Естественно, неверно говорить, что истинная регрессионная кривая находится внутри этого коридора с вероятностью $1-\alpha$.

4 Выбор ширины окна

4.1 Правило подстановки

Из выведенного выше результата следует, что асимптотическая среднеквадратическая ошибка оценивания равна

$$\mathcal{AMSE}(a) = h^4 B\left(a\right)^2 \sigma_K^4 + \frac{\sigma^2\left(a\right)}{f\left(a\right)} \frac{R_K}{nh}.$$

Если минимизировать эту величину по h, то возникнет *правило подстановки* для *(локально) оптимальной ширины окна*

$$h^*(a) = \left(\frac{\sigma^2\left(a\right)}{4f\left(a\right) B\left(a\right)^2} \frac{R_K}{\sigma_K^4}\right)^{1/5} n^{-1/5}.$$

Заметим, что скорость падения равна оптимальной, выведенной выше.

Практическое применение правила подстановки заключается в (непараметрическом!) оценивании $f\left(a\right), f'\left(a\right), g'\left(a\right), g''\left(a\right), \sigma^2\left(a\right)$, вычислении R_K, σ_K^4 и подстановке полученных результатов в формулу для $h^*(a)$. Эта трудоемкая процедура дает численное значение оптимальной ширины окна всего для одного значения a, так что ее приходится повторить для всех a. Конечно, это очень нелегко, и исследователи чувствовали бы себя комфортнее с одним-единственным значением *глобально оптимальной ширины окна* h^*, общей для всех a.

Глобально оптимальную ширину окна легко вывести, минимизируя критерий интегрированной асимптотической среднеквадратической ошибки оценивания

$$\mathcal{IAMSE}(a) = \int \mathcal{AMSE}(a) da = h^4 \sigma_K^4 \int B\left(a\right)^2 da + \frac{R_K}{nh} \int \frac{\sigma^2\left(a\right)}{f\left(a\right)} da.$$

Естественно, можно ввести какую-либо взвешивающую схему, если есть причины не использовать равномерное взвешивание для разных a. При равномерном взвешивании

$$h^* = \left(\frac{\int \sigma^2\left(a\right)/f\left(a\right) da}{4\int B\left(a\right)^2 da} \frac{R_K}{\sigma_K^4}\right)^{1/5} n^{-1/5}.$$

Данная стратегия, однако, не лишает исследователя необходимости оценивать $f\left(a\right)$, $f'\left(a\right)$, $g'\left(a\right)$, $g''\left(a\right)$, $\sigma^2\left(a\right)$, то есть процедура настолько же трудоемкая, как и прежде. Статистик Бернард Сильверман предложил универсальную формулу, основанную на вышеописанной процедуре, но специфичной для определенного частного случая, например, в предположении нормальной плотности f.[1] Вот эта универсальная формула, называемая *правилом Сильвермана*:

$$h^S = 1.364 \left(\frac{R_K}{\sigma_K^4} \right)^{1/5} \hat{\sigma}_x n^{-1/5},$$

где $\hat{\sigma}_x^2$ – выборочная дисперсия регрессора. В частности, для гауссова ядра формула выглядит как $h^S = 1.06 \hat{\sigma}_x n^{-1/5}$.

На практике правило Сильвермана обычно обеспечивает приемлемые результаты, кроме, возможно, тех случаев, когда оценивание происходит вблизи краев носителя x. Иногда, впрочем, вид оцененной регрессионной кривой неудовлетворителен, и подобную ширину окна используют лишь как стартовое значение при поиске более приемлемой.

4.2 Кросс-валидация

Радикально иное правило выбора глобальной ширины окна – это *кросс-валидация*. Она основана на качестве подгонке, а не на асимптотических свойствах оценки. Можно обычную среднеквадратическую ошибку, оцененную в точках выборки,

$$\mathcal{MSE}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2,$$

устремить к нулю, что приводит к идеальной подгонке, если положить h равным своему наименьшему значению, когда каждое наблюдение объясняется только им же самим. Естественно, это не что иное как экстремальная степень недосглаживания, и оно нас не устраивает. От источника проблемы можно легко избавиться, если запретить объяснять наблюдение самим собой. Исходя из этого, разумным критерием будет *функция кросс-валидации*

$$\mathcal{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{g}^{-i}(x_i) \right)^2,$$

где $\hat{g}^{-i}(x_i)$ – это оценка Надарайа–Уотсона в точке x_i, которая (оценка) использует все наблюдения, за исключением i-го, т.е.

$$\hat{g}^{-i}(x_i) = \frac{\sum_{j=1, j \neq i}^{n} y_i K_h (x_j - x_i)}{\sum_{j=1, j \neq i}^{n} K_h (x_j - x_i)}.$$

Оптимальная в смысле кросс-валидации ширина окна h^{CV} минимизирует $\mathcal{CV}(h)$. К сожалению, при практическом применении данная ширина окна часто приводит к сильному недосглаживанию. Следовательно, в этих случаях ее можно использовать в качестве начального приближения для приемлемой ширины окна, и окончательное решение опять же остается за визуальным анализом.

5 Многопеременная ядерная регрессия

До сих пор регрессор x был единственным. Как обобщить оценку Надарайа–Уотсона на постановку с множественными регрессорами?

Обозначим через d размерность x. У ядра теперь будет d-мерный аргумент, отображающий расстояние между a и каждым x_i в d-мерном пространстве в скалярный вес: $K : \mathbb{R}^d \to \mathbb{R}$.

[1] На самом деле Сильверман использовал оптимальную ширину окна для оценки плотности Надарайа–Уотсона. См. одну из задач в конце эссе.

Определим $d \times d$-мерную симметричную и положительно определенную *матрицу ширины окна* H. Далее, положим

$$K_H(u) = \frac{1}{\det H} K\left(H^{-1}u\right).$$

Оценка Надарайа–Уотсона выглядит по-прежнему как

$$\hat{g}(a) = \frac{\sum_{i=1}^{n} y_i K_H(x_i - a)}{\sum_{i=1}^{n} K_H(x_i - a)}.$$

Есть много способов организовать структуру K и H. На практике широко используются два способа.

Первый использует разложение d-мерного пространства на произведение d одномерных:

$$K_H(u) = \prod_{\ell=1}^{d} K_{h_\ell,\ell}(u_\ell) = \prod_{\ell=1}^{d} \frac{1}{h_\ell} K_\ell\left(\frac{u_\ell}{h_\ell}\right),$$

где K_ℓ и h_ℓ – ядро и ширина окна, соответственно, по ℓ-ой размерности. Такое $K_H(u)$ называется *ядром-произведением.* K_ℓ потенциально могут быть разными по разным координатам, но обычно они одинаковы. Матрица ширины окна равна $H = \operatorname{diag}\{h_\ell\}_{\ell=1}^{d}$, и каждая h_ℓ выбирается отдельно (например, в простейшем случае – с помощью правила Сильвермана).

Ядро-произведение игнорирует зависимость между регрессорами. Кроме того, выбор d значений ширины окна не больно-то привлекателен. У второго метода нет этих недостатков. Матрица ширины окна имеет следующую структуру:

$$H = hS^{1/2},$$

где h – единая ширина окна, S – выборочная дисперсионная матрица регрессоров, а $S^{1/2}$ – квадратный корень из нее (например, определяемый через разложение Холески).

К сожалению, точность оценивания при больших значениях d мала. Это явление называют *проклятием размерности.* Причина в том, что при прочих равных в d-мерное гиперокно попадает намного меньше наблюдений, чем в его одномерный аналог, и дисперсия оценки быстро возрастает с увеличением d. В частности, скорость сходимости (при второй схеме) равна $\sqrt{nh^d}$, а оптимальная ширина окна – $O\left(n^{-1/(d+4)}\right)$. На практике непараметрическое оценивание обычно возможно только при очень малом количестве регрессоров d вроде 1, 2, 3, редко выше, и требует больших выборок для достижения приличной надежности оценивания.

6 Локальная полиномиальная регрессия

Вернемся к скалярному x. Заметим, что оценку Надарайа–Уотсона можно представить в виде решения задачи минимизации взвешенной суммы квадратов:

$$\hat{g}(a) = \arg\min_{\beta_0} \sum_{i=1}^{n} (y_i - \beta_0)^2 K_h(x_i - a).$$

Это означает, что оценивание Надарайа–Уотсона – это локальная (в том смысле, что взвешивание основано на локальности наблюдений к a) регрессия на константе. Нет причин останавливаться на регрессии на константе. Естественной является идея расширить ее на линейную регрессию не только на константе, но также и на x. Результатом будет *локальная линейная регрессия*:

$$\hat{g}_1(a) = (1,0) \arg\min_{(\beta_0,\beta_1)'} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1(x_i - a))^2 K_h(x_i - a).$$

Вектор $(1, 0)$ отбирает первый элемент вектора 2×1. В качестве сопутствующего результата второй элемент этого вектора дает локально линейную оценку наклона регрессионной кривой в a, то есть ее первой производной $g'(a)$.

Оценка локальной линейной регрессии имеет то преимущество перед оценкой Надарайа–Уотсона, что она учитывает наклонность регрессионной кривой, что имеет значение, если наблюдения распределены неравномерно вокруг a, и таким образом уменьшает смещение. Это проявляется в асимптотических свойствах оценки, идентичных свойствам оценки Надарайа–Уотсона, кроме того момента, что теперь

$$B(a) = \frac{g''(a)}{2}.$$

Локальную полиномиальную регрессию можно обобщить далее и получить *локальную полиномиальную регрессию порядка* p:

$$\hat{g}_p(a) = (1, 0, \ldots, 0) \arg \min_{(\beta_0, \ldots, \beta_p)'} \sum_{i=1}^{n} \left(y_i - \beta_0 - \ldots - \beta_p (x_i - a)^p\right)^2 K_h(x_i - a).$$

Данная оценка учитывает не только наклонность, но и свойства кривизны регрессионной кривой в точке a, и в качестве сопутствующего результата дает оценки производных $g(x)$ до p-го порядка в точке a. Удобно, что оценку можно записать в явном виде как оценку взвешенного метода наименьших квадратов:

$$\hat{g}_p(a) = (1, 0, \ldots, 0) \left(X'WX\right)^{-1} X'WY,$$

где $Y = (y_1, \ldots, y_n)'$, $X = (X_1, \ldots, X_n)'$, $X_i = (1, x_i - a, \ldots, (x_i - a)^p)'$, и, наконец, $W = \operatorname{diag}\{K_h(x_i - a)\}_{i=1}^{n}$. Если $p > 1$, асимптотическое смещение $B(a)$ равно нулю. Это означает, что оптимальная ширина окна и результирующая скорость сходимости оценки меняются.

При практическом применении существуют серьезные недостатки использования локальной полиномиальной регрессии. Использование информации в локальных непараметрических методах очень ограничено, а увеличение p означает уменьшение степеней свободы. В качестве экстремального примера рассмотрим равномерное ядро и узкое окно, настолько узкое, что в него попадает лишь два наблюдения. Оценка Надарайа–Уотсона усредняет ординаты этих двух наблюдений, локальная линейная регрессия соединяет их прямой и берет ординату пересечения с вертикальной прямой $x = a$, а локальная полиномиальная регрессия при $p > 1$ попросту не существует.

На практике стоит ограничиваться малыми p, вроде 0, 1 или 2, не больше.

7 Временные ряды

Если вместо случайной выборки у нас стационарный временный ряд, основные методы, приведенные выше, в целом работают. Интересно, что, в отличие от параметрических задач, в асимптотической дисперсии не возникают автоковариации y_t в $x_t = a$. Причина в том, что эффект последних асимптотически исчезает на фоне дисперсии y_t в $x_t = a$ (см. Bierens, 1994). Таким образом, можно использовать те же самые формулы.

Типичное приложение непараметрических методов во временных рядах – непараметрическая авторегрессия

$$y_t = g(y_{t-1}, y_{t-2}, \ldots, y_{t-k}) + \sigma(y_{t-1}, y_{t-2}, \ldots, y_{t-k}) \eta_t$$

где $\mathbb{E}[\eta_t | y_{t-1}, y_{t-2}, \ldots] = 0$ и $\mathbb{V}[\eta_t | y_{t-1}, y_{t-2}, \ldots] = 1$.

Оценку Надарайа–Уотсона $\hat{g}\left(a_1,\ldots,a_k\right)$ авторегрессионной функции $g\left(y_{t-1},\ldots,y_{t-k}\right)$ в $\left(y_{t-1},\ldots,y_{t-k}\right)=\left(a_1,\ldots,a_k\right)$ можно построить по знакомой схеме, а оценку автоскедастичной функции $\sigma^2\left(y_{t-1},\ldots,y_{t-k}\right)$ как

$$\hat{\sigma}^2\left(y_{t-1},\ldots,y_{t-k}\right)=\hat{\delta}\left(a_1,\ldots,a_k\right)-\hat{g}\left(a_1,\ldots,a_k\right)^2,$$

где $\hat{\delta}\left(a_1,\ldots,a_k\right)$ – оценка Надарайа–Уотсона непараметрической регрессии y_t^2 на $\left(y_{t-1},\ldots,y_{t-k}\right)$ в точке $\left(y_{t-1},\ldots,y_{t-k}\right)=\left(a_1,\ldots,a_k\right)$. Если порядок авторегрессии k неизвестен, его можно оценить совместно с регрессионной функцией (см., например, Tschernig & Yang, 2000). Обзор непараметрических методов в контексте временных рядов содержится в Heiler (2001).

8 Другие непараметрические методы

Непараметрический метод может быть одного из двух типов: локальный или глобальный. Оценка Надарайа–Уотсона, локальная линейная и локальная полиномиальная регрессии принадлежат классу локальных методов, так как оценивание $g(x)$ в точке a использует информацию в наблюдениях, находящихся вблизи a. Глобальные непараметрические методы вместо этого пытаются подогнать всю кривую ко всем точкам выборки одновременно. При этом влияние одного наблюдения не так ограничено, и значения его абсциссы и ординаты влияют не только на оцененную регрессию поблизости этой точки, но и на положение и форму всей оцененной регрессионной кривой.

Независимо от того, локальный метод или глобальный, всегда наличествует *параметр сглаживания*, контролирующий степень последнего. Выбор этого параметра осуществляется исследователем. Параметр сглаживания в уже рассмотренных методах – это ширина окна.

8.1 Метод ближайших соседей

Другим локальным непараметрическим методом является *оценка k ближайших соседей*

$$\hat{g}_{NN}(a)=\frac{1}{k}\sum_{i=1}^{n}y_i 1_i,$$

где $1_i=\mathbb{I}\left\{x_i \text{ является одним из } k \text{ ближайших соседей к } a\right\}$, и о соседстве судится по тому, насколько близки абсциссы точек выборки к a. Здесь k является параметром сглаживания, и для состоятельности необходимы условия $k\to\infty$ и $k/n\to 0$ по мере того как $n\to\infty$. Другой версией является *симметризованная оценка k ближайших соседей*. Предполагая, что k четно,

$$\hat{g}_{SNN}(a)=\frac{1}{k}\sum_{i=1}^{n}y_i 1_i,$$

где $1_i=\mathbb{I}\left\{x_i \text{ является одним из левых } k/2 \text{ или правых } k/2 \text{ ближайших соседей к } a\right\}$. Асимптотические свойства этих оценок схожи со свойствами оценки Надарайа–Уотсона.

Одним из преимуществ методов ближайших соседей перед ядерными методами является существование оценки при любом раскладе, так как у любой точки есть соседи в любой выборке, в то время как в окно могут не попасть наблюдения совсем. Кроме того, количество усредняемых величин контролируется напрямую. Конечно же, обе идеи можно скомбинировать.

8.2 Разложение в серии (решето)

Оценивание сериями, или решетом, – один из глобальных методов, основанный на разложении гладкой функции по базису в функциональном пространстве. Положим, существует, предпочтительно ортогональный, упорядоченный базис $\left\{\psi_j(x)\right\}_{j=0}^{\infty}$, такой, что

$$g(x) = \sum_{j=0}^{\infty} \gamma_j \psi_j(x).$$

Термин «упорядоченный» означает, что $\psi_0(x)$ наиболее важный член, $\psi_1(x)$ менее важный, $\psi_2(x)$ еще менее важный и т.д. Например, базисом могут быть полиномы Эрмита или синусы и косинусы Фурье. Выберем параметр отсечения J, такой, что асимптотически $J \to \infty$ и $J/n \to 0$ по мере того как $n \to \infty$. Тогда

$$\hat{g}_S(x) = \sum_{j=0}^{J} \hat{\gamma}_j \psi_j(x),$$

где $(\hat{\gamma}_0, \ldots, \hat{\gamma}_J)'$ – вектор МНК-оценок в «линейной регрессии» y на $(\psi_0(x), \ldots, \psi_J(x))'$. Этот метод также иногда называют полунепараметрическим, так как он непараметрический по сути, но параметрический по технической реализации. См. обзор Chen (2007), а также статью Кристенсен (2009) в настоящем номере журнала.

8.3 Искусственные нейронные сети

Идея искусственных нейронных сетей похожа на идею решета, т.е. аппроксимацию нелинейной функции линейной комбинацией неких базовых функций, в таком количестве, в каком необходимо для достижения хорошего приближения. В качестве базисных обычно используются логистические функции:

$$\hat{g}_{ANN}(x) = \hat{\phi}_{0,0} + \hat{\phi}_{1,0}x + \sum_{j=1}^{J} \frac{\hat{\gamma}_j}{1 + \exp(\hat{\phi}_{0,j} + \hat{\phi}_{1,j}x)},$$

где коэффициенты оцениваются нелинейным методом наименьших квадратов вместо МНК, используемого в решете. Расширенные версии задействуют интересные иерархические структуры, см. Franses & van Dijk (2000, глава 5).

8.4 Сплайны

Другим глобальным методом является оценивание сплайнами. Положим, мы хотим подогнать всю регрессионную кривую, используя обычный критерий подгонки, сумму квадратов ошибок. Это, конечно же, немедленно приведет к интерполяции, т.е. недосглаживанию максимальной степени. Однако мы можем добавить штрафной член, наказывающий за недосглаживание. Интерполирующая кривая слишком извилистая, то есть обладает большой второй производной по абсолютной величине. Эти идеи приводят к следующей *оценке сплайнами*:

$$\hat{g}_{CS}(x) = \arg\min_{\hat{g}(x)} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2 + \lambda \int \left(\hat{g}''(u)\right)^2 du,$$

где параметром сглаживания является λ. Если λ мало, решение близко к интерполирующей кривой, в то время как если λ очень большое, решение близко к линейному предиктору с точки зрения критерия наименьших квадратов. Индекс CS расшифровывается как «кубический сплайн», что означает решение в виде кусочно кубического полинома с непрерывно дифференцируемыми переходами в абсциссах точек наблюдений. Серьезное пособие по сплайнам – Wahba (1990).

9 Задачи

9.1 Оценка плотности Надарайа–Уотсона

Вывести асимптотическое распределение оценки Надарайа–Уотсона плотности скалярной случайной величины x, имеющей непрерывное распределение, аналогично тому, как выведено асимптотическое распределение оценки Надарайа–Уотсона регрессионной функции, при аналогичных предположениях. Дать интерпретацию зависимости выражений для асимптотического смещения и асимптотической дисперсии от формы плотности.

9.2 Несмещенность ядерных оценок

Рассмотрим оценку Надарайа–Уотсона $\hat{g}\,(x)$ условного среднего $g\,(x)$ для случайной выборки. Показать, что если $g\,(x) = c$, где c – некоторая константа, то оценка $\hat{g}\,(x)$ несмещена. Какова интуиция за этим результатом? Выяснить, при каких обстоятельствах оценка локальной линейной регрессии $g\,(x)$ будет несмещена. Будет ли оценка плотности $f\,(x)$ несмещена?

9.3 Оценивание при ограничении на форму

Фирмы производят продукт, используя технологию $f(l, k)$. Функциональная форма f неизвестна, но известно, что она обладает свойством постоянного эффекта от масштаба. Для фирмы i наблюдается труд l_i, капитал k_i и выпуск y_i, а порождающий данные процесс принимает форму $y_i = f(l_i, k_i) + \varepsilon_i$, где $\mathbb{E}\,[\varepsilon_i] = 0$, и ошибка ε_i независима от (l_i, k_i). Для случайной выборки $\{y_i, l_i, k_i\}_{i=1}^n$ предложить непараметрическую оценку $f(l, k)$, которая бы тоже обладала свойством постоянного эффекта от масштаба.

9.4 Непараметрическая функция риска

Пусть $z_1, \ldots, z_n$ – скалярные независимые одинаково распределенные случайные величины с неизвестной плотностью $f(\cdot)$ и функцией распределения $F(\cdot)$. Предположим, что распределение z имеет носитель $\mathbb{R}$. Возьмем $t \in \mathbb{R}$, такое что $0 < F(t) < 1$. Целью является оценивание *функции риска*

$$H(t) = \frac{f(t)}{1 - F(t)}.$$

Предложить непараметрическую оценку $\hat{F}(t)$ для $F(t)$. Обозначим за $\hat{f}(t)$ оценку Надарайа–Уотсона для $f(t)$, и выберем ширину окна h так, что $nh^5 \to 0$. Предложить оценку $\hat{H}(t)$ для $H(t)$, использующую $\hat{F}(t)$ и $\hat{f}(t)$, и найти ее асимптотическое распределение.

10 Решения задач

10.1 Оценка плотности Надарайа–Уотсона

Применим знакомую технологию:

$$\begin{aligned}
\mathbb{E}[\hat{f}(a) - f(a)] &= \int K(u) \left(f(a) + huf'(a) + \frac{1}{2}(hu)^2 f''(a) + O(h^3) \right) du - f(a) \\
&= \frac{h^2}{2} f''(a) \sigma_K^2 + O(h^3)
\end{aligned}$$

и

$$\begin{aligned}
\mathbb{V}[\hat{f}\,(a)] &= \frac{1}{nh} \int K\,(u)^2\,(f\,(a) + o\,(1))\,du - \frac{1}{n} O\,(1) \\
&= \frac{1}{nh} f(a) R_K + o\left(\frac{1}{nh}\right).
\end{aligned}$$

Используя ЦПТ Линдберга–Леви, по мере того как $n \to \infty$, $h \to 0$ и $nh \to \infty$, имеем

$$\sqrt{nh}\left(\hat{f}(a) - f(a)\right) \xrightarrow{d} \mathcal{N}\left(\frac{1}{2}\lambda f''(a)\sigma_K^2, f(a)R_K\right),$$

при условии, что $\lambda \equiv \lim\limits_{n\to\infty} \sqrt{nh^5}$ существует и конечно.

Асимптотическое смещение пропорционально $f''(a)$, значение которой говорит о том, насколько плотность в окрестности a отличается от оцениваемой плотности в a. Отметим, что асимптотическое смещение не зависит от $f(a)$, т.е. как часто наблюдения попадают в данную область, и от $f'(a)$, т.е. отличаются ли плотности слева и справа от a. Асимптотическая дисперсия пропорциональна $f(a)$, плотности в a, что может показаться странным (бо́льшая частота выпадения наблюдений приводит к худшему качеству оценивания). Однако мы оцениваем $f(a)$, так что ее большее значение также означает больший разброс оценки вокруг истинной величины, и этот эффект превалирует (эффект частоты дает $\propto f(a)^{-1}$, эффект размера дает $\propto f(a)^2$).

10.2 Несмещенность ядерных оценок

Математическое ожидание равно

$$\begin{aligned}
\mathbb{E}\left[\hat{g}\left(a\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n y_i K_h\left(x_i - a\right)}{\sum_{i=1}^n K_h\left(x_i - a\right)}|x_1, \ldots, x_n\right]\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}\left[y_i|x_i\right] K_h\left(x_i - a\right)}{\sum_{i=1}^n K_h\left(x_i - a\right)}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n cK_h\left(x_i - a\right)}{\sum_{i=1}^n K_h\left(x_i - a\right)}\right] = c,
\end{aligned}$$

т.е. оценка $\hat{g}\left(a\right)$ несмещена для $c = g\left(a\right)$. Причина проста: все элементы выборки одинаково релевантны при оценивании тривиального условного среднего, так что от участия точек вдали от a смещение не возникает.

Оценка локальной линейной регрессии будет несмещена, если $g\left(x\right) = c + bx$. Тогда все элементы выборки одинаково релевантны при оценивании, так как оценивается линейная, хоть и локальная, регрессия. Действительно,

$$\hat{g}_1\left(a\right) = \bar{y} + \frac{\sum_{i=1}^n \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) K_h\left(x_i - a\right)}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2 K_h\left(x_i - a\right)}\left(a - \bar{x}\right),$$

так что

$$\begin{aligned}
\mathbb{E}\left[\hat{g}_1\left(a\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\bar{y}|x_1, \ldots, x_n\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) K_h\left(x_i - a\right)}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2 K_h\left(x_i - a\right)}|x_1, \ldots, x_n\right]\left(a - \bar{x}\right)\right] \\
&= \mathbb{E}\left[c + b\bar{x}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n \left(c + bx_i - c - b\bar{x}\right)\left(x_i - \bar{x}\right) K_h\left(x_i - a\right)}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2 K_h\left(x_i - a\right)}|x_1, \ldots, x_n\right]\left(a - \bar{x}\right)\right] \\
&= \mathbb{E}\left[c + b\bar{x} + b\left(a - \bar{x}\right)\right] = c + bx.
\end{aligned}$$

Что же касается плотности, вряд ли стоит ожидать несмещенности. Действительно,

$$\hat{f}\left(x\right) = \frac{1}{n}\sum_{i=1}^n K_h\left(x_i - a\right),$$

так что

$$\mathbb{E}[\hat{f}\left(a\right)] = \mathbb{E}\left[K_h\left(x_i - a\right)\right] = \frac{1}{h}\int K\left(\frac{x_i - a}{h}\right) f\left(x\right) dx.$$

Это матожидание сильно зависит от ширины окна и ядра, и вряд ли будет равно $f\left(x\right)$, кроме как в особых условиях (например, равномерная $f\left(x\right)$, a далеко от границ и т.д.).

10.3 Оценивание при ограничении на форму

Технология с постоянным эффектом от масштаба обладает свойством

$$f(l,k) = kf\left(\frac{l}{k},1\right).$$

Регрессия для нормированных переменных выглядит как

$$\frac{y_i}{k_i} = f\left(\frac{l_i}{k_i},1\right) + \frac{\varepsilon_i}{k_i}.$$

Поэтому можно построить (одномерную!) ядерную оценку для $f(l,k)$ как

$$\hat{f}(l,k) = k \times \frac{\sum_{i=1}^{n} \frac{y_i}{k_i} K_h\left(\frac{l_i}{k_i} - \frac{l}{k}\right)}{\sum_{i=1}^{n} K_h\left(\frac{l_i}{k_i} - \frac{l}{k}\right)}.$$

По сути, мы присваиваем больший вес тем наблюдениям, которые ближе к лучу l/k.

10.4 Непараметрическая функция риска

Простая непараметрическая оценка для $F(t) \equiv \Pr\{z \leq t\}$ – это выборочная частотность

$$\hat{F}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{z_j \leq t\}.$$

Согласно ЗБЧ, она состоятельна для $F(t)$. Согласно ЦПТ, ее скорость сходимости равна $\sqrt{n}$.

Мы знаем из разделов 9.1 и 10.1, что

$$\sqrt{nh}\left(\hat{f}(t) - f(t)\right) \xrightarrow{d} \mathcal{N}(0, R_K f(t)),$$

используя $\lambda = 0$. По принципу аналогий,

$$\hat{H}(t) = \frac{\hat{f}(t)}{1-\hat{F}(t)}.$$

По теореме Слуцкого, эта оценка состоятельна для $H(t)$, а также

$$\begin{aligned}\sqrt{nh}\left(\hat{H}(t) - H(t)\right) &= \frac{\sqrt{nh}(\hat{f}(t)-f(t))}{1-\hat{F}(t)} + \sqrt{h}f(t)\frac{\sqrt{n}(\hat{F}(t)-F(t))}{(1-\hat{F}(t))(1-F(t))} \\ &\xrightarrow{d} \frac{1}{1-F(t)}\mathcal{N}(0, R_K f(t)) + 0 \sim \mathcal{N}\left(0, R_K\frac{f(t)}{(1-F(t))^2}\right).\end{aligned}$$

Причина того, что неопределенность в $\hat{F}(t)$ не влияет на асимптотическое распределение $\hat{H}(t)$, в том, что $\hat{F}(t)$ сходится быстрее, чем $\hat{f}(t)$.

К сожалению, $\hat{H}(t)$ выглядит не очень аппетитно из-за скачков в $\hat{F}(t)$.

Список литературы

Крил, М. (2008). Некоторые ловушки параметрической инференции. *Квантиль* 4, 1–6.

Кристенсен, Д. (2009). Полупараметрическая эконометрика: вводный курс. *Квантиль* 7, 53–83.

Расин, Дж. (2008). Непараметрическая эконометрика: вводный курс. *Квантиль* 4, 7–56.

Bierens, H.J. (1994). *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. New York: Cambridge University Press.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Глава 76 в *Handbook of Econometrics* (под редакцией J.J. Heckman & E.E. Leamer), том 6/2. Elsevier Science.

Franses, P. & D. van Dijk (2000). *Nonlinear Time Series Models in Empirical Finance*. New York: Cambridge University Press.

Härdle, W. (1990). *Applied Nonparametric Regression*. New York: Cambridge University Press.

Härdle, W. & O. Linton (1994). Applied nonparametric methods. Глава 38 в *Handbook of Econometrics* (под редакцией R. Engle & D. McFadden), том 4. Elsevier Science.

Heiler, S. (2001). Nonparametric time series analysis: nonparametric regression, locally weighted regression, autoregression, and quantile regression. Глава в *A Course in Time Series Analysis* под редакцией D. Peña, G. Tiao & R. Tsay. Wiley.

Li, Q. & J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Nadaraya, E.A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* 10, 186–190.

Pagan, A. & A. Ullah (1999). *Nonparametric Econometrics*. New York: Cambridge University Press.

Tschernig, R. & L. Yang (2000). Nonparametric lag selection for time series. *Journal of Time Series Analysis* 21, 457–487.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya* 26, 359–372.

Nonparametric regression

Stanislav Anatolyev

New Economic School, Moscow, Russia

This essay covers the principles and methodology of nonparametric estimation of a mean regression. The emphasis is put on kernel smoothing, but non-kernel methods are also reviewed.

Полупараметрическое моделирование и оценивание[*]

Деннис Кристенсен[†]

Колумбийский Университет, Нью-Йорк, США
Центр эконометрического анализа временных рядов, Орхус, Дания

Полупараметрические модели характеризуются тем, что включают в себя конечномерные и бесконечномерные (функциональные) параметры. Из-за этого они обладают дополнительной гибкостью по сравнению с параметрическими моделями, и в то же время можно строить оценки для параметрических компонент, сходящиеся со скоростью, обычной для параметрических оценок. Эти две особенности делают полупараметрические модели и оценки все более и более популярными в прикладной экономике. В настоящем эссе содержится выборочный обзор литературы по полупараметрическому моделированию и оцениванию, с уклоном в полупараметрические регрессионные модели. Особое внимание уделяется построению двухшаговых полупараметрических оценок и выводу их асимптотических свойств. Также кратко обсуждаются оценивание «решетом» и полупараметрическая эффективность.

1 Введение

В течение последних тридцати лет полупараметрическое моделирование и оценивание экономических процессов привлекает много внимания. Главная причина такой популярности в том, что подобная модель является компромиссным решением между двумя крайностями, полностью параметрическим и полностью непараметрическим моделированием. В первом случае для объяснения выборочных данных используются параметрическая модель, и естественная оценка в этом случае – оценка максимального правдоподобия (ММП-оценка). При правильной спецификации ММП-оценка обладает обычными приятными свойствами, такими как максимальная эффективность. Но если некоторые компоненты модели определены неверно, ММП-оценка асимптотически смещена, и выводы, полученные на основе оцененной модели, могут быть крайне обманчивыми. В противоположность этому полностью непараметрические модели дают максимальную гибкость, сводя к минимуму вероятность неправильно специфицировать модель. С другой стороны, непараметрическое оценивание требует много исходных данных, и в малых выборках мы получим довольно неточные оценки. Особенно это проявляется в моделях большой размерности, где точность оценок падает по мере добавления новых переменных. Это явление называется «проклятие размерности».

Полупараметрические модели являются компромиссным решением между непараметрическим и параметрическим подходами. Они включают в себя как непараметрические, так и параметрические компоненты. Поэтому полупараметрическая модель сохраняет до некоторой степени гибкость полностью непараметрической модели и гораздо менее подвержена неправильной спецификации по сравнению с полностью параметрической моделью. В то же время параметрическую компоненту полупараметрической модели вообще можно оценить с точностью, сравнимой с той, которую мы бы получили, используя (правильно специфицированную) полностью параметрическую модель.

[*]Перевод О. Еремина и С. Анатольева. Цитировать как: Кристенсен, Деннис (2009). «Полупараметрическое моделирование и оценивание», Квантиль, №7, стр. 53–83. Citation: Kristensen, Dennis (2009). "Semiparametric modelling and estimation," Quantile, No.7, pp. 53–83.

[†]Адрес: Economics Department, Columbia University, 1018 International Affairs Building, 420 West 118th Street, New York, NY 10027, USA. Электронная почта: *dk2313@columbia.edu*

В настоящем эссе мы даем краткое введение и обзор методов по полупараметрическому моделированию и оцениванию, уделяя особое внимание регрессионным моделям. Мы вводим основные понятия полупараметрического моделирования и оценивания в рамках регрессионных моделей по трем причинам. Во-первых, эти модели широко используются в экономике и поэтому должны быть хорошо знакомы рядовому читателю. Во-вторых, с регрессионными моделями довольно просто работать, что позволяет легко ввести основные полупараметрические понятия и методики. В-третьих, большинство способов, которые мы рассматриваем в рамках регрессий, можно перенести на многие другие виды моделей. Чтобы продемонстрировать это, мы кратко коснемся полупараметрических копул и покажем, как представленные методы можно применить в этих условиях.

Сперва мы рассмотрим оценивание некоторых ведущих полупараметрических моделей. Затем мы определим основной подход, с помощью которого можно проанализировать асимптотические свойства этих оценок. Основной класс оценок, который мы будем рассматривать, – это так называемые двухшаговые полупараметрические оценки, в которых на первом шаге оцениваются непараметрические компоненты модели, которые затем используются для оценки параметрических компонент. Мы выведем ряд условий общего типа, при которых полупараметрическая оценка состоятельна и асимптотически нормально распределена, и далее обсудим подробнее, как эти условия можно проверить для конкретной модели.

В качестве альтернативной стратегии оценивания мы кратко рассмотрим класс полупараметрических оценок, основанных на так называемом методе «решето». В то же время мы не детализируем основную теорию этих оценок. В заключение мы немного обсудим полупараметрическую эффективность и как она используется для для создания новых оценок. Опять же, эта часть статьи не требует специальных знаний, мы лишь попробуем объяснить различные концепции на интуитивном уровне.

Мы не ставим перед собой цель охватить в данном обзоре всю литературу по данной теме. Следует иметь ввиду, что есть много замечательных обзоров литературы по полупараметрическому моделированию и оцениванию. Среди них Ichimura & Todd (2007), Härdle, Müller, Sperlich & Werwatz (2004), Horowitz (2009), Li & Racine (2007), Pagan & Ullah (1999), Powell (1994) и Robinson (1988), дополняющие и расширяющие наш обзор по ряду направлений.

Эссе организовано следующих образом. В разделах 2–4 мы рассмотрим несколько примеров полупараметрических моделей и обсудим, как осуществляется их оценивание. В разделе 5 мы проанализируем свойства основного класса двухшаговых полупараметрических оценок, который включает некоторые из оценок, рассмотренных в предыдущем разделе. Мы сосредоточимся на оценках, основанных на ядерном сглаживании, так как они просты для анализа и популярны в прикладных исследованиях. В разделе 6 мы кратко рассмотрим метод одновременного оценивания обоих компонент, используя, так называемый, метод «решето» для непараметрической компоненты. Полупараметрическую эффективность мы обсудим в разделе 7. В заключение в разделе 8 мы укажем список работ, где более детально рассмотрены упомянутые темы.

Разделы 2–4 и 6–7 не требуют глубоких знаний эконометрической теории, но раздел 5 потребует некоторых усилий от менее подготовленного в техническом плане читателя. Чтобы сохранить сложность изложения на приемлемом уровне, мы не приводим доказательств, и большинство математических выводов даны в общих чертах. Заинтересовавшийся читатель может обратиться к автору за доказательством всех теорем. Кроме того, в разделе 8 приведены статьи, содержащие более точные результаты и строгие доказательства.

2 Полупараметрические регрессии

В самой общей форме регрессионную модель можно представить в виде

$$Y = m(X) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] = 0, \tag{1}$$

где $Y \in \mathbb{R}$ – отклик (зависимая переменная), $X \in \mathbb{R}^d$ – набор из $d \geq 1$ регрессоров (независимых переменных), и $\varepsilon \in \mathbb{R}$ – ошибка. Регрессионная функция $m : \mathbb{R}^d \mapsto \mathbb{R}$ объясняет, как условное математическое ожидание Y меняется с X:

$$\mathbb{E}\left[Y|X=x\right] = m\left(x\right).$$

Также, пусть $f_{\varepsilon|X}\left(e|x\right)$ – условная плотность распределения ε при заданном $X = x$.[1] Предположим, у нас имеется случайная выборка (Y_i, X_i) из модели, где $i = 1, ..., n$. Нам интересно осуществить инференцию относительно функций $m\left(x\right)$ и $f_{\varepsilon|X}\left(e|x\right)$.

В полностью параметрическом случае мы предполагаем, что как регрессионная функция, m, так и (условная) функция распределения ошибок, $f_{\varepsilon|X}$, известны вплоть до некоторого параметра конечной размерности. То есть, у нас есть конкретные параметрические функции $m\left(x;\beta\right)$ и $f_{\varepsilon|X}\left(e|x;\sigma\right)$, где $\beta \in \mathcal{B}$ включает коэффициенты регрессии, характеризующие вид функции m, а $\sigma \in \Sigma$ – параметр, задающий вид (условной) функции распределения ошибок. Предполагая верную спецификацию модели, то есть $m\left(x\right) = m\left(x;\beta_0\right)$ и $f_{\varepsilon|X}\left(e|x\right) = f_{\varepsilon|X}\left(e|x;\sigma_0\right)$ для некоторого $\theta_0 = (\beta_0, \sigma_0)$, естественной оценкой модели будет ММП-оценка

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f_{\varepsilon|X}\left(Y_i - m\left(X_i;\beta\right)|X_i;\sigma\right).$$

Часто используют гауссовскую регрессионную модель, где вектор ошибок не зависит от X и распределен нормально: $N\left(0,\sigma^2\right)$. В этом случае ММП-оценка $\theta = \left(\beta, \sigma^2\right)$ совпадает с оценкой наименьших квадратов: $\hat{\beta}_{\mathrm{MLE}} = \hat{\beta}_{\mathrm{LS}}$ и $\hat{\sigma}^2_{\mathrm{MLE}} = \hat{\sigma}^2_{\mathrm{LS}}$ где

$$\hat{\beta}_{\mathrm{LS}} = \arg\min_{\beta\in\mathcal{B}} \sum_{i=1}^{n} \left(Y_i - m\left(X_i;\beta\right)\right)^2, \quad \hat{\sigma}^2_{\mathrm{LS}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - m(X_i;\hat{\beta})\right)^2.$$

Что касается спецификации регрессионной функции, широко используется линейная функция: $m\left(x;\beta\right) = \beta_1 x_1 + ... + \beta_d x_d$, и ММП-оценка превращается в оценку метода наименьших квадратов (МНК),

$$\hat{\beta}_{\mathrm{OLS}} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right). \tag{2}$$

При выполнении условий регулярности оценка $\hat{\beta}_{\mathrm{MLE}}$ $\sqrt{n}$-состоятельна и асимптотически нормально распределена. Например, при гауссовых ошибках ММП-оценка удовлетворяет

$$\sqrt{n}(\hat{\beta}_{\mathrm{MLE}} - \beta_0) \to^d N\left(0, V\right), \quad V = \sigma^2 \mathbb{E}\left[\dot{m}\left(x;\beta\right)\dot{m}\left(x;\beta\right)'\right]^{-1},$$

где $\dot{m}\left(x;\beta\right) = \partial m\left(x;\beta\right)/\partial\beta$ (см., например, Amemiya, 1985). В свою очередь, это значит, что регрессионную функцию можно оценить как $\hat{m}_{\mathrm{MLE}}\left(x\right) = m(x;\hat{\beta}_{\mathrm{MLE}})$.

Однако параметрическая модель может быть неверно специфицирована, то есть $m\left(x;\beta\right) \neq m\left(x\right)$ для всех $\beta \in \mathcal{B}$ и/или $f_{\varepsilon|X}\left(e|x;\sigma\right) \neq f_{\varepsilon|X}\left(e|x\right)$ для всех значений $\sigma \in \Sigma$. В этом случае оценка регрессионной функции $\hat{m}_{\mathrm{MLE}}\left(x\right)$, вообще говоря, несостоятельна и ведет к неправильному представлению о том, как X влияет на Y. Чтобы избежать этого, можно использовать полностью параметрические оценки m, например, ядерные оценки или оценки «решетом». Мы сосредоточимся на ядерных оценках и вкратце дадим общее представление о решетчатых оценках. Подробнее смотрите статьи Härdle (1992), Silverman (1986), Расин (2008) и Анатольев (2009). Оценки «решетом» мы кратко обсудим в разделе 6. Ядерные оценки

[1] Мы предполагаем, как это обычно принято в литературе по не- и полупараметрическому моделированию, что все переменные имеют непрерывное распределение.

образует особый класс непараметрических оценок, которые используют локальные данные для инференции о характеристиках распределения. Предположим, X непрерывно распределена с плотностью распределения $f(x)$. Тогда плотность можно оценить непараметрически с помощью ядерной оценки плотности: для каждого значения $x \in \mathbb{R}^d$ она вычисляется по формуле

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h (X_i - x), \tag{3}$$

где $K_h(x) = K(x/h)/h^d$, $K : \mathbb{R}^d \mapsto \mathbb{R}$ – ядерная функция, а $h > 0$ – ширина окна. Исследователь сам выбирает K и h. Ядерная оценка плотности похожа на гистограмму плотности, где ширина окна задает ширину ячейки гистограммы, а ядро – вес наблюдения для каждой ячейки. Наибольший вес имеют наблюдения, близкие к x, а расположенные далеко от x играют малую роль или вовсе не учитываются. Аналогичным образом ядерная оценка регрессии $m(x) = \mathbb{E}[Y|X = x]$ в данной точке $x \in \mathbb{R}^d$ принимает форму взвешенного среднего:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K_h (X_i - x)}{\sum_{i=1}^{n} K_h (X_i - x)}. \tag{4}$$

Повторим, что это локальная оценка, которая использует наблюдения X_i, близкие к x, для извлечения информации о значении $m(\cdot)$ в точке x.

Ядерные оценки регрессии очень робастны. Оценка $\hat{m}(x)$ состоятельна при $h \to 0$ и $nh^d \to \infty$ независимо от истинного значения регрессионной функции m. Но при этом страдает точность оценивания: в конечной выборке дисперсия оценки выше по сравнению с дисперсией параметрических оценок. На теоретическом уровне это следует из того, что скорость сходимости ядерной оценки $\sqrt{n^{4/(4+d)}}$ ниже, чем $\sqrt{n}$ – скорость сходимости параметрических оценок при $h \to 0$.[2] Заметим, что точность непараметрических оценок зависит от размерности X, $d \geq 1$: с ростом d снижается скорость сходимости непараметрической оценки (упоминавшееся ранее проклятье размерности). Вдобавок к этому, даже если удастся получить точные непараметрические оценки, ядерные оценки $\hat{f}(x)$ и $\hat{m}(x)$ сложно представить и интерпретировать при большом d.

Таким образом, выбор между разными моделями и способами оценивания – это выбор между риском неправильной спецификации модели и степенью точности оценок. ММП-оценка полностью параметрической модели обладает наибольшей точностью, но очень велика вероятность получить неверные выводы из-за неправильной спецификации. Напротив, полностью непараметрическая модель гарантирует правильную спецификацию, но дает очень низкую точность оценки. Это приводит нас к использованию полупараметрических моделей и оценок, которые обладают относительно высокой гибкостью и в то же время лучшей скоростью сходимости для некоторых параметров модели.

2.1 Одноиндексная модель

Широко распространенная полупараметрическая регрессионная модель, так называемая одноиндексная, имеет следующий вид:

$$Y = g(\beta' X) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] = 0, \tag{5}$$

где функция $g : \mathbb{R} \mapsto \mathbb{R}$ и параметр $\beta \in \mathbb{R}^d$ неизвестны. Мы не делаем никаких предположений относительно (условного) распределения ε, и считаем g и $f_{\varepsilon|X}$ непараметрическими функциями. В этом случае $\gamma = (g, f_{\varepsilon|X})$ – бесконечномерная (непараметрическая) компонента, а β – параметрическая компонента.

[2] Мы предполагаем, что интересующая нас функция дважды непрерывно дифференцируема.

Название «одноиндексная» происходит от того, что g является функцией *индекса* $\beta' X \in \mathbb{R}$, а не всего вектора $X \in \mathbb{R}^d$. Таким образом, мы полагаем, что X влияет на Y только посредством индекса $\beta' X$, что является ограничением относительно полной регрессионной функции m из уравнения (1). Поэтому по сравнению с полностью непараметрической моделью есть риск неправильной спецификации модели.

С другой стороны, легко интерпретировать влияние X на Y, которое описывается конечномерным параметром β и одномерной функцией g. При этом g имеет область определения $\mathbb{R}$, в то время как область определения функции m из уравнения (1) – $\mathbb{R}^d$. Поэтому, независимо от размерности X, оценивание g остается одномерной задачей, что по существу избавляет нас от проклятья размерности.

Данный подход применим для некоторых видов трансформационных моделей. Предположим, случайная величина Y^* удовлетворяет

$$Y^* = \beta_0' X + \eta,$$

где $\eta \sim F_\eta$ не зависит от X. Однако мы наблюдаем не Y^*, а только

$$Y = t\,(Y^*)$$

для некоторого преобразования t, которое может быть как известным, так и неизвестным. Видно, что

$$\mathbb{E}\,[Y|X = x] = \mathbb{E}\left[t\left(\beta_0' X + \eta\right)|X = x\right] = \int t\left(\beta_0' x + v\right) dF_\eta\,(v)\,.$$

Определив g и ε как

$$g\,(z) := \int t\,(z + v)\,dF_\eta\,(v)\,, \quad \varepsilon := Y - \mathbb{E}\,[Y|X = x]\,,$$

этот класс моделей можно представить в форме уравнения (5). Трансформационные модели включают в себя модели с ограниченной зависимой переменной, такие как цензурированные регрессионные модели и модели длительности. Например, при $t\,(y) = 1\,\{y > 0\}$ трансформационная модель превращается в модель бинарного выбора, такую как логит и пробит. Если преобразование $t\,(y) = y \cdot 1\,\{y > 0\}$, мы получим Тобит-модель. Преимущество полупараметрического подхода в том, что мы можем оценить β без необходимости настаивать на точной форме t и F_η.

Теперь мы хотим построить оценку одноиндексной модели. С этой целью сперва нам надо обсудить вопрос идентификации интересующих нас параметров. То есть можем ли мы однозначно определить параметры β и функции g и $f_{\varepsilon|X}$ из данных? Заметим, что параметр β нельзя идентифицировать, если $\Pr\left\{\delta' X = c\right\} = 1$ для некоторых констант $c \in \mathbb{R}$ и $\delta \in \mathbb{R}^d$. Более того, нам нужно нормировать β, чтобы идентифицировать g. Чтобы понять, почему это так, положим $\tilde{g}\,(z) = g\,(a + bz)$ для некоторых констант $(a, b) \in \mathbb{R}^2$, что эквивалентно $\tilde{g}\,(-a + 1/bz) = g\,(z)$. Обе спецификации неразличимы:

$$g\left(\beta' x\right) = \tilde{g}\left(-a + (1/b)\beta' x\right) = \tilde{g}(\tilde{\beta}' x),$$

где $\tilde{\beta} = (1/b)\beta$, то есть по исходной выборке мы не сможем различить $\tilde{g}$ и g. Поэтому потребуем, чтобы X не содержал констант, а один из коэффициентов β был равен единице; положим $\beta_1 = 1$ (всегда можно переобозначить порядок компонент X). Наконец, заметим, что если g – линейная функция, то мы не можем идентифицировать β (если g неизвестна).

При сделанных предположениях мы теперь можем вывести оценки для β и g. Сперва предположим, что функция g известна. Тогда естественной оценкой β будет МНК-оценка

$$\hat{\beta}_g = \arg\min_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - g\left(\beta' X_i\right)\right]^2. \tag{6}$$

И наоборот, предположим, что $\beta \in \mathbb{R}^d$ известна, тогда естественной непараметрической оценкой g будет стандартная ядерная регрессионная оценка

$$\hat{g}(z;\beta) = \frac{\sum_{i=1}^n Y_i K_h\left(\beta' X_i - z\right)}{\sum_{i=1}^n K_h\left(\beta' X_i - z\right)}.$$

Однако, если β и g неизвестны, обе эти оценки будут недостижимыми. Вместо этого мы предлагаем использовать комбинированный подход: подставить непараметрическую оценку $\hat{g}(z;\beta)$ в функцию наименьших квадратов уравнения (6) и получить достижимую оценку β:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^n \left[Y_i - \hat{g}\left(\beta' X_i;\beta\right)\right]^2.$$

Когда мы получили $\hat{\beta}$, очевидно, оценкой $g(z)$ будет $\hat{g}(z;\hat{\beta})$.

Альтернативный способ – метод оценивания средних производных, предложенный Powell, Stock & Stoker (1989). Пусть g дифференцируемая функция, тогда выполняется следующее равенство:

$$\frac{\partial \mathbb{E}\left[Y|X=x\right]}{\partial x} = \beta g'\left(\beta' x\right),$$

где $g'(x) = \partial g(z)/\partial z$. Следовательно, для любой ограниченной функции w

$$\mathbb{E}\left[\frac{\partial \mathbb{E}\left[Y|X\right]}{\partial x} w(X)\right] = \beta \mathbb{E}\left[w(X)\, g'\left(\theta' X\right)\right].$$

Это доказывает, что параметр δ, заданный формулой

$$\delta := \mathbb{E}\left[\frac{\partial \mathbb{E}\left[Y|X\right]}{\partial x} w(X)\right]$$

неотличим с точностью до коэффициента нормирования $\left(\mathbb{E}\left[w(X)\, g'\left(\theta' X\right)\right]\right)$ от β. Теперь построим оценку δ с использованием весовой функции w, такой что $w(x) = f(x)$, где f – плотность распределения X. Сперва рассмотрим

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \mathbb{E}\left[Y|X\right]}{\partial x} f(X)\right] &= \int_{\mathbb{R}^d} \frac{\partial \mathbb{E}\left[Y|X=x\right]}{\partial x} f^2(x)\, dx \\
&= -2\int_{\mathbb{R}^d} \mathbb{E}\left[Y|X=x\right] f(x) \frac{\partial f(x)}{\partial x} dx \\
&= -2\mathbb{E}\left[\mathbb{E}\left[Y|X\right] \frac{\partial f(X)}{\partial x}\right] \\
&= -2\mathbb{E}\left[Y \frac{\partial f(X)}{\partial x}\right].
\end{aligned}$$

Последнее выражение в правой части является основой для оценки δ: заменив популяционное математическое ожидание выборочным, а плотность f ядерной оценкой $\hat{f}$ согласно равенству (3), получим:

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^n Y_i \frac{\partial \hat{f}(X_i)}{\partial x}.$$

Преимущество $\hat{\delta}$ над $\hat{\beta}$ в том, что первая оценка имеет явный вид и не требует численной оптимизации.

Можно распространить одноиндексную модель на более общий класс моделей:

$$Y = g\left(v\left(X;\beta_0\right)\right) + \varepsilon, \quad \mathbb{E}\left[\varepsilon|X\right] = 0,$$

для некоторой функции $v : \mathbb{R} \times \mathcal{B} \mapsto \mathbb{R}$, известной с точностью до β_0. Методика оценивания, изложенная выше в общих чертах, применима и для этого более общего случая.

2.2 Частично линейная модель

Если предположить, что m из (1) линейна по некоторым своим аргументам, то можно получить другую спецификацию. Положим $X = (X_1, X_2)$, где $X_i \in \mathbb{R}^{d_i}$, $i = 1, 2$, и $d = d_1 + d_2$, так что

$$Y = \beta_0' X_1 + g(X_2) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] = 0, \tag{7}$$

для некоторых $g : \mathbb{R}^{d_2} \mapsto \mathbb{R}$ и $\beta \in \mathbb{R}^{d_1}$. Как и раньше, мы не специфицируем $\varepsilon|X$.

По сравнению с общей регрессионной моделью из уравнения (1), мы налагаем следующее ограничение на вид регрессионной функции: $m(x) = \beta_0' x_1 + g(x_2)$. То есть Y аддитивна по X_1 и X_2, причем Y зависит от X_1 линейно. Наша модель включает параметрическую компоненту β_0 и две непараметрические компонентаы g и $f_{\varepsilon|X}$, то есть является полупараметрической моделью.

Опять же, мы должны наложить ограничения на модель, чтобы g и β были идентифицируемы. Среди компонент X не должно быть константы, так как при $\tilde{g}(x_2) = g(x_2) - a$, $a \in \mathbb{R}$, мы не сможем различить $\beta' x_1 + g(x_2)$ и $\left(a + \beta' x_1\right) + \tilde{g}(x_2)$. В действительности мы должны предположить, что матрица

$$\Omega = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1|X_2])(X_1 - \mathbb{E}[X_1|X_2])'\right]$$

невырождена. Если это условие не выполняется, мы не сможем различать линейный и нелинейный члены. Чтобы показать это, рассмотрим

$$\mathbb{E}[Y|X_2] = \beta_0' \mathbb{E}[X_1|X_2] + g(X_2) + \mathbb{E}[\varepsilon|X_2],$$

что означает

$$Y - \mathbb{E}[Y|X_2] = \beta_0'(X_1 - \mathbb{E}[X_1|X_2]) + \eta, \tag{8}$$

где $\eta = \varepsilon - \mathbb{E}[\varepsilon|X_2]$ удовлетворяет условию $\mathbb{E}[\eta|X] = 0$. Таким образом, чтобы идентифицировать β_0, нужно, чтобы Ω была невырождена.

Уравнение (8) является базовым для оценки, основанной на остатках: строим ядерные оценки для $m_Y(x_2) = \mathbb{E}[Y|X_2 = x_2]$ и $m_{X_1}(x_2) = \mathbb{E}[X_1|X_2 = x_2]$,

$$\hat{m}_Y(x_2) = \frac{\sum_{i=1}^n Y_i K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)}, \quad \hat{m}_{X_1}(x_2) = \frac{\sum_{i=1}^n X_{1,i} K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)},$$

и подставляем их в (8). После этого мы можем оценить β с помощью МНК,

$$\hat{\beta} = \left(\sum_{i=1}^n \hat{Z}_i \hat{Z}_i'\right)^{-1} \sum_{i=1}^n \hat{Z}_i (Y_i - \hat{m}_Y(X_{2,i}))', \tag{9}$$

где $\hat{Z}_i = X_{1,i} - \hat{m}_{X_1}(X_{2,i})$.

Этот метод оценивания можно расширить на следующую более общую модель:

$$Y = v(X_1; \beta) + g(X_2) + \varepsilon, \quad E[\varepsilon|X] = 0, \tag{10}$$

где $v : \mathbb{R}^{d_1} \times \mathcal{B} \mapsto \mathbb{R}$ известна с точностью до $\theta \in \Theta$. Однако полученная оценка имеет неявный вид, и приходится применять численные методы оптимизации.

3 Спецификация распределения ошибок

До этого момента мы обсуждали только как можно смоделировать и оценить функциональный вид m в общей регрессионной модели с помощью полупараметрических методов. В этом разделе мы сосредоточимся на ошибке ε и обсудим, как разные предположения относительно ошибки приводят к разным (полупараметрическим) методикам оценивания регрессионной функции. В некоторых случаях можно вывести оценку интересного нам параметра, не оценивая бесконечномерные компоненты. Однако эти оценки неэффективны, и для улучшения их эффективности можно использовать методы полупараметрического оценивания.

3.1 Линейная регрессионная модель

Рассмотрим стандартную линейную регрессионную модель:

$$Y = \beta' X + \varepsilon, \tag{11}$$

где $\mathbb{E}\left[\varepsilon | X\right] = 0$. Обычно она считается полностью параметрической моделью, но в нашей терминологии это полупараметрическая модель, так как распределения $\varepsilon|X$, $f_{\varepsilon|X}$, определены не полностью. Если $f_{\varepsilon|X}$ не задана, мы имеем параметрическую компоненту, θ, и непараметрическую компоненту, $f_{\varepsilon|X}$.

Если предположить, что ошибка имеет нормальное распределение, то, как было показано в предыдущем разделе, ММП-оценка превращается в обычную МНК-оценку как в уравнении (2). Однако МНК-оценку можно рассматривать как полупараметрическую оценку θ, так как она остается $\sqrt{n}$-состоятельной независимо от точности спецификации $f_{\varepsilon|X}$. Более того, привлекательное свойство МНК состоит в том, что для его применения не нужно оценивать $f_{\varepsilon|X}$ в отличие от полупараметрических оценок, рассмотренных в предыдущем разделе, где нужно было получить предварительную оценку непараметрической компоненты, чтобы затем оценить параметрический компонент.

Однако читателю может быть интересно, существуют ли другие оценки, получше. Очевидно, если мы наложим (правильную) параметрическую структуру на $f_{\varepsilon|X}$, мы можем использовать ММП, который вообще более эффективен, чем МНК. Но даже если мы не налагаем никакой формы на распределение, то, как мы увидим далее, МНК-оценка в общем случае неэффективна в классе полупараметрических оценок.

3.2 Гетероскедастичность неизвестной формы

Мы рассматриваем линейную модель (11), но теперь предполагаем, что ошибки гетероскедастичны,

$$\mathbb{E}\left[\varepsilon^2 | X = x\right] = \sigma^2\left(x\right), \tag{12}$$

где условная функция дисперсии, $\sigma^2\left(\cdot\right)$, *неизвестна.*

Обычная МНК-оценка (2) по-прежнему состоятельна и асимптотически нормально распределена, но теперь ее асимптотическое распределение следующее:

$$\sqrt{n}(\hat{\theta}_{\mathrm{OLS}} - \theta) \to^d N\left(0, \mathbb{E}\left[XX'\right]^{-1} \mathbb{E}\left[\sigma^2\left(X\right)XX'\right]^{-1} \mathbb{E}\left[XX'\right]^{-1}\right).$$

В частности, она уже не эффективна, как мы далее увидим. Сперва рассмотрим случай, когда условная дисперсионная функция $\sigma^2\left(x\right)$ *известна.* Тогда мы можем применить взвешенный метод наименьших квадратов (ВМНК),

$$\tilde{\theta}_{\mathrm{WLS}} = \left(\sum_{i=1}^{n} \sigma^{-2}\left(X_i\right) X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \sigma^{-2}\left(X_i\right) X_i Y_i\right), \tag{13}$$

который дает меньшую асимптотическую дисперсию оценки по сравнению с МНК:

$$\sqrt{n}(\tilde{\theta}_{\mathrm{WLS}} - \theta) \to^d N\left(0, \mathbb{E}\left[\sigma^{-2}\left(X\right)XX'\right]^{-1}\right),$$

где

$$\mathbb{E}\left[\sigma^{-2}\left(X\right)XX'\right]^{-1} \leq \mathbb{E}\left[XX'\right]^{-1} \mathbb{E}\left[\sigma^2\left(X\right)XX'\right]^{-1} \mathbb{E}\left[XX'\right]^{-1}$$

которое превращается в равенство тогда и только тогда, когда $\sigma^2\left(X\right) = \sigma^2 = \mathbb{E}\left[\varepsilon^2\right]$ константа почти наверное.

Если функция условной дисперсии $\sigma^2(x)$ *неизвестна*, то $\tilde{\theta}_{\mathrm{WLS}}$ недостижима. В этом случае можно положить, что $\sigma^2(x)$ зависит от параметров, и оценить эти параметры, используя стандартные методы. Однако в этом случае нужно правильно специфицировать функциональный вид условной дисперсии. Чтобы избежать неправильной спецификации, следует использовать непараметрическую оценку $\sigma^2(x)$. Чтобы мотивировать использование оценки, заметим сперва, что $\sigma^2(x)$ по определению является обычным условным средним ε^2, ср. (12). Естественной оценкой условного среднего является ядерная регрессионная оценка, представленная в (4). В идеале хотелось бы вычислить $\hat{\sigma}^2(x) = \sum_{i=1}^n \varepsilon_i^2 K_h(X_i - x) / \sum_{i=1}^n K_h(X_i - x)$. Однако, так как ε_i, $i = 1, ..., n$, ненаблюдаемы, мы заменяем их остатками. Это приводит к следующей трехшаговой процедуре:

1. Рассчитать МНК-оценку, $\hat{\theta}_{\mathrm{OLS}}$, по формуле (2).

2. Рассчитать остатки, $\hat{\varepsilon}_i = Y_i - \hat{\theta}'_{\mathrm{OLS}} X_i$, $i = 1, ..., n$, и с их помощью непараметрически оценить условную дисперсию:

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2 K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}.$$

3. Получить ВМНК-оценку по формуле (13), но вместо $\sigma^2(x)$ подставить $\hat{\sigma}^2(x)$:

$$\hat{\theta}_{\mathrm{WLS}} = \left(\sum_{i=1}^n \hat{\sigma}^{-2}(X_i) X_i X_i' \right)^{-1} \left(\sum_{i=1}^n \hat{\sigma}^{-2}(X_i) X_i Y_i \right), \tag{14}$$

И вновь этот метод оценивания можно обобщить для более сложной параметрической модели

$$Y = g(X; \theta) + \varepsilon, \quad \mathbb{E}[\varepsilon | X] = 0,$$

где $g : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}$ известна с точностью до $\theta \in \Theta$.

3.3 Предположение о независимости

Можно использовать изложенную выше идею, чтобы получить ММП-оценки в случае, когда распределение ошибок имеет неизвестный вид. Мы сохраняем линейную спецификацию (11), но теперь предполагаем, что

$$\varepsilon \text{ и } X \text{ независимы},$$

так что $f_{\varepsilon|X}(\varepsilon|X) = f_\varepsilon(\varepsilon)$, где

$$\mathbb{E}[\varepsilon] = \int_{\mathbb{R}} z f_\varepsilon(z)\, dz = 0, \quad \sigma^2 = \int_{\mathbb{R}} z^2 f_\varepsilon(z)\, dz < \infty.$$

По сравнению с предыдущим разделом мы ввели дополнительное предположение о независимости регрессоров и ошибок. Однако мы не предполагаем, что известно распределение ε, так что и в этом случае модель остается полупараметрической.

Предположение о независимости позволяет оценить параметрическую компоненту с помощью полупараметрической ММП-оценки: допустим, плотность f_ε *известна*, тогда можно использовать ММП-оценку

$$\tilde{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_i \log f_\varepsilon \left(Y_i - \theta' X_i \right), \tag{15}$$

которая при условии регулярности будет удовлетворять

$$\sqrt{n}(\tilde{\theta}_{\text{MLE}} - \theta) \to^d N\left(0, H_0^{-1}\right),$$

где

$$H_0 = \mathbb{E}\left[\frac{\partial \log f_\varepsilon\left(Y_i - \theta' X_i\right)}{\partial \theta} \frac{\partial \log f_\varepsilon\left(Y_i - \theta' X_i\right)}{\partial \theta'}\right] = \int \frac{f_\varepsilon'(z)^2}{f_\varepsilon(z)} dz \mathbb{E}\left[XX'\right].$$

Однако плотность f_ε *неизвестна*, и, следовательно, $\tilde{\theta}_{\text{MLE}}$ недоступна. С другой стороны, заметим, что МНК по-прежнему применим и дает состоятельную оценку. Однако МНК-оценка будет неэффективна по сравнению с ММП-оценкой, так как $\int f_\varepsilon'(z)^2 / f_\varepsilon(z)\, dz \leq \sigma^2$, что превращается в равенство тогда и только тогда, когда f_ε – плотность нормального распределения $N\left(0, \sigma^2\right)$.

Чтобы улучшить эффективность ММП, мы собираемся получить полупараметрическую версию ММП-оценки с помощью трехшаговой процедуры:

1. Рассчитать МНК-оценку, $\hat{\theta}_{\text{OLS}}$, по формуле (2).

2. Рассчитать остатки, $\hat{\varepsilon}_i = Y_i - \hat{\theta}_{\text{OLS}}' X_i$, $i = 1, ..., n$, и с их помощью непараметрически оценить плотность безусловного распределения f_ε, например, как

$$\hat{f}_\varepsilon(x) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(\hat{\varepsilon}_i - x\right). \tag{16}$$

3. Получить ММП-оценку по формуле (15), но вместо f_ε подставить $\hat{f}_\varepsilon$,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_\varepsilon\left(Y_i - \theta' X_i\right). \tag{17}$$

И вновь данный метод оценивания можно обобщить для более сложной параметрической модели. Предположим, например, что

$$Y = g(X; \theta) + \sigma(X; \theta)\, \varepsilon,$$

где $g, \sigma : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}$ известны с точностью до $\theta \in \Theta$, ε и X независимы, и вдобавок

$$\mathbb{E}[\varepsilon] = \int_{\mathbb{R}} z f_\varepsilon(z)\, dz = 0, \quad \mathbb{E}\left[\varepsilon^2\right] = \int_{\mathbb{R}} z^2 f_\varepsilon(z)\, dz = 1.$$

Предположим, мы получили предварительную оценку θ, например ММП-оценку при нормальных ошибках, $\hat{\theta}_{\text{QMLE}}$, которая остается состоятельной даже если ошибки не нормально распределены. Тогда мы можем рассчитать остатки

$$\hat{\varepsilon}_i = \frac{Y_i - g(X_i; \hat{\theta}_{\text{QMLE}})}{\sigma(X_i; \hat{\theta}_{\text{QMLE}})}, \quad i = 1, ..., n,$$

и затем непараметрически оценить плотность по формуле (16). На заключительном этапе мы определим

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left\{\log \hat{f}_\varepsilon\left(\frac{Y - g(X; \theta)}{\sigma(X; \theta)}\right) + \log(\sigma(X; \theta))\right\}.$$

Мы ожидаем, что $\hat{\theta}$ будет наиболее эффективной оценкой, в отличие от $\hat{\theta}_{\text{QMLE}}$.

4 Копулы

Чтобы показать, что полупараметрическое моделирование имеет приложения за пределами регрессионного подхода, рассмотрим последний пример, касающийся копул. Копулы оказались удобным инструментом при моделировании ситуаций с многомерной зависимостью. В частности, их применяют в финансах, см., например, Genest, Gendron & Bourdeau-Brien (2009). В этом разделе мы рассмотрим полупараметрическое семейство копул и соответствующие оценки.

Пусть $Z = (Z_1, Z_2) \in \mathbb{R}^2$ – двумерная непрерывная случайная величина. Обозначим плотность функции совместного распределения (ФПР) и кумулятивную функцию распределения (КФР), соответственно, за f и F:

$$\Pr\{Z_1 \leq z_1, Z_2 \leq z_2\} = F(z_1, z_2) = \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} f(v_1, v_2)\, dv_1 dv_2,$$

и пусть f_k и F_k – маржинальные ФПР и КФР Z_k соответственно:

$$\Pr\{Z_k \leq z\} = F_k(z) = \int_{-\infty}^{z} f_k(v)\, dv, \quad k = 1, 2.$$

Так называемые копулы используют, чтобы смоделировать структурную зависимость между Z_1 и Z_2. При этом используется следующий известный факт: существует единственная функция $C : [0,1]^2 \mapsto [0,1]$, такая что

$$F(z_1, z_2) = C(F_1(z_1), F_2(z_2)),$$

(см. Joe, 1997). Функция C называется *копулой* для Z. Легко видеть, что C – КФР равномерно распределенной случайной величины $U := (F_1(Z_1), F_2(Z_2))$:

$$C(u_1, u_2) = \Pr\{F_1(Z_1) \leq u_1, F_2(Z_2) \leq u_2\}.$$

Кроме того, плотность совместного распределения Z можно записать в виде

$$f(z_1, z_2) = c(F_1(x_1), F_2(z_2)) f_1(z_1) f_2(z_2),$$

где $c : [0,1] \times [0,1] \mapsto \mathbb{R}_+$ – функция плотности распределения U.

Теперь можно смоделировать совместное распределение Z, задав два маржинальных распределения и копулу. В полностью параметрическом подходе это можно сделать, например, по формуле

$$f(z_1, z_2; \xi) = c(F_1(z_1; \alpha_1), F_2(z_2; \alpha_1); \theta) f_1(z_1; \alpha_1) f_2(z_2; \alpha_1),$$

где $\xi = \left(\theta', \alpha_1', \alpha_2'\right)' \in \Theta \times \mathcal{A}_1 \times \mathcal{A}_1$ – конечномерный параметр. После чего можно оценить ξ с помощью ММП:

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \{\log c(F_1(Z_{1,i}; \alpha_1), F_2(Z_{2,i}; \alpha_1); \theta) + \log f_1(Z_{1,i}; \alpha_1) + \log f_2(Z_{2,i}; \alpha_1)\}.$$

При большой размерности ξ это может оказаться сложной численной задачей. Вместо этого можно оценить параметры с помощью двухшаговой процедуры оценивания: сперва оценим $(\alpha_1', \alpha_2')'$ как

$$\hat{\alpha}_k = \arg\max_{\alpha_k \in \mathcal{A}_k} \frac{1}{n} \sum_{i=1}^{n} \log f_k(Z_{k,i}; \alpha_k), \quad k = 1, 2,$$

а затем –

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log c\left(F_1\left(Z_{1,i};\hat{\alpha}_1\right), F_2\left(Z_{2,i};\hat{\alpha}_2\right);\theta\right).$$

Эта двухшаговая оценка менее эффективна по сравнению с полной ММП-оценкой, но ее проще реализовать.

Тривиальная полупараметрическая модель с копулой имеет следующий вид. Мы, как и раньше, задаем параметрическое семейство копул, $c\left(u_1, u_2;\theta\right)$, но теперь оставляем оба маржинальных распределения незаданными. Мы хотим непараметрически оценить маржинальные распределения и на их основе сделать выводы относительно θ. Пусть

$$\hat{F}_k\left(z_k\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{Z_{k,i}\leq z_k\right\}, \quad k=1,2,$$

– выборочная кумулятивная функция распределения. Тогда естественной оценкой θ будет

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log c\left(\hat{F}_1\left(Z_{1,i}\right), \hat{F}_2\left(Z_{2,i}\right);\theta\right).$$

5 Класс двухшаговых оценок

В предыдущих двух разделах мы рассмотрели несколько примеров полупараметрических моделей и вывели оценки интересующих нас параметров. В этом разделе мы исследуем методы, с помощью которых можно проанализировать асимптотические свойства этих оценок. В частности, мы выясним, при каких условиях оценки будут $\sqrt{n}$-состоятельны и асимптотически нормально распределены.

Мы начнем с представления основного класса полупараметрических двухшаговых оценок. На первом шаге вычисляется предварительная непараметрическая оценка. На втором шаге эта непараметрическая оценка подставляется в целевую функцию, которая минимизируется с целью получить оценку параметрической компоненты. Этот достаточно большой класс оценок среди прочих включает в себя оценки, рассмотренные в предыдущих разделах. Мы дадим общие условия для состоятельности и асимптотической нормальности оценок параметрических компонент при соответствующих условиях регулярности.

Наша задача оценивания имеет много общего со стандартной параметрической двухшаговой задачей оценивания, где сперва оценивается шумовой параметр, а затем эту оценку используют его для получения оценки другого, интересующего нас, параметра. Единственное отличие заключается в том, что в нашем случае предварительная оценка является функцией, бесконечномерным параметром. Тем не менее, с небольшими видоизменениями можно использовать идею доказательства для параметрической двухшаговой оценки.

5.1 Структура

Мы хотим оценить конечномерный параметр $\theta\in\Theta\subseteq\mathbb{R}^k$ с помощью случайной целевой функции $Q_n\left(\theta,\gamma\right)$, где $\gamma\in\Gamma$ – некоторый бесконечномерный параметр, чаще всего функция. Целевая функция как правило будет являться функцией наблюдаемых данных, $\left(Y_i, X_i\right)$, где $i=1,...,n$, однако мы не будет явно задавать эту зависимость и обозначим ее лишь с помощью индекса n. Мы предполагаем, что пространство параметров Γ – линейное пространство с нормой $\|\cdot\|$. Это может быть, например, супремум-норма, $\|\gamma\| = \sup_x\left|\gamma\left(x\right)\right|$, или норма L_q, $\|\gamma\| = \left(\int\left|\gamma\left(x\right)\right|^q w\left(x\right)dx\right)^{1/q}$ для некоторой функции весов $w\left(x\right)\geq 0$.

Если известно истинное значение параметра γ, которое мы обозначим за γ_0, то можно оценить θ как

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta, \gamma_0). \tag{18}$$

В этом случае, чтобы вывести асимптотические свойства $\tilde{\theta}$, можно использовать стандартный результат для параметрических оценок (см., например, Newey & McFadden, 1994).

Мы рассмотрим случай, когда γ_0 неизвестна, и, следовательно, $\tilde{\theta}$ недостижима. Однако предположим, что доступна предварительная оценка, $\hat{\gamma}$. Тогда, заменяя γ_0 на $\hat{\gamma}$, мы можем записать

$$\hat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta, \hat{\gamma}). \tag{19}$$

Назовем $\hat{\theta}$ полупараметрической двухшаговой оценкой.

Прежде всего мы сделаем минимальные предположения относительно вида $Q_n(\theta,\gamma)$ и $\hat{\gamma}$ и потребуем лишь, чтобы последняя была состоятельной оценкой γ_0 и достаточно быстро сходилась. В большинстве случаев целевая функция имеет вид

$$Q_n(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} q(Z_i;\theta,\gamma), \tag{20}$$

но мы не будем ограничиваться только такими функциями.

Прежде чем перейти к анализу общей двухшаговой оценки, сперва покажем, как с помощью подходящего выбора $q(z;\theta,\gamma)$ можно представить оценки из предыдущих разделов в виде (19)–(20):

Пример 1: Одноиндексная модель. При $\gamma = g$ оценку для этой модели можно записать в виде (19)–(20), где q имеет вид

$$q(z;\theta,\gamma) = \left[y - \gamma\left(\theta' x\right)\right]^2,$$

а оценку $\hat{\gamma}_\theta$ можно выбрать как

$$\hat{\gamma}_\theta(z) = \frac{\sum_{i=1}^n Y_i K_h\left(\theta' X_i - z\right)}{\sum_{i=1}^n K_h\left(\theta' X_i - z\right)}.$$

Пример 2: Частично линейная модель. В этом случае, чтобы оценка приняла желаемый вид, надо взять

$$q(z;\theta,\gamma) = \left[y - \gamma_1(x_2) - \theta'(x_1 - \gamma_2(x_2))\right]^2,$$

где $\gamma_1(x_2) = \mathbb{E}[Y|X_2 = x_2]$, $\gamma_2(x_2) = \mathbb{E}[X_1|X_2 = x_2]$. Предварительными оценками будут

$$\hat{\gamma}_1(x_2) = \frac{\sum_{i=1}^n Y_i K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)}, \quad \hat{\gamma}_2(x_2) = \frac{\sum_{i=1}^n X_{1,i} K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)}.$$

Пример 3: Эффективное оценивание при гетероскедастичности. Определим функцию q как

$$q(z;\theta,\gamma) = \gamma^{-1}(x)\left[y - \theta' x\right]^2, \tag{21}$$

где $\gamma(x) = \sigma^2(x)$. ВМНК-оценка является частным случаем двухшаговой оценки. Здесь предварительными оценками будут

$$\hat{\gamma}(x_2) = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2 K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}. \tag{22}$$

Пример 4: Полупараметрические копулы. Функция q, определяющая оценку параметра копулы θ, задается формулой

$$q(z;\theta,\gamma) = \log c\left(\gamma_1(z_1), \gamma_2(z_2); \theta\right),$$

где $\gamma_k(z) = F_k(z)$ – маржинальная кумулятивная функция распределения Z_k, $k = 1, 2$, которая оценивается как

$$\hat{\gamma}_k(z) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{Z_{k,i} \leq z\right\}, \quad k = 1, 2.$$

В двух из этих примеров, а именно в частично линейной модели и регрессионной модели с неизвестной гетероскедастичностью, можно вывести в явном виде выражение для $\hat{\theta}$. То есть, можно провести прямой анализ этих конкретных оценок, и он будет гораздо проще по сравнению с косвенным анализом, предложенным ниже. Но в общем случае явные выражения для оценок недоступны, и все сводится к анализу свойств целевой функции $Q_n(\theta,\gamma)$.

В рамках этого подхода мы сперва установим общие условия, при которых $\hat{\theta}$ состоятельна и сходится к нормальному распределению. Налагая ограничения на вид целевой функции $Q_n(\theta,\gamma)$ и оценку $\hat{\gamma}$, мы в общих чертах покажем, как эти условия можно проверить в том конкретном случае, когда $\hat{\gamma}$ – ядерная оценка. Наконец, мы выясним асимптотические свойства первого порядка ВМНК-оценки, определенной в разделе 3.2, проверив выполнение общих условий для этой оценки.

5.2 Состоятельность

Доказательство состоятельности более или менее идентично доказательству в случае параметрических двухшаговых оценок. Единственное концептуальное отличие заключается в том, что теперь мы работаем с бесконечномерным параметром. Мы наложим следующие ограничения на целевую функцию:

C.1 Существует функция $Q(\theta,\gamma)$, такая что $\sup_{\theta\in\Theta}|Q_n(\theta,\gamma_0) - Q(\theta,\gamma_0)| \to^P 0$.

C.2 Для всех $\varepsilon > 0$, $\inf_{\|\theta-\theta_0\|>\varepsilon} Q(\theta,\gamma_0) > Q(\theta_0,\gamma_0)$.

C.3 Для некоторого $\lambda > 0$ и $B_n = O_P(1)$,

$$\sup_{\theta\in\Theta}|Q_n(\theta,\gamma) - Q_n(\theta,\gamma_0)| \leq B_n\|\gamma-\gamma_0\|^{\lambda}$$

для γ из окрестности γ_0.

Условие (C.1) означает, что недостижимая конечномерная целевая функция, $Q_n(\theta,\gamma_0)$, имеет определенный предел, $Q(\theta,\gamma_0)$. (C.2) – условие идентификации, то есть того, что целевая функция имеет единственный минимум $\theta_0 = \arg\min_{\theta\in\Theta} Q(\theta,\gamma_0)$. Легко показать, что условие (C.2) следует из следующих трех условий: Θ компакт, $\theta \mapsto Q(\theta,\gamma_0)$ непрерывна, и $Q(\theta,\gamma_0) > Q(\theta_0,\gamma_0)$ для всех $\theta \neq \theta_0$. Более примитивные условия для (C.1) можно найти в Newey (1991). Условия (C.1)–(C.2) означают, что недостижимая оценка $\tilde{\theta}$, заданная уравнением (18), состоятельна, $\tilde{\theta} \to^P \theta_0$; см., например, Newey & McFadden (1994, Теорема 2.1).

Последнее условие (C.3) утверждает, что разница между двумя целевыми функциями, $Q_n(\theta,\hat{\gamma})$ и $Q_n(\theta,\gamma_0)$, асимптотически пренебрежима: $Q_n(\theta,\hat{\gamma}) \to^P Q_n(\theta,\gamma_0)$ при $\hat{\gamma} \to^P \gamma_0$. Заметим, что норма $\|\gamma-\gamma_0\|$ функциональная, что обсуждалось в начале этого раздела. При этих предположениях достижимая оценка сходится к недостижимой, $\hat{\theta} = \tilde{\theta} + o_P(1)$.

Условия (C.1) и (C.3) можно заменить на следующие два условия:

C.1' $\sup_{\theta\in\Theta,\|\gamma-\gamma_0\|<\delta}|Q_n(\theta,\gamma)-Q(\theta,\gamma)|\to^P 0$, $\delta>0$.

C.3' $\sup_{\theta\in\Theta}|Q(\theta,\gamma)-Q(\theta,\gamma')|\to 0$ при $\gamma\to\gamma'$.

Для проверки условий (C.1') и (C.3') можно воспользоваться теорией эмпирических процессов, см. Andrews (1994a,b), Chen, Linton & van Keilegom (2003), van der Vaart & Wellner (1996). Эта проверка обычно включает в себя проверку условия Липшица, выраженное в (C.3).

Итоговые результаты о состоятельности сформулированы в следующих теоремах:

Теорема 1 *Пусть $Q_n(\theta,\gamma)$ удовлетворяет (C.1)–(C.3). Если $\hat\gamma\in\Gamma$ с некоторого момента по мере $\hat\gamma\to^P\gamma_0$, тогда $\hat\theta\to^P\theta_0$.*

Доказательство Доступно по запросу к автору.

Замечание 2 *В случае, когда $\hat\gamma$ зависит от θ, нужно усилить условие состоятельности до $\sup_{\theta\in\Theta}\|\hat\gamma_\theta-\gamma_\theta\|\to^P 0$.*

Теперь проверим условия (C.1)–(C.3) для ВМНК-оценки:

Пример 3 (продолжение). Обозначим как σ_0^2 и θ_0 истинные значения параметров. Мы предполагаем, что $X\in\mathcal{X}$, где $\mathcal{X}\subset\mathbb{R}^d$ компакт, и $\mathbb{E}\left[Y^2\right]<\infty$. Для идентификации нужно, чтобы $\mathbb{E}\left[XX'\sigma_0^{-2}(X)\right]$ была невырожденной. Предположение о компактности носителя $\mathcal{X}$ можно снять, но тогда нужно будет ввести подравнивание оценки, см. Robinson (1987).

Также предположим, что $\sigma_0^2(x), f(x)>0$ дважды непрерывно дифференцируемы. В частности, $\underline{\sigma}^2:=\inf_{x\in\mathcal{X}}\sigma_0^2(x)>0$. Введем ограничение, что Θ компакт, то есть существует константа c, такая что $\left|\theta'x\right|\le c$ для любых $\theta\in\Theta$ и $x\in\mathcal{X}$. Определим норму σ^2 как $\left\|\sigma^2\right\|_\infty=\sup_{x\in\mathcal{X}}\left|\sigma^2(x)\right|$ и предположим, что выполнено

$$\left\|\hat\sigma^2-\sigma_0^2\right\|_\infty\to^P 0, \tag{23}$$

где $\hat\sigma^2$ – ядерная оценка из формулы (22); это, например, можно сделать, используя результаты в Kristensen (2009).

Сперва мы покажем (C.3). Целевая функция принимает вид (21). Разложение в ряд Тейлора до первого порядка малости функции $a\mapsto 1/a$ дает

$$\begin{aligned} q(z;\theta,\hat\sigma^2)-q\left(z;\theta,\sigma_0^2\right) &= \left[y-\theta'x\right]^2\left(\frac{1}{\hat\sigma^2(x)}-\frac{1}{\sigma_0^2(x)}\right)\\ &= \left[y-\theta'x\right]^2\frac{-1}{\left[\lambda_x\hat\sigma^2(x)+(1-\lambda_x)\,\sigma_0^2(x)\right]^2}\left(\hat\sigma^2(x)-\sigma_0^2(x)\right)\end{aligned}$$

для некоторого $\lambda_x\in[0,1]$. В силу (23) $\inf_{x\in\mathcal{X}}\hat\sigma^2(x)\ge\underline{\sigma}^2/2$ почти наверное с некоторого момента по мере $n\to\infty$. Следовательно,

$$\frac{1}{\left[\lambda_x\hat\sigma^2(x)+(1-\lambda_x)\,\sigma_0^2(x)\right]^2}\le\frac{1}{[\lambda_x\underline{\sigma}^2/2+(1-\lambda_x)\,\underline{\sigma}^2/2]^2}=\frac{4}{\underline{\sigma}^4}<\infty.$$

Также, так как $\theta\mapsto\left(Y-\theta'X\right)^2$ непрерывна и $\left(Y-\theta'X\right)^2\le 3Y^2+3c^2$, где $\mathbb{E}\left[Y^2\right]<\infty$, из известных фактов о равномерной сходимости (см., например, Newey, 1991) следует, что

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\left(Y_i-\theta'X_i\right)^2-E[\left(Y_i-\theta'X_i\right)^2]\right|\to^P 0$$

и $\sup_{\theta\in\Theta} E[\left(Y-\theta' X\right)^2]<\infty$. Следовательно, $\sup_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \left(Y_i-\theta' X_i\right)^2 = O_P\left(1\right)$. Далее перепишем

$$\begin{aligned}\sup_{\theta\in\Theta}\left|Q_n\left(\theta,\hat\sigma^2\right)-Q_n\left(\theta,\sigma_0^2\right)\right| &\leq \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left|q(Z_i;\theta,\hat\sigma^2)-q\left(Z_i;\theta,\sigma_0^2\right)\right|\\ &\leq \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left[Y_i-\theta' X_i\right]^2\times\sup_x\left|\frac{1}{\hat\sigma^2\left(x\right)}-\frac{1}{\sigma_0^2\left(x\right)}\right|,\end{aligned}$$

где

$$\sup_x\left|\frac{1}{\hat\sigma^2\left(x\right)}-\frac{1}{\sigma_0^2\left(x\right)}\right|\leq\frac{4\left\|\hat\sigma^2-\sigma_0^2\right\|_\infty}{\underline{\sigma}^4}=o_P\left(1\right).$$

Следовательно, (C.3) выполняется для

$$B_n=\frac{4}{\underline{\sigma}^2}\times\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left[Y_i-\theta' X_i\right]^2,$$

и $\lambda=1$.

Теперь проверим (C.1). С помощью еще одного приложения равномерного закона больших чисел легко показать, что

$$\sup_{\theta\in\Theta}\left|Q_n\left(\theta,\sigma_0^2\right)-Q\left(\theta,\sigma_0^2\right)\right|\to^P 0,$$

где $\theta\mapsto Q\left(\theta,\sigma_0^2\right)=\mathbb{E}\left[\left[Y-\theta' X\right]^2\sigma_0^{-2}\left(X\right)\right]$ непрерывна.

Наконец, чтобы проверить (C.2), заметим, что для любых $\theta\neq\theta_0$,

$$\begin{aligned}Q\left(\theta,\sigma_0^2\right) &= \mathbb{E}\left[\left[\left(\theta_0-\theta\right)' X+\varepsilon\right]^2\sigma_0^{-2}\left(X\right)\right]\\ &= \left(\theta_0-\theta\right)'\mathbb{E}\left[XX'\sigma_0^{-2}\left(X\right)\right]\left(\theta_0-\theta\right)+\mathbb{E}\left[\varepsilon^2\sigma_0^{-2}\left(X\right)\right]\\ &> \mathbb{E}\left[\varepsilon^2\sigma_0^{-2}\left(X\right)\right]\\ &= Q\left(\theta_0,\sigma_0^2\right).\end{aligned}$$

Согласно замечаниям после условий (C.1)–(C.3), отсюда следует (C.2). Мы проверили все условия и согласно теореме 1 $\hat\theta$ состоятельна.

5.3 Асимптотическая нормальность

Для доказательства асимптотической нормальности мы воспользуемся тем же принципом, что и в случае параметрических двухшаговых оценок. Можно было бы разложить оценку, полученную на первом шаге, в ряд Тейлора, таким образом принимая во внимание дополнительную выборочную ошибку из-за оценивания на первом шаге. Однако, в наших условиях оценка первого шага является функцией, то есть бесконечномерным объектом. Поэтому, чтобы воспользоваться тем же методом, что и для параметрических двухшаговых оценок, нам сперва нужно обобщить понятие производных с конечномерного случая на бесконечномерный.

Пусть $T:\Gamma\mapsto\mathbb{R}^d$ – функционал, который ставит в соответствие $\gamma\in\Gamma$ некоторый евклидовый вектор. Например , $T\left(\gamma\right)=\int\gamma\left(x\right)dx$, $T_x\left(\gamma\right)=\partial\gamma\left(x\right)/\partial x$ и $T_x\left(\gamma_1,\gamma_2\right)=\gamma_1\left(x\right)\gamma_2\left(x\right)$.

Определение 3 *Говорят, что T дифференцируема по направлению в точке $\gamma\in\Gamma$, если существует линейный непрерывный функционал $\dot T\left(\gamma\right)\left[\cdot\right]:\Gamma\mapsto\mathbb{R}^d$ такой что*

$$\dot T\left(\gamma\right)\left[h\right]=\lim_{t\to 0}\frac{T\left(\gamma+th\right)-T\left(\gamma\right)}{t},$$

для всех $h\in\Gamma$.

$\dot{T}$ называется *производной по направлению* от T. В конечномерном случае, если T дифференцируема с производной $\partial T(\gamma)/\partial\gamma$, тогда $\dot{T}(\gamma)[h]$ – дифференциал T:

$$\dot{T}(\gamma)[h] = \frac{\partial T(\gamma)}{\partial\gamma}h.$$

Как правило, можно перенести результаты с конечномерного случая, используя производную по направлению. В частности, по-прежнему работает правило дифференцирования сложной функции.

Примеры функционалов. (i) $\Gamma = \{\gamma : \int|\gamma(x)|\,dx < \infty\}$ и $T(\gamma) = \int\gamma(x)\,dx$. Тогда $\dot{T}(\gamma)[h] = \int h(x)\,dx$ – линейная, непрерывная в пространстве L_1, и

$$T(\gamma + th) - T(\gamma) = \int(\gamma + th)(x)\,dx - \int\gamma(x)\,dx = t\int h(x)\,dx = t\dot{T}(\gamma)[h].$$

(ii) $\Gamma = \{\gamma|\exists\partial\gamma(x)/\partial x\}$ и $T(\gamma) = \partial\gamma(x)/\partial x$. Тогда $\dot{T}(\gamma)[h] = \partial h(x)/\partial x$:

$$T(\gamma + th) - T(\gamma) = \frac{\partial(\gamma + h)}{\partial x} - \frac{\partial\gamma}{\partial x} = t\frac{\partial h}{\partial x} = t\dot{T}(\gamma)[h].$$

Эти два случая просты, так как T – линейный функционал.

(iii) $T_x(\gamma) = F(\gamma(x))$, тогда $\dot{T}_x(\gamma)[h] = F'(\gamma(x))h(x)$.

(iv) $T(\gamma) = \int F(\gamma(x))\,dx$. Тогда $\dot{T}(\gamma)[h] = \int F'(\gamma(x))h(x)\,dx$ при определенных ограничениях на F и Γ.

Теперь мы хотим использовать производные по направлению для вычисления дополнительной выборочной дисперсии нашей оценки $\hat{\theta}$ из-за присутствия $\hat{\gamma}$. Сперва рассмотрим следующие функционалы: скор и Гессиан целевой функции

$$S_n(\theta,\gamma) = \frac{\partial Q_n(\theta,\gamma)}{\partial\theta}, \quad H_n(\theta,\gamma) = \frac{\partial^2 Q_n(\theta,\gamma)}{\partial\theta\partial\theta'}.$$

Теперь предположим, что существует производная по направлению $h \in \Gamma$ функции $S_n(\theta,\gamma)$ по γ в $(\theta,\gamma) = (\theta_0,\gamma_0)$, которую обозначим $\dot{S}_n(\theta_0;\gamma_0)[h]$. Тогда мы можем сформулировать условия, при которых имеет место асимптотическая нормальность:

N.1 $\|\hat{\gamma} - \gamma_0\| = o_P(n^{-1/4})$ и $\hat{\theta} \to^P \theta_0$.

N.2 $\theta_0 \in \mathrm{int}(\Theta)$.

N.3 $Q_n(\theta,\gamma)$ дважды непрерывно дифференцируемая по θ в окрестности $\mathcal{N}$ точки θ_0.

N.4 Производная по направлению $\dot{S}_n(\theta_0;\gamma_0)[h]$ существует и удовлетворяет условию

$$\left\|S_n(\theta_0,\gamma) - S_n(\theta_0,\gamma_0) - \dot{S}_n(\theta_0;\gamma_0)[\gamma - \gamma_0]\right\| \le B_n\|\gamma - \gamma_0\|^2,$$

где $B_n = O_P(1)$.

N.5 $\sqrt{n}\left\{S_n(\theta_0,\gamma_0) + \dot{S}_n(\theta_0;\gamma_0)[\hat{\gamma} - \gamma_0]\right\} \to^d N(0,\Omega_0)$.

N.6 $||H_n(\theta,\gamma) - H_n(\theta,\gamma_0)|| \le B_n\|\gamma - \gamma_0\|^\lambda$, где $B_n = O_P(1)$.

N.7 $\sup_{\theta\in\mathcal{N}}\|H_n(\theta,\gamma_0) - H(\theta,\gamma_0)\| \to^P 0$, где $H_0 = H(\theta_0,\gamma_0)$ невырожденна.

Как и в случае состоятельности, эти условия состоят из двух частей. (N.2), (N.3), (N.5) (положив $\dot{S}_n(\theta_0; \hat{\gamma} - \gamma_0) = 0$) и (N.7) означают, что недостижимая оценка, $\tilde{\theta}$, при известном γ_0 $\sqrt{n}$-асимптотически нормальна распределена (см., например, Newey & McFadden, 1994, Теорема 3.1). Оставшиеся условия, (N.1), (N.4) и (N.6), позволяют показать, что достижимая оценка также $\sqrt{n}$-асимптотически нормально распределена.

Теорема 4 *Пусть $\hat{\theta} \to^P \theta_0$, и выполняются (N.1)–(N.7). Тогда*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, H_0^{-1}\Omega_0 H_0^{-1}\right),$$

где Ω_0 и H_0 определены как в (N.5) и (N.7).

Доказательство Доступно по запросу к автору.

Как мы упоминали до теоремы 4, недостижимая оценка, $\tilde{\theta}$, также $\sqrt{n}$-асимптотически нормально распределена, если выполнены (N.1)–(N.7). Однако, $\tilde{\theta}$ в общем случае имеет меньшую асимптотическую дисперсию и потому эффективнее, чем $\hat{\theta}$. Асимптотические дисперсии этих двух оценок равны только если дополнительный член асимптотически исчезает, то есть $\sqrt{n}\dot{S}_n(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0] = o_P(1)$. В этом случае $\tilde{\theta}$ и $\hat{\theta}$ эквивалентны в первом приближении. Однако, в большинстве случаев из-за того, что мы не знаем γ_0, эффективность страдает, то есть $\mathbb{V}[\hat{\theta}] > \mathbb{V}[\tilde{\theta}]$.

Иногда вместо условий (N.3)–(N.7) проще проверить альтернативные условия:

N.3’ $Q_n(\theta, \gamma)$ непрерывно дифференцируема по θ в окрестности $\mathcal{N}$ точки θ_0.

N.4’ Существует функционал $S(\theta, \gamma)$, такой что $\nu_n(\theta, \gamma) := S_n(\theta, \gamma) - S(\theta, \gamma)$ удовлетворяет

$$\sup_{\|\theta - \theta_0\| < \delta, \|\gamma - \gamma_0\| < \delta} \|\nu_n(\theta, \gamma) - \nu_n(\theta_0, \gamma_0)\| = o_P\left(1/\sqrt{n}\right),$$

и $S(\theta_0, \gamma_0) = 0$.

N.5’ Производная по направлению $\dot{S}(\theta, \gamma)[h]$ от функции $S(\theta, \gamma_0)$ существует и удовлетворяет

$$\left\| S(\theta_0, \gamma) - S(\theta_0, \gamma_0) - \dot{S}(\theta_0, \gamma_0)[\gamma - \gamma_0] \right\| \le B \|\gamma - \gamma_0\|^2$$

для некоторой константы $B < \infty$.

N.6’ $\sqrt{n}(S_n(\theta_0, \gamma_0) + (\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0]) \to^d N(0, \Omega_0)$.

N.7’ Функция $S(\theta, \gamma)$ непрерывно дифференцируема по θ в окрестности $\mathcal{N}$ точки θ_0 с непрерывной производной $H(\theta, \gamma)$, которая удовлетворяет

$$\sup_{\theta \in \mathcal{N}} \|H(\theta, \gamma) - H(\theta, \gamma_0)\| \le B \|\gamma - \gamma_0\|^\lambda,$$

где $H_0 = H(\theta_0, \gamma_0)$ невырожденна.

Заметим, что мы ослабили условие (N.3), так что теперь требуется только, чтобы $Q_n(\theta, \gamma)$ имела одну производную. Условие (N.4’) более общее, но его можно проверить эмпирическими методами (см. необходимые условия, например, в Chen, Linton & van Keilegom, 2003). В большинстве случаев $S(\theta, \gamma)$ можно выбрать как $S(\theta, \gamma) = \partial Q(\theta, \gamma)/\partial\theta$, при этом условие идентификации в (C.2) обычно гарантирует, что $S(\theta_0, \gamma_0) = 0$.

Также заметим, что условия (N.5’) и (N.6’) включают в себя предельную скор-функцию $S(\theta, \gamma)$, а не $S_n(\theta, \gamma)$, как в (N.5) и (N.6).

Теорема 5 *Пусть $\hat{\theta} \to^P \theta_0$, и выполняются условия (N.1)–(N.2) и (N.3')–(N.7'). Тогда заключение теоремы 4 остается верным.*

Доказательство Доступно по запросу к автору.

Если в теореме 4 требуется выполнение ЦПТ для $\sqrt{n}(S_n(\theta_0, \gamma_0) + \dot{S}_n(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0])$, то в теореме 5 – для $\sqrt{n}(S_n(\theta_0, \gamma_0) + \dot{S}(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0])$. При применении этих теорем основная проблема заключается в том, чтобы проверить выполнение ЦПТ для соответствующего слагаемого. На первый взгляд может показаться, что это невозможно из-за наличия $\dot{S}_n(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0]$ и $\dot{S}(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0]$, так как оба слагаемых включают в себя непараметрическую оценку, $\hat{\gamma}$, которая вообще сходится медленнее, чем $\sqrt{n}$. Однако, если подставить $\hat{\gamma}$ в производную по направлению, то происходит дополнительное полное сглаживание непараметрической оценки. Как мы увидим далее, это сглаживание в общем случае увеличивает скорость сходимости и делает возможным проверить условия (N.5) или (N.6').

Чтобы продемонстрировать, как можно проверить $\sqrt{n}$-сходимость, мы ограничимся случаем, когда скор-функция имеет вид

$$S_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} s(Z_i; \theta, \gamma) + o_P(n^{-1/2}),$$

где $Z_i \in \mathbb{R}^d$, $i = 1, ..., n$, независимо и одинаково распределены. Это ограничение выполняется, когда, например, $Q_n(\theta, \gamma)$ задано как в (20), при этом $s(z; \theta, \gamma) = \partial q(z; \theta, \gamma)/\partial\theta$. При этих ограничениях производные по направлению $S_n(\theta, \gamma)$ и $S(\theta, \gamma)$ задаются как

$$\dot{S}_n(\theta, \gamma)[h] = \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i; \theta, \gamma)[h], \quad \dot{S}(\theta, \gamma)[h] = \mathbb{E}[\dot{s}(Z; \theta, \gamma)[h]],$$

где $\dot{s}(z; \theta, \gamma)[h]$ – производная по направлению h функции $s(z; \theta, \gamma)[h]$ по γ.

Сперва мы сформулируем достаточные условия, при которых выполняются (N.4)–(N.5):

N.4.i $\|s(z; \theta_0, \gamma) - s(z; \theta_0, \gamma_0) - \dot{s}(z; \theta_0, \gamma_0)[\gamma - \gamma_0]\| \leq b(z)\|\gamma - \gamma_0\|^2$, причем $\mathbb{E}[b(Z)] < \infty$.

N.5.i $n^{-1}\sum_{i=1}^{n} \dot{s}(Z_i; \theta_0, \gamma_0)[\hat{\gamma} - \gamma_0] = \dot{S}(\theta_0, \gamma_0)[\hat{\gamma} - \gamma_0] + o_P(1/\sqrt{n})$.

N.5.ii Существует функция $\delta : \mathbb{R}^d \mapsto \mathbb{R}^k$ с $\mathbb{E}[\delta(Z)] = 0$ и $E[\|\delta(Z)\|^2] < \infty$, такая что

$$\dot{S}(\theta_0, \gamma_0)[\hat{\gamma} - \gamma_0] = \frac{1}{n}\sum_{i=1}^{n} \delta(Z_i) + o_P\left(1/\sqrt{n}\right).$$

N.5.iii $\mathbb{E}[s(Z; \theta_0, \gamma_0)] = 0$ и $E[\|s(Z; \theta_0, \gamma_0)\|^2] < \infty$.

Лемма 6 *Пусть $\dot{s}(z; \theta, \gamma)[h]$ существует и удовлетворяет (N.4.i)–(N.5.iii). Тогда условия (N.4) и (N.5) выполняются, причем*

$$\Omega = \mathbb{E}\left[[s(Z; \theta_0, \gamma_0) + \delta(Z)][s(Z; \theta_0, \gamma_0) + \delta(Z)]'\right].$$

Доказательство Доступно по запросу к автору.

Хотя (N.4.i)–(N.5.iii) более простые условия, во многих случаях по-прежнему не очевидно, как их проверить. В частности, не так очевидно доказать, что выполняются предположения (N.5.i)–(N.5.ii). Поэтому в дальнейшем мы предположим, что $\hat{\gamma}$ может быть записана в виде

$$\hat{\gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} w_n(x, Z_i) + o_P(n^{-1/4}),$$

для некоторой функции w_n, которая может зависеть от размера выборки n. Это ограничение выполняется, например, для ядерной оценки, если положить $w_n(x, Z_i) = Y_i K_h(X_i - x)$. Можно легко убедиться, что оценки решетом тоже сюда подпадают (см., например, Newey, 1997). Далее, мы опустим зависимость от (θ_0, γ_0), и будем писать, например, просто $\dot{S}[\hat{\gamma} - \gamma_0]$ вместо $\dot{S}(\theta_0, \gamma_0)[\hat{\gamma} - \gamma_0]$.

Проверка (N.5.i). Сперва заметим, что так как $\dot{s}$ линейная функция, можно записать

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\hat{\gamma} - \gamma_0] - \dot{S}[\hat{\gamma} - \gamma_0] &= \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\hat{\gamma} - \bar{\gamma}] - \dot{S}[\hat{\gamma} - \bar{\gamma}] \\ &+ \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\bar{\gamma} - \gamma_0] - \dot{S}[\bar{\gamma} - \gamma_0] \\ &=: I_{n,1} + I_{n,2}, \end{aligned}$$

где $\bar{\gamma}(x) = \mathbb{E}[\hat{\gamma}(x)]$. Определив

$$V_n(x, x') = \dot{s}(z)[w_n(\cdot, x')], \quad V_n = \mathbb{E}[V_n(Z_1, Z_2)],$$

$$V_{n,1}(x) = \mathbb{E}[V_n(x, Z)], \quad V_{n,2}(x) = \mathbb{E}[V_n(Z, x)],$$

и опять пользуясь тем, что $\dot{s}$ – линейный функционал, запишем:

$$\begin{aligned} I_{n,1} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \dot{s}(Z_i)[w_n(\cdot, Z_j)] - \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[E[w_n(\cdot, Z)]] \\ &\quad -\frac{1}{n}\sum_{i=1}^{n} \dot{S}[w_n(\cdot, Z_j)] + \dot{S}[E[w_n(\cdot, Z_2)]] \\ &= \frac{1}{n}\sum_{i,j=1}^{n} V_n(Z_i, Z_j) - \frac{1}{n}\sum_{i=1}^{n} V_{n,1}(Z_i) - \frac{1}{n}\sum_{j=1}^{n} V_{n,2}(Z_j) - V_n. \end{aligned}$$

Теперь можно использовать результаты для так называемых U-статистик (см., например, Lee, 1990 в качестве введения),чтобы показать, что правая часть есть $o_P(n^{-1/2})$ в самом общем случае. Итак, с первым членом порядок. Что касается второго, обычно можно показать, что

$$\|\dot{s}(z; \theta_0, \gamma_0)[h]\| \le b(z)\|h\|,$$

в случае чего

$$\mathbb{E}[\|I_{n,2}\|] \le \mathbb{E}[b(Z)]\|[E[\hat{\gamma}] - \gamma_0]\|,$$

и теперь нужно показать, что смещение достаточно быстро стремится к нулю: $\|E[\hat{\gamma}] - \gamma_0\| = o_P(n^{-1/2})$. В случае, когда $\hat{\gamma}$ – ядерная оценка, это условие можно проверить в самом общем случае, комбинируя так называемые ядра высокого порядка с недостаточным сглаживанием.

Проверка (N.5.ii). Это условие обычно проверяют, сначала показав, что существует функция d, такая что

$$\dot{S}[h] = \int d(x)\, h(x)\, dx.$$

Как правило, это выводится непосредственно, если известно явное выражение для $\dot{S}$, см. Newey (1994b). В качестве альтернативы можно воспользоваться теоремой представления Рисса, как это сделано в Aït-Sahalia (1993).

При заданном представлении, как правило, можно найти функцию δ. Предположим, например, что $\gamma_0(x) = f_X(x)\mathbb{E}[Y|X=x]$, $\hat{\gamma}(x) = 1/n\sum_{i=1}^n Y_i K((X_i - x)/h)/h^d$, и $\dot{S}[h] = \int d(x)h(x)\,dx$. Тогда сперва запишем

$$\dot{S}[\hat{\gamma} - \gamma_0] = \dot{S}[\hat{\gamma}] - \dot{S}[\gamma_0].$$

Первое слагаемое удовлетворяет

$$\dot{S}[\hat{\gamma}] = \int d(x)\hat{\gamma}(x)\,dx = \frac{1}{n}\sum_{i=1}^n Y_i \frac{1}{h^d}\int d(x) K\left(\frac{X_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^n Y_i d(X_i) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

где последнее равенство следует из соответствующего условия регулярности, так как

$$h^{-d}\int d(x) K\left(\frac{X_i - x}{h}\right)dx \to d(X_i)$$

при $h \to 0$. Второе слагаемое можно записать как

$$\dot{S}[\gamma_0] = \int d(x)\gamma_0(x)\,dx = \int d(x) f_X(x)\mathbb{E}[Y|X=x]\,dx = \mathbb{E}[Yd(X)].$$

Полагая $\delta(z)$ равным

$$\delta(z) := yd(x) - \mathbb{E}[Yd(X)],$$

мы получим желаемое:

$$\dot{S}[\hat{\gamma} - \gamma_0] = \frac{1}{n}\sum_{i=1}^n \delta(Z_i) + o_P\left(1/\sqrt{n}\right).$$

Дальнейшие шаги по проверке (N.5.ii) для ядерной оценки можно найти в Newey (1994b).

Пример 3 (продолжение). Допустим, мы уже проверили, что

$$\left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty = o_P(n^{-1/4}),$$

для некоторой последовательности ширины окна (см., например, Kristensen, 2009).

Чтобы вывести асимптотическое распределение $\hat{\theta}$, сперва найдем скор-функцию и Гессиан:

$$\begin{aligned} S_n(\theta, \sigma^2) &= \frac{\partial Q_n(\theta, \sigma^2)}{\partial\theta} = -\frac{2}{n}\sum_{i=1}^n \sigma^{-2}(X_i)\left[Y_i - \theta' X_i\right]X_i, \\ H_n(\theta, \sigma^2) &= \frac{\partial^2 Q_n(\theta, \sigma^2)}{\partial\theta\partial\theta'} = \frac{2}{n}\sum_{i=1}^n \sigma^{-2}(X_i) X_i X_i'. \end{aligned}$$

Сделаем следующую догадку относительно производной по направлению скор-функции:

$$\dot{S}_n[h] = \frac{1}{n}\sum_{i=1}^n \dot{s}(Z_i)[h],$$

где

$$\dot{s}(z)[h] = \left[y - \theta_0' x\right]x'\sigma_0^{-4}(x)h(x) = d(y,x)h(x),$$

$$d(y,x) = \left[y - \theta_0' x\right]x'\sigma_0^{-4}(x),$$

и h – направление производной по σ^2. Здесь мы опустили зависимость $\dot{S}_n[h]$ от θ_0 и $\sigma_0^2(x)$. Заметим, что

$$d(Y,X) = \left[Y - \theta_0' X\right]X\sigma_0^{-4}(X)h(X) = \varepsilon X\sigma_0^{-4}(X)h(X).$$

Проверяем, что необходимые условия удовлетворены. Во-первых, $h \mapsto \dot{S}_n[h]$ линейна; во-вторых, разложение в ряд Тейлора до второго порядка малости функции $a \mapsto 1/a$ дает

$$\frac{1}{a} - \frac{1}{a_0} = -\frac{1}{a_0^2}(a - a_0) + \frac{2}{(\lambda a + (1-\lambda) a_0)^3}(a - a_0)^2$$

для некоторого $\lambda \in [0,1]$. Полагая $a = \hat{\sigma}^2(x)$, $a_0 = \sigma_0^2(x)$, получим

$$\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) = -\sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right) + \frac{2\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)^2}{\left(\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right)^3},$$

где

$$\left|\frac{2}{\left(\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right)^3}\right| \leq \frac{16}{\underline{\sigma}^6},$$

с той же самой аргументацией, что и при доказательстве состоятельности. Следовательно,

$$\begin{aligned}
&\left\|S_n\left(\theta_0, \hat{\sigma}^2\right) - S_n\left(\theta_0, \sigma_0^2\right) - \dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right]\right\| \\
\leq\ & \frac{2}{n}\sum_{i=1}^{n}\left\|\varepsilon_i X_i'\right\|\left|\hat{\sigma}^{-2}(X_i) - \sigma_0^{-2}(X_i) + \sigma_0^{-4}(X_i)\left(\hat{\sigma}^2(X_i) - \sigma_0^2(X_i)\right)\right| \\
\leq\ & \left(\frac{2}{n}\sum_{i=1}^{n}\left\|\varepsilon_i X_i'\right\|\right)\sup_{x\in\mathcal{X}}\left|\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) + \sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)\right|,
\end{aligned}$$

где $n^{-1}\sum_{i=1}^{n}\|\varepsilon X'\| = \mathbb{E}\left[\|\varepsilon X'\|\right] + o_P(1)$ по ЗБЧ, в то время как

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left|\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) + \sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)\right| &\leq \frac{16}{\underline{\sigma}^6}\sup_{x\in\mathcal{X}}\left|\hat{\sigma}^2(x) - \sigma_0^2(x)\right|^2 \\
&= \frac{16}{\underline{\sigma}^6}\left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty^2 \\
&= o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Делаем вывод, что

$$\sqrt{n}S_n\left(\theta_0, \sigma^2\right) = \sqrt{n}S_n\left(\theta_0, \sigma_0^2\right) + \sqrt{n}\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] + o_P(1).$$

Далее, для любого h,

$$\begin{aligned}
\dot{S}[h] &= \mathbb{E}\left[\dot{s}(Z)[h]\right] = \mathbb{E}\left[d(Y,X)\,h(X)\right] \\
&= \mathbb{E}\left[\varepsilon\sigma_0^{-4}(X)\,h(X)\,X'\right] = \mathbb{E}\left[\mathbb{E}[\varepsilon|X]\,\sigma_0^{-4}(X)\,h(X)\,X'\right] \\
&= 0,
\end{aligned}$$

так как $\mathbb{E}[\varepsilon|X] = 0$. Это означает, что добавочное слагаемое $\delta(Z) \equiv 0$. Таким образом, если мы сможем проверить, что

$$\sqrt{n}\left(\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] - \dot{S}\left[\hat{\sigma}^2 - \sigma_0^2\right]\right) \to^P 0,$$

то сможем заключить, что

$$\sqrt{n}\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] = o_P(1).$$

Гессиан удовлетворяет

$$\begin{aligned}
\left\|H_n\left(\hat{\sigma}^2\right)-H_n\left(\sigma_0^2\right)\right\| &\leq \frac{2}{n}\sum_{i=1}^{n}\|X_i\|^2\left|\hat{\sigma}^{-2}\left(X_i\right)-\sigma_0^{-2}\left(X_i\right)\right| \\
&\leq \left(\frac{2}{n}\sum_{i=1}^{n}\|X_i\|^2\right)\sup_{x\in\mathcal{X}}\left|\hat{\sigma}^{-2}\left(x\right)-\sigma_0^{-2}\left(x\right)\right| \\
&= O_P\left(1\right)\times o_P\left(1\right),
\end{aligned}$$

и $H_n\left(\sigma_0^2\right)\to^P H_0=\mathbb{E}\left[\sigma_0^{-2}\left(X\right)XX'\right]$ по ЗБЧ.

Собрав полученные результаты, получим

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\sqrt{n}S_n\left(\theta_0,\hat{\sigma}^2\right) \\
&= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\left\{\sqrt{n}S_n\left(\theta_0,\sigma_0^2\right)+\sqrt{n}\dot{S}_n\left[\hat{\sigma}^2-\sigma_0^2\right]+o_P\left(1\right)\right\} \\
&= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\left\{\sqrt{n}S_n\left(\theta_0,\sigma_0^2\right)+o_P\left(1\right)\right\} \\
&\to^d N\left(0,H_0^{-1}\Omega_0H_0^{-1}\right),
\end{aligned}$$

где

$$\Omega_0=\mathbb{E}\left[\sigma^{-4}\left(X\right)\varepsilon^2XX'\right]=\mathbb{E}\left[\sigma^{-2}\left(X\right)XX'\right],$$

так что

$$H_0^{-1}\Omega_0H_0^{-1}=\mathbb{E}\left[\sigma^{-2}\left(X\right)XX'\right]^{-1}.$$

Заметим, что недостижимая ВМНК-оценка

$$\tilde{\theta}=\arg\min_{\theta\in\Theta}Q_n\left(\theta,\sigma_0^2\right)$$

имеет такое же асимптотическое распределение, что и $\hat{\theta}$. Следовательно, достижимая ВМНК-оценка, основанная на непараметрической оценке $\hat{\sigma}^2\left(x\right)$, асимптотически эквивалентна недостижимой ВМНК-оценке. Так что при замене σ_0^2 на $\hat{\sigma}^2$ в оценке θ мы не теряем в эффективности. Однако заметим, что это не означает, что $\hat{\theta}$ обязательно наиболее эффективная из доступных оценок, так как не выполняется равенство Рао–Крамера (за исключением случая, когда нормированные ошибки $\sigma^{-1}\left(X\right)\varepsilon$ имеют независимое и одинаковое стандартное нормальное распределение).

5.4 Оценивание дисперсий

Положив $\hat{H}=H_n(\hat{\theta},\hat{\gamma})$, можно оценить H_0, при этом оценка будет состоятельна при выполнении условий теоремы 4. Если нам известна оценка $\hat{\delta}$ параметра δ, то при условиях регулярности

$$\hat{\Omega}=\frac{1}{n}\sum_{i=1}^{n}\left[s(Z_i;\hat{\theta},\hat{\gamma})+\hat{\delta}(Z_i)\right]\left[s(Z_i;\hat{\theta},\hat{\gamma})+\hat{\delta}(Z_i)\right]'$$

будет состоятельной оценкой для Ω_0. Обычно можно вывести явное выражение для δ, так как $\delta\left(z\right)=\delta\left(z;\theta_0,\gamma_0\right)$, при этом естественно взять $\hat{\delta}(z)=\delta(z;\hat{\theta},\hat{\gamma})$. Стандартными способами можно показать, что оценка дисперсии будет состоятельной, если s и δ удовлетворяют условию Липшица по (θ,γ), см., например, Newey & McFadden (1994, Теорема 8.13).

Однако в сложных моделях нельзя получить выражение для δ в явном виде (см., например, Kristensen, 2008). В этом случае можно воспользоваться бутстрапом (Chen, Linton & van Keilegom, 2003) или численными методами (Newey, 1994a).

6 Оценивание «решетом»

Несмотря на то, что во многих случаях полупараметрические модели можно оценивать двухшаговой процедурой, существует альтернативный подход, который состоит в одновременном оценивании как параметрической, так и непараметрической компонент. Обсудим, как это реализуется в контексте метода «решето».

Как и в предыдущем разделе, мы хотим оценить параметр $\theta \in \Theta \subseteq \mathbb{R}^k$, используя целевую функцию $Q_n(\theta, \gamma)$, где $\gamma \in \Gamma$ – бесконечномерный параметр. Вместо того, чтобы использовать предварительную оценку (если такая вообще доступна) и воспользоваться двухшаговой процедурой, можно оценить θ и γ одновременно, используя так называемое решето. Метод «решето» – это общий непараметрический метод, в котором бесконечномерные функциональные пространства заменяются на приближающие их конечномерные (так называемое решето) в конечных выборках. Ошибка приближения, возникающая из-за перехода к конечномерным пространствам, асимптотически исчезает с ростом объема выборки ввиду увеличения размерности решета.

Для того, чтобы дать определение полупараметрической оценке методом решето, сначала введем дополнительные обозначения. Пусть выбрана последовательность аппроксимирующих конечномерных пространств $\{\Gamma_J\}$, такая что $\Gamma_J \subseteq \Gamma$, $J \geq 1$, и $\overline{\bigcup_{J=1}^{\infty} \Gamma_J} = \Gamma$. Определим

$$(\hat{\theta}, \hat{\gamma}) = \arg \min_{\theta \in \Theta, \gamma \in \Gamma_{J_n}} Q_n(\theta, \gamma) \tag{24}$$

для некоторой последовательности $J_n \to \infty$ при $n \to \infty$. Здесь мы оцениваем θ и γ одновременно, используя ту же целевую функцию, Q_n. При использовании же двухшаговых оценок, рассмотренных в предыдущем разделе, использовались две разные целевые функции для получения оценок θ и γ соответственно.

Общие результаты относительно состоятельности и скорости сходимости для случая, где $Q_n(\theta, \gamma)$ принимает вид выборочного среднего как в уравнении (20), можно найти в работах Shen & Wong (1994) и Shen (1997). Более того, в этих статьях выводятся условия, при которых оценка $\hat{\theta}$ $\sqrt{n}$-асимптотически нормальна. ОММ-подобные решетчатые оценки для моделей, определяемых через условия на условные моменты, разработаны и проанализированы в Ai & Chen (2003); см. также Blundell, Chen & Kristensen (2007). Поскольку условия, необходимые для получения указанных результатов, являются сугубо техническими (как и доказательства), мы не станем здесь вдаваться в детали.

Один из недостатков подхода, рассмотренного выше, связан с его практической реализацией. При двухшаговом оценивании $\hat{\gamma}$ используется в качестве предварительной оценки, так что достаточно решить оптимизационную задачу малой размерности $\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(\theta, \hat{\gamma})$. С другой стороны, решетчатые оценки требуют одновременной оптимизации как по θ, так и по γ. В частности, размерность γ может оказаться достаточно большой в стандартных задачах и будет расти экспоненциально с ростом числа переменных, от которых берется функция. Таким образом, численная задача решения $(\hat{\theta}, \hat{\gamma})$ в уравнении (24) является задачей «высокой» размерности и может оказаться расчетно невыполнимой. Тем не менее, во многих случаях можно найти решение в явном виде, что упрощает вычисления.

Рассмотрим два примера, в которых показано, как можно получить полупараметрические решетчатые оценки.

Пример 1 (продолжение). Когда $\gamma = g$, целевая функция для одноиндексной модели имеет форму (20), где q задается как

$$q(z; \theta, \gamma) = \left[y - \gamma\left(\beta' x\right)\right]^2.$$

Пусть Γ – некоторое функциональное пространство, для которого существует решето в форме

$$\Gamma_J = \left\{ \gamma_J(z) = \sum_{j=1}^{J} \alpha_j \varphi_j(z) : \alpha_j \in \mathbb{R},\ j = 1, ..., J \right\} \tag{25}$$

где $\varphi_1(z), \varphi_2(z), ...$ – *известные* базисные функции. Решетчатая оценка, определенная в уравнении (24), принимает вид

$$(\hat{\beta}, \hat{\gamma}) = \arg \min_{\beta, A_{J_n}} \sum_{i=1}^{n} \left[Y_i - A'_{J_n} \Phi_{J_n}(\beta' X_i) \right]^2,$$

где $A_{J_n} = (\alpha_1, ..., \alpha_{J_n})'$ и $\Phi_{J_n}(z) = (\varphi_1(z), ..., \varphi_{J_n}(z))'$. Для любого заданного значения β условия первого порядка по A_{J_n} имеют вид

$$\sum_{i=1}^{n} \left[Y_i - A'_{J_n} \varphi_{J_n}(\theta' X_i) \right] \Phi_{J_n}(\beta' X_i) = 0,$$

откуда получаем решение:

$$\hat{A}_{J_n}(\beta) = \left(\sum_{i=1}^{n} \Phi_{J_n}(\beta' X_i) \Phi_{J_n}(\beta' X_i)' \right)^{-1} \sum_{i=1}^{n} \Phi_{J_n}(\beta' X_i) Y_i.$$

Подставив, получаем «профильную» оценку:

$$\hat{\beta} = \arg \min_{\theta} \sum_{i=1}^{n} \left[Y_i - \hat{A}_{J_n}(\beta)' \Phi_{J_n}(\beta' X_i) \right]^2.$$

Здесь сложность вычислений ограничивается лишь численной оптимизацией по θ. Заметим, что в данном случае одновременная оценка совпадает с оценкой по двухшаговому методу, где оценка сериями используется в качестве предварительной оценки γ.

Пример 4 (продолжение). В случае полупараметрической модели с копулой $\gamma = (f_1, f_2)$, и целевую функцию снова можно записать в виде (20), где q задается:

$$q(z; \theta, \gamma) = \{ \log c(F_1(z_1), F_2(z_2); \theta) + \log f_1(z_1) + \log f_2(z_2) \}.$$

Эта оценка была предложена в Chen, Fan & Tsyrennikov (2006), где также указан способ, которым можно построить решетчатое пространство для двух плотностей. Решетчатую оценку в общем случае нельзя записать в явном виде, и необходимо решать задачу численной оптимизации, что делает этот метод не слишком привлекательным. Решетчатая оценка, составленная таким образом, в общем случае будет более эффективной, чем двухшаговая оценка, построенная в разделе 4.

Хотя в первом из этих примеров решетчатая и двухшаговая ядерная оценки очень похожи, в общем случае решетчатая оценка будет отличаться. В частности, решетчатые оценки будут более эффективными по сравнению с двухшаговыми ядерными оценками в силу их построения, поскольку параметрическая и непараметрическая компоненты оцениваются одновременно. Это приводит нас к вопросу об эффективности полупараметрических оценок.

7 Полупараметрическая эффективность

Итак, любая полупараметрическая модель полностью характеризуется параметрической компонентой, θ_0, и непараметрической, $\gamma_0(\cdot)$. Нас интересуют параметр θ_0 и насколько эффективно можно его оценить, ничего не зная априори о непараметрической компоненте, $\gamma_0(\cdot)$. В общем случае ответить на этот вопрос нелегко, но существует конструктивный подход, с помощью которого можно рассчитать границы уровня эффективности для θ_0.

Содержательный смысл описываемых границ следующий. Рассмотрим оценивание двух статистических моделей, в которых вторая модель содержится в первой (встроена в нее). Естественно ожидать, что оценивание второй модели будет проще по сравнению с первой. В частности, если обе модели содержат общий параметр, например θ, то следует ожидать, что этот параметр во второй модели будет оценен точнее. Таким образом, если мы сможем рассчитать эффективность оценивания θ во второй модели, то мы получим границы эффективности θ для первой модели.

Stein (1956) использовал эту идею для построения границ эффективности в задачах полупараметрического оценивания. В качестве первой, более сложной модели он выбрал полупараметрическую модель, которая характеризуется набором $(\theta_0, \gamma_0(\cdot))$. А в качестве второй, более простой модели он использовал полностью параметрическую подмодель. Выберем некоторое параметрическое семейство функций, $\gamma(\cdot;\alpha)$, где $\alpha \in \mathcal{A} \subseteq \mathbb{R}^l$ – параметр, и предположим, что параметрическая подмодель содержит истинную функцию $\gamma_0(\cdot)$, $\gamma_0(\cdot) = \gamma(\cdot;\alpha_0)$ для некторого $\alpha_0 \in \mathcal{A}$. Таким образом, вторая модель характеризуется набором (θ_0, α_0).

Оценивание полупараметрической модели должно быть как минимум не легче, чем полностью параметрической подмодели. Таким образом, следует ожидать, что в полупараметрической модели мы не сможем оценить θ_0 точнее. Поскольку параметрическая модель полностью специфицирована в терминах (θ, α), мы можем записать плотность модели как функцию (θ, α), $(\theta, \alpha) \mapsto p(z; \theta, \gamma(\cdot;\alpha))$. Естественной в таком случае будет ММП-оценка, а ее точность определяется соответствующей информационной матрицей Фишера,

$$\mathcal{I} = \mathbb{E}\left[\frac{\partial^2 \log p(Z;\theta,\gamma(\cdot;\alpha))}{\partial(\theta,\alpha)\,\partial(\theta,\alpha)'}\right] = \begin{bmatrix} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\alpha} \\ \mathcal{I}_{\theta\alpha} & \mathcal{I}_{\alpha\alpha} \end{bmatrix}.$$

Для любой заданной параметрической спецификации, $\gamma(\cdot;\alpha)$, уровень эффективности для θ задается неравенством Рао–Крамера,

$$\mathcal{I}_{\mathrm{p}} = \mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\alpha}\mathcal{I}_{\alpha\alpha}^{-1}\mathcal{I}_{\theta\alpha}.$$

То есть, асимптотическая дисперсия ММП-оценки равна $\mathcal{I}_{\mathrm{p}}^{-1}$. Это значение дисперсии выражает цену, которую мы вынуждены платить за незнание $\gamma(\cdot)$ (что соответствует α в параметрической модели). Если α известно, θ можно оценить с асимптотической дисперсией $\mathcal{I}_{\theta\theta}^{-1}$. Если α неизвестно и его нужно оценивать, дисперсия будет $\mathcal{I}_{\mathrm{p}}^{-1} \geq \mathcal{I}_{\theta\theta}^{-1}$, где равенство возможно тогда и только тогда, когда $\mathcal{I}_{\theta\alpha} = 0$.

Рассмотрим теперь оценку, не полагающуюся на параметрическую информацию о $\gamma(\cdot)$, и пусть $\mathcal{I}_{\mathrm{sp}}^{-1}$ – ее асимптотическая дисперсия. Тогда должно выполняться $\mathcal{I}_{\mathrm{p}}^{-1} \leq \mathcal{I}_{\mathrm{sp}}^{-1}$. Утверждение верно независимо от выбора параметрической подмодели, так что $\sup_{\gamma(\cdot;\alpha)} \mathcal{I}_{\mathrm{p}}^{-1} \leq \mathcal{I}_{\mathrm{sp}}^{-1}$. Из этого следует следующее определение границ эффективности полупараметрической оценки как асимптотической дисперсии «наименее благоприятной» параметрической подмодели:

$$\text{semiparametric efficiency bound (SEB)} = \sup_{\gamma(\cdot;\alpha)} \mathcal{I}_{\mathrm{p}}^{-1}.$$

Одним из привлекательных является класс полупараметрических оценок, который хорош настолько же, как если бы наличествовала полная информация о непараметрической компоненте. Назовем полупараметрическую оценку $\hat{\theta}$ *адаптивной*, если

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, \mathcal{I}_{\theta\theta}^{-1}\right).$$

Необходимым условием для этого является $\mathcal{I}_{\theta\alpha} = 0$ для всех параметрических подмоделей, что в общем случае не выполняется. Например, ни одна из полупараметрических оценок, рассмотренных в разделе 2, не является адаптивной. С другой стороны, очевидно, что ММП-подобные оценки, рассмотренные в разделе 3.3, адаптивны при выполнении определенных условий регулярности на распределение ошибок.

Чтобы понять, как связаны границы эффективности с асимптотическими результатами предыдущего раздела, напомним, что дисперсия полупараметрической оценки имеет вид $H_0^{-1}\Omega_0 H_0^{-1}$, где $H_0 = \mathcal{I}_{\theta\theta}$ и

$$\Omega_0 = \mathbb{E}\left[\left\{s\left(Z;\theta_0,\gamma_0\right)+\delta\left(Z\right)\right\}\left\{s\left(Z;\theta_0,\gamma_0\right)+\delta\left(Z\right)\right\}'\right].$$

Здесь $s\left(Z;\theta_0,\gamma_0\right) = \partial\log p\left(Z;\theta,\gamma\right)/\partial\theta$, где $\delta\left(Z\right)$ – добавочный член, связанный с использованием оценки γ_0 вместо ее истинного значения. Вопрос о границе эффективности полупараметрической оценки, грубо говоря, – это вопрос о поиске оценки с минимально возможным δ в терминах дисперсии. В частности, если $\delta = 0$, оценка *адаптивная,* ср. определение выше, поскольку дисперсия полностью совпадает с дисперсией в ситуации, когда нам известно γ_0.

Следует отметить, что полупараметрическая оценка, достигающая границы эффективности, может не существовать, поскольку такая граница не обязательно является точной. Примеры такого рода можно найти в Ritov & Bickel (1987), где граница эффективности полупараметрической оценки четко определена, но $\sqrt{n}$-состоятельной полупараметрической оценки не существует.

Несмотря на то, что граница эффективности имеет интуитивный смысл, в общем случае в полупараметрических задачах для нее сложно получить явное выражение. Поэтому мы не станем далее рассматривать этот вопрос. Вместо этого рассмотрим подробнее интуитивный смысл такой границы и покажем, как она меняется в зависимости от предположений исследователя о виде модели. В качестве простого примера рассмотрим следующую полупараметрическую регрессионную модель:

$$Y = m\left(X;\theta\right)+\varepsilon,$$

где условное среднее полностью параметризовано, и единственными условиями на ошибки являются ограничения $\mathbb{E}\left[\varepsilon|X\right] = 0$ и $\mathbb{E}\left[\varepsilon^2\right] < \infty$. В этом случае $\gamma = F_{\varepsilon|X}\left(e|x\right)$ – непараметрическая компонента. В зависимости от дополнительных предположений исследователя касательно $F_{\varepsilon|X}$ возникают различные границы эффективности. Например, если учитывать только ограничение на условное среднее $\mathbb{E}\left[\varepsilon|X\right] = 0$, то SEB принимает вид

$$\mathrm{SEB} = \mathbb{E}\left[\sigma^{-2}\left(X\right)\dot{m}\left(X;\theta,\gamma\right)\dot{m}\left(X;\theta,\gamma\right)'\right]^{-1},$$

где $\dot{m}\left(X;\theta,\gamma\right) = \partial m\left(X;\theta,\gamma\right)/\partial\theta$, а $\sigma^2\left(X\right) = \mathbb{E}\left[\varepsilon^2|X\right]$ – условная дисперсия. Этого можно достичь, например, используя решетчатую оценку Ai & Chen (2003).

С другой стороны, если предположить, что ε и X независимы и что ε имеет симметричное распределение, граница эффективности приобретает вид

$$\mathrm{SEB} = \mathbb{E}\left[\frac{\partial\log f_\varepsilon\left(Y-m\left(X;\theta\right)\right)}{\partial\theta}\frac{\partial\log f_\varepsilon\left(Y-m\left(X;\theta\right)\right)}{\partial\theta}\right]^{-1}.$$

В этом случае граница эффективности совпадает со значением в неравенстве Рао–Крамера. Адаптивная оценка может быть получена способом, аналогичным описанному в разделе 3.3.

8 Примечания

Оценка одноиндексной модели была предложена в работе Ichimura (1993), где также выведены ее теоретические свойства. Асимптотическая теория для оценки средних производных была получена в Powell, Stock & Stoker (1989); см. также Hristache, Juditsky & Spokoiny (2001). В случае модели бинарного выбора в качестве альтернативы можно использовать ММП-оценки β_0 в одноиндексной модели, см. Klein & Spady (1993).

Robinson (1988b) и Speckman (1988) предложили оценку, основанную на остатках частично-линейной модели, которая задается уравнением (7), и получили ее асимптотическое распределение. Andrews (1994a) получил результат для расширенной версии (10).

Robinson (1987) вывел асимптотику для ВМНК-оценки для случая гетероскедастичности неизвестного вида и показал, что эта оценка достигает границы полупараметрической эффективности. Ai & Chen (2003) предложили полупараметрические решетчатые оценки для класса полупараметрических моделей, задаваемых ограничениями на условные моменты.

Введение и описание общих результатов для копул содержится в Joe (1997). Свойства полупараметрических оценок с копулами из раздела 4 были получены в работе Genest, Ghoudi & Rivest (1995), в то время как свойства решетчатых оценок из раздела 6 проанализированы в Chen, Fan & Tsyrennikov (2006).

Для дальнейшего чтения по теме функциональных производных см. Luenberger (1969) Kantorovich & Akilov (1982).

Наши асимптотические результаты для двухшаговых оценок похожи на полученные в работах Andrews (1994a), Chen, Linton & van Keilegom (2003), Newey & McFadden (1994), Newey (1994b), Pakes & Olley (1995). В этих статьях рассматриваются общие условия состоятельности и асимптотической нормальности двухшаговых полупараметрических оценок. Что касается свойств более высоких порядков для полупараметрических оценок, сошлемся на Linton (1995, 1996).

Результаты касательно непараметрических решетчатых оценок можно найти в Andrews (1991), Fenton & Gallant (1996), Gallant & Nychka (1987), Newey (1997), Shen & Wong (1994). Про их использование в полупараметрическом оценивании см. Ai & Chen (2003), Shen (1997). Chen (2007) приводит обзор как непараметрических, так и полупараметрических методов оценивания с использованием решетчатых оценок.

Newey (1990) является хорошим введением в теорию границ эффективности полупараметрических оценок, а так же методов их получения; общий подход к расчету границ эффективности можно найти в Bickel, Klaassen, Ritov & Wellner (1993), а также в Severini & Tripathi (2001). Chamberlain (1987, 1992) выводит границы эффективности для случая ограничений на условные моменты и для полупараметрических регрессий. Manski (1984) рассматривает границы эффективности для адаптивных оценок нелинейных регрессионных моделей при наличии предположения о независимости; см. Drost & Klaassen (1997) относительно схожих результатов для случая моделей гетероскедастичных временных рядов.

Мы не обсуждаем вопросы практического применения полупараметрических оценок и отсылаем читателя к обзору Ichimura & Todd (2007). Cattaneo, Crump & Jansson (2009) детально обсуждают выбор границ для оценок средних производных, тогда как Härdle, Hall & Ichimura (1993) предлагают специальный метод для одноиндексных моделей.

Всюду в нашей работе предполагалась независимость и одинаковая распределенность данных. Большинство результатов для предложенных оценок подходят для стационарных последовательностей и последовательностей с перемешиванием; см., например, Ang & Kristensen (2009), Chen, Wu & Yu (2009), Hidalgo (1992), Kristensen (2008), а также Li & Wooldridge (2002). Вопрос границ полупараметрической эффективности для временных рядов тем не менее недостаточно хорошо исследован, если отбросить условие марковости и допустить произвольную форму зависимости; см. Bickel & Kwon (2001), а также Schick & Wefelmeyer (2005), где приводятся некоторые результаты и обсуждения.

Благодарности

Автор хотел бы поблагодарить Бруно Джиованети и Шин Канайа за полезные комментарии и предложения.

Список литературы

Анатольев, С. (2009). Непараметрическая регрессия. *Квантиль* 7, 37–52.

Расин, Дж. (2008). Непараметрическая эконометрика: вводный курс. *Квантиль* 4, 7–56.

Ai, C. & X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795–1844.

Aït-Sahalia, Y. (1993). The Delta method for nonparametric kernel functionals. Manuscript, University of Chicago.

Amemiya, T. (1985). Non-linear regression models. Глава 6 в *Handbook of Econometrics* (под редакцией M.D. Intriligator & Z. Griliches), том 1, 333–389. Elsevier Science.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–45.

Andrews, D.W.K. (1994a). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62, 43–72.

Andrews, D.W.K. (1994b). Empirical process methods in econometrics. Глава 37 в *Handbook of Econometrics* (под редакцией R. Engle & D. McFadden), том 4, 2246–2294. Elsevier Science.

Ang, A. & D. Kristensen (2009). Testing conditional factor models. CREATES Research Papers 2009–09, University of Aarhus.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press.

Bickel, P.J. & J. Kwon (2001). Inference for semiparametric models: Some questions and an answer. *Statistica Sinica* 11, 863–960.

Blundell, R., X. Chen & D. Kristensen (2007). Semi-nonparametric IV estimation of shape invariant Engel curves. *Econometrica* 75, 1613–1670.

Cattaneo, M.D., R.K. Crump & M. Jansson (2009). Small bandwidth asymptotics for density-weighted average derivatives. Manuscript, University of California–Berkeley.

Chamberlain, G. (1987). Asymptotic efficiency in estimation of conditional moment restrictions. *Journal of Econometrics* 34, 305–334.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* 60, 567–596.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Глава 76 в *Handbook of Econometrics* (под редакцией J.J. Heckman & E.E. Leamer), том 6/2. Elsevier Science.

Chen, X., Y. Fan & V. Tsyrennikov (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of American Statistical Association* 101, 1228–1240.

Chen, X., O. Linton & I. van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71, 1591–1608.

Chen, X., W.B. Wu & Y. Yu (2009). Efficient estimation of copula-based semiparametric Markov models. Cowles Foundation Discussion Papers, No. 1691.

Drost, F.C. & C.A.J. Klaassen (1997). Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* 81, 193–221.

Fenton, V.M. & A.R. Gallant (1996). Convergence rates of SNP density estimators. *Econometrica* 64, 719–727.

Gallant, A.R. & D.W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* 55, 363–390.

Genest, C., K. Ghoudi & L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 82, 543–552.

Genest, C., M. Gendron & M. Bourdeau-Brien (2009). The advent of copulas in finance. *European Journal of Finance* 15, в печати.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W., P. Hall & H. Ichimura (1993). Optimal smoothing in single-index models. *Annals of Statistics* 21, 157–178.

Härdle, W. & O. Linton (1994). Applied nonparametric methods. Глава 38 в *Handbook of Econometrics* (под редакцией R. Engle & D. McFadden), том 4. Elsevier Science.

Härdle, W, M. Müller, S. Sperlich & A. Werwatz (2004). *Nonparametric and Semiparametric Models*. New York: Springer-Verlag.

Hidalgo, J. (1992). Adaptive estimation in time series regression models with heteroskedasticity of unknown form. *Econometric Theory* 8, 161–187.

Horowitz, J. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer-Verlag.

Hristache, M., A. Juditsky & V. Spokoiny (2001). Direct estimation of the index coefficients in a single-index model. *Annals of Statistics* 29, 595–623.

Ichimura, H. (1993). Semiparametic least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

Ichimura, H. & P.E. Todd (2007). Implementing nonparametric and semiparametric estimators. Глава в *Handbook of Econometrics* (под редакцией J.J. Heckman & E.E. Leamer), том 6/2, 5369–5468. Elsevier Science.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Kantorovich, L.V. & G.P. Akilov (1982). *Functional Analysis*. Pergamon Press, Oxford.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley.

Klein, R. & R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–421.

Kristensen, D. (2008). Pseudo-maximum-likelihood estimation in two classes of semiparametric diffusion models. Manuscript, Columbia University.

Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous, dependent data. *Econometric Theory* 25, в печати.

Lee, A.J. (1990). *U-Statistics, Theory and Practice*. Marcel Dekker.

Li, Q. & J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. & J.M. Wooldridge (2002). Semiparametric estimation of partially linear models for dependent data with generated regressors. *Econometric Theory* 18, 625–645.

Linton, O.B. (1995). Second order approximation in the partially linear regression model. *Econometrica* 63, 1079–1112.

Linton, O.B. (1996). Edgeworth approximation for MINPIN estimators in semiparametric regression models. *Econometric Theory* 12, 30–60.

Manski, C. (1984). Adaptive estimation of non-linear regression. *Econometric Reviews* 3, 145–194.

Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.

Newey, W.K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica* 59, 1161–1167.

Newey, W.K. (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10, 233–253.

Newey, W.K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–1362.

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147–168.

Newey, W.K. & D.L. McFadden (1994). Large Sample Estimation and Hypothesis Testing. Глава 36 в *Handbook of Econometrics* (под редакцией R. Engle & D. McFadden), том 4, 2111–2245. Elsevier Science.

Pagan, A. & A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

Pakes, A. & S. Olley (1995). A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics* 65, 295–332.

Powell, J.L. (1994). Estimation of Semiparametric Models. Глава 41 в *Handbook of Econometrics* (под редакцией R. Engle & D. McFadden), том 4, 2443–2521. Elsevier Science.

Powell, J.L., J.H. Stock & T.M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 51, 1403–1430.

Robinson, P.M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 55, 875–891.

Robinson, P.M. (1988a). Semiparametric econometrics: A survey. *Journal of Applied Econometrics* 3, 35–51.

Robinson, P.M. (1988b). Root-N-consistent semiparametric regression. *Econometrica* 56, 931–954.

Severini, T.A. & G. Tripathi (2001). A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics* 102, 23–66.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics* 25, 2555–2591.

Shen, X. & W.H. Wong (1994). Convergence rate of sieve estimates. *Annals of Statistics* 22, 580–615.

Schick, A. & W. Wefelmeyer (2006). Efficient estimators for time series. Глава в *Frontiers in Statistics* (под редакцией J. Fan & H. L. Koul), 45–62. Imperial College Press.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society B* 50, 413–436.

van der Vaart, A & J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag.

Semiparametric modelling and estimation

Dennis Kristensen

Columbia University, New York, USA
Center for Research in Econometric Analysis of TimE Series, Aarhus, Denmark

Semiparametric models are characterized by a finite- and infinite-dimensional (functional) component. As such they allow for added flexibility over fully parametric models, and at the same time estimators of parametric components can be developed that exhibit standard parametric convergence rates. These two features have made semiparametric models and estimators increasingly popular in applied economics. We give a partial overview over the literature on semiparametric modelling and estimation with particular emphasis on semiparametric regression models. The main focus is on developing two-step semiparametric estimators and deriving their asymptotic properties. We do however also briefly discuss sieve-based estimators and semiparametric efficiency.

Задачи и решения

Задачи

Задача 7.1

Временной ряд $\{X_t\}_{t=-\infty}^{+\infty}$ со свойством $\mathbb{E}[X_t] < \infty$ называется линейным (или имеет свойство линейной регрессии), если для всех $s \geq 0$ выполнено

$$\mathbb{E}[X_{t+s}|X_{t-1}, X_{t-2}, \ldots, X_{t-k}] = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_k X_{t-k},$$

то есть если матожидания X_t, условные на конечном количестве прошлых X-ов, линейны. Покажите с помощью контрпримера, что линейность временного ряда – не то же самое, что требование линейности условного матожидания на всей предыстории $\mathbb{E}[X_{t+s}|X_{t-1}, X_{t-2}, \ldots]$.

Задача 7.2

Предположим, что

$$y_i = \alpha + \beta x_i + u_i,$$

где u_i независимо и одинаково распределены, причем $\mathbb{E}\left[u_i\right] = 0$, $\mathbb{E}\left[u_i^2\right] = \sigma^2$ и $\mathbb{E}\left[u_i^3\right] = \nu$, в то время как регрессор x_i детерминистически ужимается: $x_i = \rho^i$, где $\rho \in (0, 1)$. Пусть размер выборки равен n. Выясните асимптотическое поведение МНК-оценок $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ для $(\alpha, \beta, \sigma^2)$ по мере того как $n \to \infty$.

Задача 7.3

Рассмотрим модель

$$y = \alpha z^2 + u, \quad z = \pi x + v,$$

где

$$\mathbb{E}\left[\begin{pmatrix} u \\ v \end{pmatrix} | x\right] = 0, \quad \mathbb{V}\left[\begin{pmatrix} u \\ v \end{pmatrix} | x\right] = \Sigma,$$

причем матрица Σ неизвестна. Набор троек $\{(x_i, z_i, y_i)\}_{i=1}^n$ образует случайную выборку.

1. Рассмотрим следующий двушаговый метод оценивания. На первом шаге мы регрессируем z на x и определяем $\hat{z} = \hat{\pi} x$, где $\hat{\pi}$ – МНК-оценка. На втором шаге мы регрессируем y на $\hat{z}^2$ и получаем МНК-оценку α. Покажите, что такая оценка α несостоятельна.

2. Предложите метод состоятельного оценивания α в духе 2ШМНК.

Список литературы

Newey, W.K. & K.D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Решения

Решение 6.1

Рассмотрим линейную модель, где среди регрессоров присутствует линейный тренд:

$$y_t = \alpha + \beta t + \varepsilon_t,$$

и где ε_t независимы и одинаково распределены согласно некоторому распределению $\mathcal{D}$ с нулевым средним и дисперсией σ^2. Объектом интереса является β.

1. Выпишите МНК-оценку β (назовем ее $\hat{\beta}$) в форме отклонений и найдите ее асимптотическое распределение.
2. Исследователь предлагает избавиться от тренда в регрессорах с помощью взятия первых разностей:
$$y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}$$
с последующим оцениванием β с помощью МНК. Выпишите эту оценку (назовем ее $\check{\beta}$) и найдите ее асимптотическое распределение.
3. Сравните оценки $\hat{\beta}$ и $\check{\beta}$ по асимптотической эффективности.

1. МНК-оценка в данном случае равна

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(t - \bar{t})}{\sum_{t=1}^{T}(t - \bar{t})^2}.$$

Рассмотрим теперь разницу

$$\hat{\beta} - \beta = \left(\frac{1}{T^{-3}\sum_t t^2 - \left(T^{-2}\sum_t t\right)^2}, -\frac{T^{-2}\sum_t t}{T^{-3}\sum_t t^2 - \left(T^{-2}\sum_t t\right)^2} \right) \begin{pmatrix} T^{-3}\sum_t \varepsilon_t t \\ T^{-2}\sum_t \varepsilon_t \end{pmatrix}.$$

Поскольку

$$\sum_{t=1}^{T} t = \frac{T(T+1)}{2}, \quad \sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6},$$

легко видеть, что первый вектор сходится к $(12, -6)$. Поэтому

$$T^{3/2}(\hat{\beta} - \beta) \stackrel{A}{=} (12, -6)\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \begin{pmatrix} \frac{t}{T}\varepsilon_t \\ \varepsilon_t \end{pmatrix}.$$

Предполагая, что условия для ЦПТ для неоднородных мартингальных последовательностей (см., например, Proposition 7.8 в Hamilton, 1994) выполнены, находим, что

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \begin{pmatrix} \frac{t}{T}\varepsilon_t \\ \varepsilon_t \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \right),$$

поскольку

$$\lim \frac{1}{T} \sum_{t=1}^{T} \mathbb{V}\left[\frac{t}{T}\varepsilon_t\right] = \sigma^2 \lim \frac{1}{T} \sum_{t=1}^{T} \left(\frac{t}{T}\right)^2 = \frac{1}{3},$$

$$\lim \frac{1}{T}\sum_{t=1}^{T}\mathbb{V}\left[\varepsilon_t\right]=\sigma^2,$$

$$\lim \frac{1}{T}\sum_{t=1}^{T}\mathbb{C}\left[\frac{t}{T}\varepsilon_t,\varepsilon_t\right]=\sigma^2\lim \frac{1}{T}\sum_{t=1}^{T}\frac{t}{T}=\frac{1}{2}.$$

Следовательно,

$$T^{3/2}(\hat{\beta}-\beta)\to(12,-6)\cdot\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\sigma^2\begin{pmatrix}\frac{1}{3}&\frac{1}{2}\\\frac{1}{2}&1\end{pmatrix}\right)=\mathcal{N}(0,12\sigma^2).$$

2. Для регрессии в приращениях МНК-оценка равна

$$\check{\beta}=\frac{1}{T}\sum_{t=1}^{T}(y_t-y_{t-1})=\beta+\frac{\varepsilon_T-\varepsilon_0}{T}.$$

В результате $T(\check{\beta}-\beta)=\varepsilon_T-\varepsilon_0\sim\mathcal{D}(0,2\sigma^2)$.

3. Если T достаточно велико, $\hat{\beta}\overset{A}{\sim}\mathcal{N}\left(\beta,12\sigma^2/T^3\right)$ и $\check{\beta}\overset{A}{\sim}\mathcal{D}\left(\beta,2\sigma^2/T^2\right)$. Легко видеть, что для больших T приближенная дисперсия у первой оценки меньше, чем у второй. Гораздо легче идентифицировать тренд у растущей последовательности, чем снос у дрейфующей.

Решение 6.2

Пусть асимптотическое смещение второго порядка некоторой состоятельной асимптотически нормальной (с асимптотической дисперсией $V_{\hat{\theta}}$) оценки $\hat{\theta}$ скалярного параметра θ равно $B_{\hat{\theta}}$. Выведите асимптотическое смещение второго порядка для $g(\hat{\theta})$ как оценки $g(\theta)$, где $g(\cdot)$ – гладкая нелинейная функция.

Пусть стохастическим разложением для $\hat{\theta}$ будет

$$\hat{\theta}=\theta+\frac{A}{\sqrt{n}}+\frac{B}{n}+o_p\left(\frac{1}{n}\right),$$

где $A\sim N\left(0,V_\theta\right)$ и $\mathbb{E}\left[B/n\right]=B_{\hat{\theta}}$. Тогда стохастическим разложением для $g(\hat{\theta})$ будет

$$g(\hat{\theta})=g(\theta)+g'(\theta)\left(\frac{A}{\sqrt{n}}+\frac{B}{n}\right)+\frac{g''(\theta)}{2}\frac{A^2}{n}+o_p\left(\frac{1}{n}\right).$$

Поэтому асимптотическое смещение второго порядка для $g(\hat{\theta})$ равно

$$\mathbb{B}_2[g(\hat{\theta})]=\mathbb{E}_2[g(\hat{\theta})]-g(\theta)=\mathbb{E}\left[g'(\theta)\left(\frac{A}{\sqrt{n}}+\frac{B}{n}\right)+\frac{g''(\theta)}{2}\frac{A^2}{n}\right]=g'(\theta)B_{\hat{\theta}}+\frac{g''(\theta)V_{\hat{\theta}}}{2n}.$$

Решение 6.3

Рассмотрим регрессию в матричной форме

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{E},$$

где регрессоры $\mathcal{X}$ коррелируют с ошибками $\mathcal{E}$, но эта корреляция слаба. Рассмотрим разложение $\mathcal{E}$ на проекцию на $\mathcal{X}$ и ей ортогональную компоненту $\mathcal{U}$:

$$\mathcal{E} = \mathcal{X}\pi + \mathcal{U}.$$

Предположим, что $\left(n^{-1}\mathcal{X}'\mathcal{X}, n^{-1/2}\mathcal{X}'\mathcal{U}\right) \xrightarrow{p} (Q,\xi)$, где $\xi \sim \mathcal{N}\left(0, \sigma_u^2 Q\right)$, и матрица Q полного ранга. Покажите, что в предположении о дрейфующем параметре $\pi = c/\sqrt{n}$, где n – размер выборки, а c фиксировано, МНК-оценка для β состоятельна и асимптотически нецентрировано нормальна, и выведите асимптотическое распределение статистика Вальда для тестирования системы линейных ограничений $R\beta = r$, где R имеет полный ранг q.

МНК-оценка состоятельна:

$$\hat\beta - \beta = \left(n^{-1}\mathcal{X}'\mathcal{X}\right)^{-1} n^{-1}\mathcal{X}'\mathcal{E} = \frac{c}{\sqrt{n}} + \left(n^{-1}\mathcal{X}'\mathcal{X}\right)^{-1} n^{-1}\mathcal{X}'\mathcal{U} \xrightarrow{p} 0,$$

и асимптотически нормальна:

$$\sqrt{n}\left(\hat\beta - \beta\right) = c + \left(n^{-1}\mathcal{X}'\mathcal{X}\right)^{-1} n^{-1/2}\mathcal{X}'\mathcal{U} \xrightarrow{p} c + Q^{-1}\xi \sim N\left(c, \sigma_u^2 Q^{-1}\right).$$

Статистика Вальда имеет следующую асимптотику:

$$\mathcal{W} = \frac{n\left(R\hat\beta - r\right)'\left(R\left(\mathcal{X}'\mathcal{X}\right)^{-1}R'\right)^{-1}\left(R\hat\beta - r\right)}{\left(\mathcal{Y} - \mathcal{X}\hat\beta\right)'\left(\mathcal{Y} - \mathcal{X}\hat\beta\right)}$$

$$\xrightarrow{p} \frac{\left(c + Q^{-1}\xi\right)' R'\left(RQ^{-1}R'\right)^{-1} R\left(c + Q^{-1}\xi\right)}{\sigma_u^2} \sim \chi_q^2(\delta),$$

где δ – параметр нецентральности:

$$\delta = \frac{c'R'\left(RQ^{-1}R'\right)^{-1}Rc}{\sigma_u^2}.$$

Список литературы

Hamilton, J. (1994). *Time Series Analysis.* Princeton University Press.

Quantile

No. 7, September 2009

English page in the world wide web: http://quantile.ru/eng
Electronic mail address: quantile@quantile.ru
Access to the journal is free and unlimited

GUIDE TO AUTHORS

Manuscripts for publication in the "Articles" section should be submitted by electronic mail to the address *submit@quantile.ru*. Submitted work may be drawn from any applied field within the economics profession. The main requirement is correct usage of adequate econometric methodology. The manuscript should be written in Russian (for Russian-speaking persons) or in English (for all others) in the *Microsoft Word* or (preferably) *LaTeX* formats, and not exceed 30 double-spaced A4 pages. All submissions are subject to quality control by the editorial counsel and independent referees. A promising manuscript may be returned to the author(s) for polishing or rewriting. The editor also invites econometrics experts worldwide to contribute to the methodological sections of the journal.

Solutions to the problems from the "Problems and Solutions" section and new problems can be sent to the address *ps@quantile.ru*.

Quantile

international econometric journal in Russian language

No. 7
September 2009

IN THIS ISSUE

Econometric literacy: topics in microeconometrics

Econometric literacy: limited dependent variables

Econometric literacy: nonparametric and semiparametric methods

Problems and solutions