Эконометрический ликбез: инструментальные переменные

Инструментальные переменные и эндогенность: нетехнический обзор*

Питер Эббес†

Университет штата Пенсильвания, Юниверсити Парк, США

Настоящее эссе представляет собой нетехнический обзор наиболее свежих результатов, появившихся в эконометрической литературе по инструментальному оцениванию линейной регрессионной модели. Стандартные методы инференции, такие как МНК, дают смещенные и несостоятельные оценки, если регрессоры и ошибки коррелированны. Для преодоления этой проблемы были разработаны методы, использующие инструментальные переменные, однако поиск хороших инструментов всегда затруднителен, и зачастую исследователи-практики имеют дело со слабыми инструментами. В эссе дается обзор последних исследований, связанных со слабыми инструментами, а также рассматриваются некоторые методы, предложенные для работы с такими инструментами, включая «экономные» методы инструментальных переменных, которые не полагаются на наблюдаемые инструменты для идентификации регрессионных параметров при зависимости регрессоров и ошибок.

1 Введение

В (прикладной) статистике стандартная модель линейной регрессии $y = X\beta + \epsilon$ является важным инструментом моделирования влияния набора объясняющих переменных на зависимую переменную. Пусть $y = (y_1, ..., y_n)'$ – $n \times 1$ вектор наблюдений для зависимой переменной, $X \in \mathbb{R}^{n\times k}$ – $n \times k$ матрица объясняющих переменных (регрессоров), β – неизвестный $k \times 1$ вектор регрессионных параметров, и $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ – ненаблюдаемый случайный шок. Для идентифицируемости параметров предполагается, что $\operatorname{rank} X = k < n$. Стандартная модель линейной регрессии часто применяется в случае кросс-секционных данных. Во многих ситуациях данные имеют иерархическую структуру, например, данные о сотрудниках фирм или продажах товаров в различных магазинах, или панельные данные о потребителях. Для таких типов данных применяются многоуровневые модели, методы анализа панельных данных, или иерархические линейные модели, которые являются обобщением стандартной модели линейной регрессии (Snijders & Bosker, 1999, Wooldridge, 2002).

Важным предположением в этих моделях является некоррелированность объясняющих переменных (X) и случайной ошибки (ϵ). В этом случае регрессоры называют экзогенными; предполагается, что они определяются вне модели. Невыполнение этого предположения может привести к смещенным и несостоятельным оценкам при применении стандартных методов инференции (например, МНК или ОМНК), а следовательно, к ошибочным выводам и принятию неправильных решений. К сожалению, во многих ситуациях предпосылка о некоррелированности регрессоров и ошибок нарушается. Тогда регрессоры называют

*Перевод Б. Гершмана и С. Анатольева. Цитировать как: Эббес, Питер (2007) «Инструментальные переменные и эндогенность: нетехнический обзор», Квантиль, №2, стр. 3–20. Citation: Ebbes, Peter (2007) "A non-technical guide to instrumental variables and regressor–error dependencies," Quantile, No.2, pp. 3–20.

†Адрес: Smeal College of Business, Penn State University, University Park, PA 16802, USA. Электронная почта: *pebbes@psu.edu*

эндогенными. Зависимость регрессоров и ошибок может возникать по разным причинам: (1) пропущенные существенные переменные, (2) ошибки измерения регрессоров, (3) самоотбор, (4) одновременность, (5) серийная корреляция ошибок при наличии лагированной зависимой переменной среди регрессоров. Ruud (2000) показывает, что проблемы (2)–(5) можно рассматривать как частные случаи проблемы (1). Похожее замечание сделал Wooldridge (2002), заметивший, что различия между возможными источниками эндогенности не всегда ясны, и эти источники могут как усиливать, так и компенсировать друг друга в зависимости от конкретного практического контекста.

В настоящем эссе дается нетехническое изложение проблемы эндогенности в линейных регрессионных моделях. Кратко рассматриваются методы инструментальных переменных (ИП), имеющие долгую историю в эконометрике (Bowden & Turkington, 1984) и разработанные для преодоления этой проблемы. Инструментальные переменные – это переменные, которые коррелируют с «проблемными» регрессорам, но не коррелируют с ошибкой. Когда инструменты доступны, для оценки регрессионных параметров могут применяться двухшаговый метод наименьших квадратов (2ШМНК) или метод максимального правдоподобия с ограниченной информацией. В статье обсуждаются некоторые важные вопросы реализации методов инструментальных переменных, касающиеся, в частности, качества используемых инструментов. Методы инструментального оценивания могут потенциально давать результаты хуже, чем просто МНК при игнорировании зависимости между регрессорами и ошибками, если инструменты плохого качества, то есть являются слабыми или/и эндогенными. Эссе охватывает большинство новейших исследований по этим вопросам и рассматривает ряд методов, предложенных для работы со слабыми инструментами, а также недавно предложенные «экономные» методы инструментального оценивания, которые не основаны на наблюдаемых инструментах для учета зависимости регрессоров и ошибок. Последние используют характеристики распределения эндогенных регрессоров для построения «латентных» инструментов (Ebbes, Wedel & Böckenholt, 2006).

Структура эссе такова. В разделе 2 обсуждается ряд практических приложений линейной регрессионной модели, в которых МНК-оценка является смещенной и несостоятельной при наличии зависимости между регрессорами и ошибкой. Эти приложения мотивируют применение техники инструментальных переменных, которая обсуждается в разделе 3. В разделе 4 кратко представлены альтернативные методы, которые не основаны на использовании наблюдаемых инструментальных переменных, а используют информацию иного рода для оценки регрессионных параметров. В разделе 5 подводятся итоги.

2 Когда МНК-оценку не спасти

Важным предположением регрессионных моделей является некоррелированность регрессоров и ошибок, то есть условие $\mathbb{E}[X\epsilon] = 0$. Рассмотрим далее пять важных приложений, когда это предположение нарушено: пропущенные переменные, ошибки измерения, самоотбор, одновременность, серийная корреляции ошибок вместе с наличием лагированных значений зависимой переменной среди регрессоров.

Пропущенные существенные объясняющие переменные

Проблема пропущенных переменных широко исследуется при оценивании влияния образования на заработную плату (Card, 1999, 2001), когда «способности» являются пропущенной переменной. Индивиды с лучшими способностями, с одной стороны, имеют больший успех на рынке труда и зарабатывают больше, а с другой, более склонны к получению образования. Способности как таковые влияют как на уровень образования, так и на доходы, а регрессор «уровень образования» и регрессионная ошибка не являются независимыми. Аналогично, в

области маркетинга исследователи часто сталкиваются с проблемой пропущенных переменных. Wansbeek & Wedel (1999) утверждают, что предположение об экзогенности регрессоров, включая цену, является недостатком стандартных моделей реакции рынка. Shugan (2004) отмечает растущее внимание рецензентов к проблеме эндогенности в большинстве журналов по маркетингу. Отсутствие экзогенности регрессоров из-за пропущенных ключевых аспектов маркетинговых моделей вызывает все больший интерес при проведении исследований в области маркетинга. Когда менеджеры магазинов устанавливают показатели маркетингового комплекса (например, цену товара или объем рекламы), они основываются на информации о локальном рынке или характеристиках товаров, неизвестных исследователю, например, на знаниях об уровне конкуренции, слухах, изменениях во вкусах, долях различных игроков на локальном рынке или наличии скидок. Эта ненаблюдаемая информация может повлиять на поведение потребителей, что приводит к корреляции ошибок с регрессорами, обычно с ценой, в типичной маркетинговой модели (Villas-Boas & Winer, 1999, Chintagunta, 2001).

Модель с пропущенными переменными выглядит следующим образом (Judge, Griffiths, Hill, Lütkepohl & Lee, 1985)

$$\mathbb{E}[y_i|x_i, w_i] \quad = \quad x_i'\beta + w_i'\gamma, \tag{1}$$

где w_i's – латентные или ненаблюдаемые переменные. Когда ожидание берется условно только на наблюдаемой переменной x_i без учета w_i, имеем

$$\mathbb{E}[y_i|x_i] \quad = \quad x_i'\beta + \mathbb{E}[w_i'|x_i]\gamma, \tag{2}$$

что не равняется $x_i'\beta$, если: (i) $\mathbb{E}[w_i'|x_i] \neq 0$ (то есть пропущенные и включенные регрессоры не ортогональны) и (ii) $\gamma \neq 0$ (то есть пропущенные регрессоры оказывают влияние на y_i). Смещение МНК-оценки для параметра β в этом случае равно $\mathbb{E}[\hat{\beta}_n^{\text{OLS}} - \beta] = \Pi\gamma$, где величина и знак смещения зависят от $\Pi = (X'X)^{-1}X'W$ и γ. Как нетрудно увидеть, пропуск существенных объясняющих переменных затрагивает все оцениваемые коэффициенты вектора β.

Ошибки измерения

Зависимость регрессоров и регрессионных ошибок также возникает, если переменные, участвующие в модели, измеряются с ошибкой. Ошибки измерения могут возникать, например, когда метод или инструмент измерения дают ошибку, величина не имеет физической единицы измерения (например, IQ, способности, или восприятие), или данные из различных источников неправильно агрегируются и компонуются.

Griliches (1977), например, рассматривает проблему ошибок измерения при оценивании влияния образования на доходы. В данном случае нужна адекватная мера образования, которая отражала бы качества работников, за которые работодатели готовы платить. Обычной практикой является использование переменной «количество законченных лет обучения» в качестве меры для «общего уровня образования». Помимо ошибочного сообщения или записи величины «количество лет обучения», сомнительно, что она полностью отражает уровень образования, так как индивиды могут, например, самообразовываться на вечерних курсах или на работе. Кроме того, поскольку большинство исследований по экономике труда основываются на выборочных обследованиях домашних хозяйств, все переменные содержат некоторые ошибки. Даже если эти ошибки малы, их влияние может возрастать при добавлении дополнительных переменных в целях уменьшить смещение из-за пропущенной переменной «способности» (Card, 1999, 2001). Аналогично, Nevo (2000) и Sudhir (2001) утверждают, что мера цены, обычно используемая при оценки (логит) моделей совокупного спроса для измерения степени конкуренции, может измеряется с ошибкой. Переменная, которая используется в большинстве случаев, – это «прейскурантная цена» или агрегированная мера цены, то

есть предполагается, что все потребители сталкиваются с одной и той же ценой (и другими маркетинговыми и товарными характеристиками). Однако это предположение нарушается, если потребители совершают покупки в разных магазинах, регионах или в разные недели, и переменная «цена» может измеряться с ошибкой. В качестве альтернативы было бы лучше использовать цену сделки как меру цены (Sudhir, 2001). Bagozzi, Yi & Nassen (1999) исследуют ошибки измерения в данных, используемых в маркетинговых исследованиях. Например, пункты анкет или шкалы рейтингов, используемые для измерения восприятия, вер, отношения, суждений, или другие теоретические конструкции наверняка ведут к ошибкам измерения, поскольку нет физических единиц для точного замера соответствующих величин. Кроме того, данные для исследований в области маркетинга могут содержать ошибки метода измерения, такие как эффект ореола, эффекты интервьюирующего, или искажения из-за эффекта социальной желательности. Их результаты говорят о том, что ошибки измерения в маркетинговых данных могут быть большими и требуют учета в эмпирических приложениях для улучшения качества принимаемых решений и статистических выводов.

Рассмотрим следующую регрессионную модель с ошибками измерения:

$$y_i = \beta_0 + \beta_1 \chi_i + \epsilon_i.$$

Здесь χ_i – это истинный ненаблюдаемый компонент модели. Вместо него наблюдается x_i, $x_i = \chi_i + \nu_i$, причем $\mathbb{E}[\epsilon_i] = \mathbb{E}[\nu_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \sigma_\epsilon^2 > 0$, $\mathbb{E}[\nu_i^2] = \sigma_\nu^2 > 0$, и $\mathbb{E}[\epsilon_i \nu_i] = \mathbb{E}[\chi_i \epsilon_i] = \mathbb{E}[\chi_i \nu_i] = 0$. Эти два уравнения можно объединить в $y_i = \beta_0 + \beta_1 x_i + u_i$, где $u_i = \epsilon_i - \beta \nu_i$. МНК-оценка параметра β_1 смещена в сторону нуля, поскольку $\mathbb{E}[u_i x_i] = -\beta_1 \sigma_\nu^2 \neq 0$, так что $\mathbb{E}[u_i | x_i] \neq 0$ (см., например, Wansbeek & Meijer, 2000).

Самоотбор

Проблема самоотбора возникает, когда индивиды выбирают себе определенное состояние, например, быть или нет членом профсоюза (Vella & Verbeek, 1998), лечиться или нет (Angrist, Imbens & Rubin, 1996), на основании экономических или других, обычно неизвестных, причин. Например, Angrist (1990) рассматривает влияние статуса ветерана войны во Вьетнаме на доход граждан, чтобы понять, следует ли правительству США давать им компенсацию за возможную потерю личного дохода, вызванную службой в армии. Однако доходы непросто сравнить учитывая лишь статус ветерана, потому что индивиды с меньшими возможностями «на гражданке» скорее поступят на военную службу, и такие индивиды зарабатывали бы меньше независимо от службы в армии.

Hamilton & Nickerson (2003) дают обзор эндогенного принятия решений в стратегическом менеджменте, когда менеджеры осуществляют организационный выбор из нескольких конкурирующих стратегий не случайно, а на основании ожиданий и опыта. Аналогично данные, собранные в интернете, могут страдать от проблемы самоотбора. Определенного рода индивиды чаще бывают в сети и, следовательно, чаще принимают участие в онлайн-опросах, заходят на вебсайты или делают покупки в интернет-магазинах. Если эти ненаблюдаемые индивидуальные характеристики влияют на поведение в сети, предпочтения или восприятие, то часть влияния этих скрытых характеристик неправильно приписывается использованию интернета. Можно предположить, что эти индивиды вели бы себя иначе независимо от частоты посещения интернета. Эти явления важны, например, при исследовании решений о количестве приобретаемого товара в интернет-магазинах по сравнению с обычными («оффлайновыми») магазинами или о покупке товаров определенного брэнда в зависимости от категории товара и характеристик магазина.

Проиллюстрируем простую модель с самоотбором:

$$\begin{aligned} y_i &= x_i'(\beta + \delta) + \epsilon_i \qquad \text{если } i \in \text{I}, \\ &= x_i'\beta + \epsilon_i \qquad\quad \text{если } i \in \text{II}, \end{aligned}$$

где I и II обозначают определенные состояния (например, интернет-пользователь или нет). Более компактная запись:

$$y_i = x_i'\beta + d_i x_i'\delta + \epsilon_i,$$

где $d_i = 1$, если $i \in$ I, и $d_i = 0$ в противном случае. Из этой записи видно, что d_i является фиктивной переменной, и стандартное оценивание не проходит, если $\mathbb{E}[\epsilon_i|d_i] \neq 0$. Это предположение нарушается в приведенных выше примерах. Более детально о проблеме самоотбора можно узнать, например, из Vella (1998).

Системы одновременных уравнений

Обычный (или иерархичный) регрессионный анализ не подходит в случае, когда переменные в правой части модели определяются одновременно с зависимыми переменными. Однако часто бывает трудно избавиться от подобного взаимного влияния. Примером может быть экономический агент, принимающий решения относительно образования или участия на рынке труда (Card, 1999, 2001) или установление цен фирмами в условиях конкуренции. Некоторые исследования рассматривают одновременность цены и величины спроса на рынках с дифференцированным продуктом при данной структуре конкуренции. Ценовая политика фирм, обусловленная, например, ненаблюдаемыми характеристиками товара, такими как наличие скидок и общенациональной рекламы, расположение точек продаж и другими параметрами розничной торговли, или же реакция со стороны конкурентов, ведет к эндогенности. Работа Berry (1994) по борьбе с эндогенностью цен в агрегированных моделях с использованием инструментальных переменных широко используется и адаптируется. Например, Nevo (2001) оценивает структурную модель спроса и предложения для отрасли готовых к употреблению зерновых завтраков; Berry, Levinsohn & Pakes (1995) и Sudhir (2001) разрабатывают модель рыночного равновесия с конкурентным ценообразованием на рынке автомобилей для исследования ценообразования на автомобили и уровня конкуренции. Свежий обзор структурного моделирования в маркетинге содержится в Chintagunta, Erdem, Rossi & Wedel (2006).

Простая модель спроса и предложения для продукта или товара выглядит следующим образом:

$$\begin{aligned} y_t^d &= (x_t^d)'\beta^d + \gamma^d p_t + \epsilon_t^d, \\ y_t^s &= (x_t^s)'\beta^s + \gamma^s p_t + \epsilon_t^s, \end{aligned}$$

где компоненты вектора x_t^d – это факторы, влияющие на спрос или поведение потребителей, а компоненты x_t^s влияют только на поведения производителей. Цена p_t определяется из равенства $y_t^d = y_t^s$. Когда оценивается уравнение спроса $y_t^d = (x_t^d)'\beta + \gamma^d p_t + \epsilon_t^d$, нельзя предполагать, что $\mathbb{E}[\epsilon_t^d|p_t] = 0$, так как цена и величина спроса определяются одновременно, то есть ненаблюдаемые положительные шоки спроса или действия конкурентов сдвигают кривую спроса вверх, что (при прочих равных) означает более высокую равновесную цену. В этом случае МНК нельзя использовать для получения оценок параметров уравнения спроса.

Лагированные значения зависимой переменной среди регрессоров

Присутствие лагированных значений зависимой переменной среди регрессоров нарушает предположение об экзогенности, если имеет место серийная корреляция ошибок. Хорошо известно, что в этом случае МНК нельзя применять (см., например, White, 2001). Рассмотрим следующую модель:

$$\begin{aligned} y_t &= x_t'\beta_1 + y_{t-1}\beta_2 + \epsilon_t, \\ \epsilon_t &= \phi\epsilon_{t-1} + v_t, \end{aligned} \tag{3}$$

где y_t – это, например, объем продаж в момент t, x_t – рекламная деятельность в момент t (детерминированная переменная для простоты), и переменная y_{t-1} включена для отслеживания отложенного эффекта прошлой рекламной компании. Предположим, что $|\phi| < 1$, $|\beta_2| < 1$, v_t – независимые одинаково распределенные случайные величины, $\mathbb{E}[v_t] = 0$, и v_t не зависят от y_t и x_t, а также пусть существуют все моменты второго порядка. Преобразуем $\epsilon_t y_{t-1} = \phi\epsilon_{t-1}y_{t-1} + v_t y_{t-1}$, так что $\mathbb{E}[\epsilon_t y_{t-1}] = \phi\mathbb{E}[\epsilon_{t-1}y_{t-1}]$. Далее, $\mathbb{E}[\epsilon_t y_t] = x_t'\beta_1\mathbb{E}[\epsilon_t] + \beta_2\mathbb{E}[y_{t-1}\epsilon_t] + \mathbb{V}[\epsilon_t]$. Пользуясь стационарностью ϵ_t, получаем

$$\mathbb{E}[y_{t-1}\epsilon_t] = \frac{\phi}{1-\phi\beta_2}\mathbb{V}[\epsilon_t],$$

и $\mathbb{E}[\epsilon_t|y_{t-1}] \neq 0$, если только не выполнено $\phi = 0$. Davidson & MacKinnon (1993) делают более сильное утверждение, указывая на то, что МНК-оценка смещена во всех моделях с лагированными зависимыми переменными среди регрессоров (однако состоятельна при $\phi = 0$). В определенных случаях объясняющие переменные могут играть роль лагированных зависимых переменных, что легко упустить. Подобная ситуация описана в Gönül, Kim & Shi (2000), которые исследуют влияние продажи каталогов на вероятность покупки товаров из этих каталогов. Переменная почтовой рассылки и другие рекламные воздействия на клиентов зачастую являются функциями от прошлых продаж, что неявно создает проблемы наподобие описанных выше.

Из приведенных примеров видно, что зависимость между регрессорами и ошибками возникает в ряде стандартных приложений. Напрямую следует, что МНК-оценка

$$\hat{\beta}_n^{\text{OLS}} = \beta + (X'X)^{-1}X'\epsilon$$

смещена, если $\mathbb{E}[\epsilon|X] \neq 0$, и теряет свою привлекательность. Более того, обычная оценка дисперсии ошибок в этом случае смещена, и истинная дисперсия недооценивается (Greene, 2000). Смещение МНК-оценки не снижается при увеличении выборки, она несостоятельна, так как $\operatorname{plim}\hat{\beta}_n^{\text{OLS}} \neq \beta$ и $\operatorname{plim}\hat{\sigma}^2_{n,\text{OLS}} < \sigma^2$. Эти проблемы можно уменьшить, по крайней мере в больших выборках, при использовании инструментальных переменных (Bowden & Turkington, 1984, White, 2001). Обсудим далее метод инструментальных переменных.

3 Метод инструментальных переменных

Метод инструментальных переменных предполагает наличие набора переменных Z, называемых инструментами. Инструменты должны быть некоррелированными с ошибкой ϵ, т.е. $\mathbb{E}[\epsilon|Z] = 0$, и объяснять часть вариации эндогенных регрессоров. Следовательно, инструменты Z не должны иметь прямого влияния на y, т.е. быть экзогенными. Стандартная модель регрессии с инструментальными переменными получается добавлением к стандартной линейной регрессионной модели уравнения, связывающего эндогенные регрессоры и инструменты, а именно:

$$\begin{aligned} y &= X\beta + \epsilon, \\ X &= Z\Pi + V, \end{aligned} \tag{4}$$

где y, X, и β определяются так же, как и раньше, Z – матрица инструментальных переменных размера $n\times q$, а V – матрица ошибок размера $n\times k$. Матрица Π отражает эффект влияния инструментов на эндогенные регрессоры. Экзогенные переменные среди X также должны содержаться в наборе Z (Wooldridge, 2002). Для идентифицируемости предполагается, что $q \geq k$ и $\operatorname{rank} Z = q < n$. Корреляция между X и ϵ (эндогенность) возникает из-за ненулевой корреляции между ϵ и V. Эта модель с инструментальными переменными является

частным случаем модели с одновременными уравнениями, широко известной в эконометрике. Наиболее популярными методами оценки β являются двухшаговый МНК (2ШМНК) и метод максимального правдоподобия с ограниченной информацией (МНКОИ), который является ММП-оценкой (4) при нормально распределенных ошибках. Двухшаговый МНК используется чаще, поскольку он реализован во многих стандартных программных пакетах.

Как только набор инструментов доступен, инструментальная оценка для β в (4) вычисляется следующим образом:

$$\hat{\beta}_n^{\text{IV}} = (X' P_Z X)^{-1} X' P_Z y, \tag{5}$$

где $P_Z = Z(Z'Z)^{-1}Z'$. Она состоятельна и асимптотически нормальна, если plim $Z'\epsilon/n = 0$, и plim $Z'Z/n$ и plim $Z'X/n$ существуют и имеют полный ранг по столбцам. Несмещенность инструментальной оценки обсуждается ниже. При рассмотрении этой оценки часто используется асимптотический подход, поскольку ее математическое ожидание не существует в случае, когда число инструментов равно числу объясняющих переменных (Wooldridge, 2002). Стандартные методы инференции могут быть использованы для инференции относительно неизвестных параметров или тестирования гипотез. ММПОИ-оценка считается немного сложнее, чем 2ШМНК-оценка. Однако, если только инструменты не являются слишком слабыми, обе оценки имеют одинаковые асимптотические свойства (Davidson & MacKinnon, 1993).

Соображения по использованию инструментальных переменных

Проблемой в практических приложениях является поиск качественных инструментов. В целом, нет четких направлений поиска, и вообще инструменты нелегко найти. Кроме того, получение дополнительных данных может обойтись очень дорого. Как таковые, инструменты часто выбираются исходя из предположений или даже просто доступности, что потенциально приводит к их негодности. Условие $\mathbb{E}[\epsilon|Z] = 0$ требует отсутствия прямой связи между инструментами и зависимой переменной, что во многих практических ситуациях весьма спорно.

Wooldridge (2002), например, обсуждает годность инструмента «порядковый номер призыва в армию», используемого в Angrist (1990) для оценки влияния статуса ветерана войны во Вьетнаме на личный доход. Хотя порядковый номер призыва в армию, определяемый из лотереи, является случайным, индивиды, с наибольшей вероятностью подпадающие под призыв, могут продолжить образование для повышения шансов отсрочки от призыва, или же работодатели могут захотеть вкладывать в образование и тренинги работников, с наименьшей вероятностью подпадающих под призыв. Bound, Jaeger & Baker (1995) ставят под вопрос экзогенность инструментов, связанных с «кварталом рождения», используемых в Angrist & Krueger (1991), которые оценивают влияние образования на доход. Они предоставляют свидетельства наличия слабой корреляции между инструментом «квартал рождения» и зависимой переменной «заработная плата», не зависящей от эффекта квартала рождения на образование и достаточно сильной для смещения инструментальной оценки. Card (1999, 2001) дает обширный обзор дискуссии о годности переменных, связанных с биографией семьи (например, образования родителей), и институциональных характеристик системы образования (например, близкое расположение колледжа) как инструментов для эндогенного регрессора «образование». При оценивании спроса лагированные значения показателей рекламной деятельности часто используются в качестве инструментов в моделях рыночной реакции, но они не являются годными, когда существуют эталонные цены, формируемые исторически (см. Bronnenberg & Mahajan, 2001). Yang, Chen & Allenby (2003) замечают, что лагированные значения цены могут не быть подходящими инструментами из-за заблаговременных закупок и накопления запасов. Кроме того, использование лагированных переменных как экзогенных само по себе является потенциальным источником эндогенности (Arellano, 2002). Nevo (2001) использует данные по ценам других рынков в качестве инструментов для цены, но замечает,

что эти инструменты не являются годными, когда имеют место общие (страновые) шоки спроса, или когда рекламная или промоутерская деятельность скоординирована по рынкам. Это вероятно, если один и тот же производитель или продавец действует на нескольких рынках. Хотя носитель издержек мог бы быть потенциальным инструментом для цены, Nevo (2000) заключает, что он редко наблюдаем, а прокси-переменные для издержек обычно недостаточно вариабельны.

Экзогенность инструментов – это лишь один из двух критериев качества инструментов. Вдобавок, доступные инструменты могут быть «слабыми», в том смысле, что они слабо коррелируют с эндогенными регрессорами. Stock, Wright & Yogo (2002) утверждают: «Исследователи-практики часто сталкиваются со слабыми инструментами. Поиск экзогенных инструментов – тяжелый труд, и те свойства, которые делают инструменты экзогенными, [...] также могут делать их и слабыми». К сожалению, статистические свойства инструментальных оценок и основанная на этих оценках инференция являются чувствительными к выбору и корректности инструментов, даже в больших выборках. Следовательно, исследователи, изучающие один и тот же вопрос, но использующие разные наборы инструментов, могут прийти к различным выводам.

Arellano (2002) замечает, что «многие темы [по инструментальному оцениванию], появившиеся [...] в эконометрической литературе в 80-х и 90-х, были на удивление зрелым образом разработаны в статьях Саргана 1958 и 1959 гг.» (Sargan, 1958, 1959). Далее представлен обзор некоторых свежих результатов о слабых инструментах, появившихся в эконометрической литературе (технические детали более подробно изложены в Stock, Wright & Yogo, 2002 и Hahn & Hausman, 2003).

Слабые инструменты

Последние результаты в эконометрической литературе показали, что присутствие слабых инструментов не только снижает точность ИП-оценок, но также может привести к несостоятельности и смещению инструментальной оценки, превосходящему смещение МНК-оценки. Более того, стандартные асимптотические приближения не срабатывают (Staiger & Stock, 1997, Bound, Jaeger & Baker, 1995, Hahn & Hausman, 2002, 2003). Как следствие, на традиционное тестирование гипотез и построение доверительных интервалов нельзя полагаться. Слабые инструменты могут возникать в случаях, когда инструменты не обладают высокой объясняющей силой по отношению к эндогенным регрессорам, или когда число инструментов велико. Обсудим далее три потенциальные ловушки использования инструментальной оценки при наличии слабых инструментов: (1) смещение 2ШМНК-оценки в конечных выборках, (2) ситуации, в которых инструменты потенциально коррелируют с ϵ, и (3) плохое асимптотическое приближение фактического распределения инструментальной оценки.

В конечных выборках ИП- или 2ШМНК-оценка смещена в том же направлении, что и МНК-оценка. На этот факт часто не обращают внимание при проведении эмпирических исследований. Даже когда $\mathbb{E}[\epsilon|Z] = 0$, $\hat{\beta}_n^{\text{IV}} = \beta + (X'P_ZX)^{-1}X'P_Z\epsilon$ является в общем случае смещенной, так как $\mathbb{E}[(X'P_ZX)^{-1}X'P_Z\epsilon] \neq 0$. Смещение возникает из-за того, что коэффициенты Π в (4) ненаблюдаемы. Если бы $Z\Pi$ были наблюдаемы, МНК-регрессия y на $Z\Pi$ давала бы несмещенные результаты, но в действительности оценка Π получается из регрессии X на Z. Buse (1992) среди прочих показывает, что это смещение в конечных выборках является функцией от количества инструментов, так что увеличение их числа может увеличить смещение инструментальной оценки. Однако это смещение будет расти лишь пропорционально при росте количества инструментов более быстром, чем доля объясненной вариации в эндогенных регрессорах. Следовательно, добавление важного или сильного инструмента необязательно увеличивает смещение, но добавление менее важных инструментов или наличие слабых инструментов, несомненно, приведет к увеличению смещения. Bound,

Jaeger & Baker (1995) и Hahn & Hausman (2003) показывают, что смещение обратно связано с F-статистикой в регрессии эндогенных объясняющих переменных на инструменты. Их результаты говорят о том, что (частный) R^2 и F-статистика для регрессии первого шага (то есть регрессии X на Z) полезны как грубые показатели качества инструментальных оценок и обязательно должны сообщаться. Распределение ММПОИ-оценки в конечных выборках не имеет конечных моментов, характеризуется толстыми хвостами, а также обычно является менее чувствительной к добавлению избыточных инструментов, чем 2ШМНК-оценка. Тем не менее, когда инструменты слабые, даже эта оценка не всегда решает проблему (Hahn & Hausman, 2003, Kleibergen & Zivot, 2003).

Вторая проблема, связанная со слабыми инструментами, – это относительная несостоятельность инструментальной оценки по сравнению с МНК-оценкой в случае, когда инструменты потенциально коррелируют с ϵ, то есть сами являются эндогенными. Bound, Jaeger & Baker (1995) показывают, что относительная несостоятельность инструментальной оценки по сравнению с МНК равна (для простоты предположим, что $k = q = 1$)

$$\frac{\operatorname{plim} \hat{\beta}_n^{\mathrm{IV}} - \beta}{\operatorname{plim} \hat{\beta}_n^{\mathrm{OLS}} - \beta} = \frac{\rho_{z,\epsilon}/\rho_{x,\epsilon}}{\rho_{x,z}},$$

где $\rho_{x,z}$ – коэффициент корреляции между x и z, а другие элементы определяются аналогично. Когда инструменты слабые, $\rho_{x,z} \to 0$, то есть даже слабая корреляция между z и ϵ может давать большую относительную несостоятельность инструментальной оценки, приводя к *большей* несостоятельности последней, чем для МНК.

В-третьих, если инструменты слабые, то даже в больших выборках классические (первого порядка) асимптотические приближения неадекватны. Этот демонстрируют (в том числе) Nelson & Startz (1990), которые замечают, что при слабых инструментах классическая асимптотическая дисперсионная матрица будет больше, а асимптотическое распределение β обладает большим разбросом. Однако также показано, что асимптотическое распределение очень плохо приближает реальную плотность распределения в конечной выборке (распределение является бимодальным, с тяжелыми хвостами и сконцентрировано вокруг предела по вероятности МНК-оценки, а не вокруг истинного значения). Если асимптотическая дисперсия $\hat{\beta}_n^{\mathrm{IV}}$ ниже, то есть когда инструменты более сильные, классическое приближение лучше. Как следствие, в присутствии слабых инструментов процедуры инференции, основанные на классической асимптотике, дают неверные результаты. Хотя методы для анализа конечных выборок могут быть использованы в таких ситуациях, их использование на практике затруднено из-за ограничительных предположений, вычислительных трудностей при работе с распределениями и отсутствия четкой схемы тестирования гипотез и построения доверительных интервалов. Проблема слабых инструментов важна не только в малых выборках, ее нельзя игнорировать и в больших. Это продемонстрировано в Bound, Jaeger & Baker (1995), которые показывают, что для исследования Angrist & Krueger (1991) можно получить аналогичные результаты при использовании искусственных случайных инструментальных переменных, несмотря на 329500 наблюдений в выборке.

Проверка инструментов на пригодность

Поскольку (асимптотические) свойства инструментальной оценки чувствительны к выбору инструментов, желательно иметь какую-то меру слабости. Результаты недавних исследований предлагают обязательно сообщать R^2- и F-статистики, полученные для регрессии первого шага. Stock, Wright & Yogo (2002), например, заключают, что F-статистика для регрессии первой стадии должна быть больше 10, чтобы инференция, основанная на 2ШМНК-оценке, была достоверной. Bowden & Turkington (1984) утверждают, что следует искать инструменты, максимизирующие все канонические корреляции с X. Staiger & Stock (1997) разработали

основанную на данных меру относительного смещения, большие значения которой должны предупредить исследователя о возможных проблемах, связанных с корреляцией инструментов и ошибок. Bowden & Turkington (1984) и Verbeek (2000) (среди прочих) предлагают тест на приемлемость инструментов, когда $q > k$. Тест отвергает нулевую гипотезу, когда данные свидетельствуют против совместной годности инструментов, хотя невозможно определить, какие именно инструменты негодные. Метод Bowden & Turkington (1984) может быть использован для проверки приемлемости дополнительного набора инструментов, но этот тест ничего не говорит о возможной слабости инструментов. Hahn & Hausman (2003) утверждают, что этот тест отвергает нулевую гипотезу слишком часто в присутствии слабых инструментов, что является существенным недостатком теста, так как он часто применяется для тестирования экономической теории, заключенной в модели.

Hahn & Hausman (2002) разработали тест на пригодность инструментальных переменных, который одновременно проверяет как экзогенность, так и релевантность инструментов. Тест сравнивает 2ШМНК-оценки в прямой и обратной ИП-регрессиях, которые, как показывается, эквивалентны при нулевой гипотезе о правильности стандартной асимптотики. Тест-статистка легко считается и имеет t-распределение при нулевой гипотезе. Отвержение нулевой гипотезы может означать либо нарушение предположения об экзогенности инструментов, либо их слабость. Hahn & Hausman (2002) предлагают двухшаговый метод, основанный на этом тесте, для принятия решения об использовании либо 2ШМНК-оценки, либо ММПОИ-оценки, либо ни одной из них. Ebbes (2004) разработал другой метод для проверки приемлемости наблюдаемых инструментов, который основан на методе латентных инструментальных переменных (Ebbes, Wedel, Böckenholt & Steerneman, 2005). В отличие от теста Hahn & Hausman (2002), он может быть использован для проверки инструментов на слабость и эндогенность как по отдельности, так и одновременно. Более того, даже если инструменты непригодны, оценки регрессионных параметров тем не менее можно использовать, так как метод латентных инструментальных переменных не основан ни на качестве, ни на доступности наблюдаемых инструментов. Симуляции показывают, что этот метод не ведет к проблемам с размером теста в присутствии слабых инструментов в отличие от классического теста на сверхидентифицирующие ограничения.

Выбор количества инструментов

Смещение инструментальной оценки в конечных выборках является функцией от количества инструментов, откуда следует, что не стоит использовать их слишком много, хотя условие идентифицируемости требует их по крайней мере столько же, сколько имеется регрессоров ($q \geq k$). Более того, увеличение количества инструментов ведет к потере степеней свободы, и регрессия первого шага (X на Z) страдает от чрезмерной подгонки. Sargan (1958) заключает, что «если первые несколько инструментов хорошо подобраны, обычно наблюдается не улучшение, а наоборот, ухудшение доверительных интервалов по мере роста числа инструментов свыше трех или четырех». Кроме того, он отмечает, что «оценки могут быть сильно смещены, если инструментов становится слишком много». В отличие от этих конечновыборочных результатов, асимптотическая теория говорит о том, что инструментальная оценка с одним дополнительным инструментом по крайней мере настолько же эффективна, то есть инструменты можно добавлять без ухудшений (Davidson & MacKinnon, 1993).

Bowden & Turkington (1984) предлагают применять метод главных компонент для снижения размерности матрицы $Z'Z$ и выбора первых p главных компонент в качестве инструментов. Однако этот подход не учитывает корреляцию между X и Z, то есть силу инструментов. Donald & Newey (2001) разработали критерий среднеквадратичной ошибки, который следует минимизировать выбором инструментов. Авторы обнаружили, что их метод выбора инструментов обычно дает улучшение качества оценки. В большинстве случаев ММПОИ-оценка лучше, чем 2ШМНК-оценка, хотя последняя дает лучший результат, когда степень эндо-

генности низкая. В случае слабых инструментов явно следует использовать меньшее число инструментов.

Тестирование на зависимость между регрессорами и ошибками

Учитывая возможные опасности при использовании инструментальной оценки и проблему поиска инструментов, желательно для начала провести тест на коррелированность регрессоров с ошибками. К сожалению, невозможно напрямую рассматривать $X'\epsilon$, так как ϵ не наблюдается, а МНК-оценивание дает $X'\hat{\epsilon} = 0$ по построению. Исключением является тестирование гипотезы $X'\alpha = 0$ в многоуровневых моделях со случайными эффектами, где $\alpha = (\alpha_1, ..., \alpha_n)'$ являются случайными эффектами, так как имеется готовый статистический тест (см., например, Ebbes, Böckenholt & Wedel, 2004). Для тестирования на эндогенность в стандартной регрессионной модели, однако, необходимы инструменты хорошего качества. Тогда можно применить тест на основе общей процедуры Хаусмана (Hausman, 1978), который основан на разности между $\hat{\beta}_n^{\mathrm{OLS}}$ и $\hat{\beta}_n^{\mathrm{IV}}$. Хаусман предложил тестовую статистику, имеющую асимптотическое χ^2-распределение при нулевой гипотезе о некоррелированности регрессоров и ошибок. Недостаток этой процедуры в том, что для расчета $\hat{\beta}_n^{\mathrm{IV}}$ необходимы инструменты, хотя впоследствии исследователь может заключить, что они не нужны. Более того, этот тест чувствителен к слабым инструментам (Staiger & Stock, 1997). Действительно, тест Хаусмана может неправильно принять гипотезу о возможности использования МНК-оценки, так как она смещена при наличии слабых инструментов (Hahn & Hausman, 2003). Ebbes, Wedel, Böckenholt & Steerneman (2005) предлагают тест без использования инструментов для проверки на коррелированность регрессоров и ошибок в линейной регрессионной модели с одним эндогенным регрессором и нормальными ошибками и показывают, что этот тест обладает достаточной мощностью для широкого класса условий.

Решения проблемы слабых инструментов, основанные на инструментальных переменных

Hahn & Hausman (2003) и Stock, Wright & Yogo (2002) дают обзор большей части эконометрических исследований о способах решения проблемы слабых инструментов в практических приложениях. Далее следует краткое изложение этих способов.

Хорошо известно, что асимптотическое приближение первого порядка для распределения инструментальной оценки является плохим при наличии слабых инструментов. Некоторые исследователи предлагают улучшенные асимптотические приближения для распределения оценки в конечных выборках в такой ситуации. Staiger & Stock (1997) разработали альтернативную асимптотику, которая моделирует коэффициенты регрессии первого шага как локально нулевые, то есть слабо коррелированные, без предположения о нормальности. В рамках этой асимптотики они показали, что если инструменты являются слабыми, то 2ШМНК- и ММПОИ-оценки несостоятельны и имеют нестандартные асимптотические распределения, причем смещение ММПОИ-оценки меньше такового у 2ШМНК-оценки, особенно в малых выборках. Более того, авторы выводят свойства процедур инференции (t-тестирование, нормы покрытия доверительных интервалов и тестирование на сверхидентифицирующие ограничения). Bekker (1994) разработал альтернативную асимптотику для моделей с нормальными ошибками, в которых растут и размер выборки, и количество инструментов. Симуляции показывают, что эта асимптотика дает лучшее приближение для среднего и большого числа инструментов, и что ММПОИ-оценке следует отдавать предпочтение перед обычной инструментальной оценкой. Однако результаты Беккера применимы лишь к случаю с нормальными ошибками и не улавливают ненормальность, наблюдаемую в точном распределении в конечной выборке, когда присутствуют слабые инструменты (Staiger & Stock, 1997).

Помимо работы по поиску лучших альтернатив обычной асимптотике первого порядка, разработаны «полностью робастные» тесты для проверки гипотез и методы построения доверительных интервалов для β с приблизительно правильным размером и нормой покрытия при наличии слабых инструментов. Один из таких робастных тестов на проверку гипотезы $\beta = \beta_0$ – это статистика Андерсона–Рубина (Anderson & Rubin, 1949), которая не зависит от степени недоидентификации. Однако, ей может не хватать мощности из-за потери степеней свободы при увеличении количества инструментов. K-статистика (Kleibergen, 2002) имеет схожие асимптотические свойства с минимальным числом степеней свободы. Bekker & Kleibergen (2003) исследуют ее распределение в конечных выборках при нормальности ошибок. Предлагались также и другие тесты, см., например, Staiger & Stock (1997). Stock, Wright & Yogo (2002) сравнивают мощности нескольких тестов при разных условиях. Учитывая двойственность проверки гипотез и построения доверительных интервалов, робастные тесты могут быть использованы и для получения доверительных интервалов. Когда инструменты слабые, эти интервалы могут быть неограничены, указывая на очень ограниченное количество информации для проведения инференции о β.

Указанные методы проверки гипотез и построения доверительных интервалов не дают точечной оценки β. Кроме того, они могут быть вычислительно сложными. Были предложены некоторые альтернативы 2ШМНК-оценке, которые являются более робастными и потенциально более надежными, если инструменты слабые. Оценки с нулевым смещением второго порядка, такие как ММПОИ-оценка, часто предлагаются в качестве робастных альтернатив. Эти оценки, однако, не имеют моментов в конечных выборках, что может стать проблемой при практической реализации (Hahn & Hausman, 2003). Другими альтернативами являются инструментальные переменные «складного ножа» (Angrist, Imbens & Krueger, 1999), k-оценка Фуллера (Fuller, 1977) или подправленная на смещение 2ШМНК-оценка (Donald & Newey, 2001). Stock, Wright & Yogo (2002) находят, что эти частично робастные оценки являются более надежной альтернативой 2ШМНК-оценке в случаях со слабыми инструментами. Тем не менее, Hahn & Hausman (2003) рекомендуют быть очень острожными при применении оценок, распределение которых не имеет моментов (например, ММПОИ-оценки). Они обнаруживают, что 2ШМНК-оценка, 2ШМНК-оценка «складного ножа» и k-оценка Фуллера ведут себя лучше, и утверждают, что «пессимизм по поводу инструментов преувеличен для 2ШМНК-оценки, поэтому, возможно, 2ШМНК-оценка часто ведет себя лучше, чем ожидается, в смысле среднеквадратической ошибки в ситуации со слабыми инструментами». Тест на спецификацию, разработанный в Hahn & Hausman (2002) можно использовать для выбора из альтернативных инструментальных оценок. Как Stock, Wright & Yogo (2002), так и Hahn & Hausman (2003) подчеркивают, что большая часть литературы по слабым инструментам предполагает экзогенность инструментов. Невыполнение этого условия, особенно вкупе со слабыми инструментами, ведет к дополнительным сложностям, и в этом случае МНК может давать наилучшие результаты по сравнению с предложенными выше способами учета слабых инструментов.

4 Альтернативные подходы к решению проблемы эндогенности

В некоторых приложениях сама природа процесса, порождающего данные, или же сама причина эндогенности подразумевают подходящие инструменты или даже другой подход к оцениванию. Wooldridge (2002) предлагает три альтернативных способа решения проблемы пропущенных переменных, включая МНК с прокси-переменными и использование индикаторов ненаблюдаемых переменных. Последний метод требует инструментального оценивания, но отличается от классического метода ИП. Индикаторный подход предполагает существование возможно неправильно измеренной прокси-переменной для пропущенной переменной w, которая нуждается в инструменте, в то время как классический метод инструментальных

переменных оставляет пропущенную переменную w внутри ошибки, и все элементы x, коррелированные с w, требуют инструментов. Когда одни и те же кросс-секционные объекты наблюдаемы во времени, инструментальные переменные не нужны, так как можно применить оценку с фиксированными эффектами для учета пропущенных переменных, если корреляция между регрессорами и ошибкой возникает из независимых от времени источников (см., например, Ebbes, Böckenholt & Wedel, 2004). Аналогично, Card (1999) предлагает обзор исследований, использующих данные по братьям и близнецам для оценки отдачи от образования, и утверждает, что проблема пропущенной переменной «способности» устраняется при оценивании внутрисемейных данных. Для моделей с ошибками измерения, авторегрессионных моделей и систем одновременных уравнений процесс, порождающий данные, может предложить подходящие инструменты. Например, модели с ошибками измерения можно оценивать, используя инструментальные переменные, когда инструменты измеряются с ошибкой (Wansbeek & Meijer, 2000, White, 2001). Другой метод оценивания моделей с ошибками измерения, предложенный в Wald (1940), использует информацию по наблюдаемой переменной группировки. Эта группировка должна быть независима от ошибок и различать высокие и низкие значения ненаблюдаемой истинной переменной. Аналогично, в авторегрессионных моделях лаги более высокого порядка могут служить инструментами для модели с лагированными зависимыми переменными в качестве регрессоров. В системах одновременных уравнений экзогенные переменные, не включенные в уравнение, часто могут служить инструментами и легко доступны (Greene, 2000).

Далее мы рассмотрим три других подхода к проблеме зависимости между регрессорами и ошибками: (1) «экономные» методы инструментальных переменных, (2) методы для моделирования спроса, издержек и конкуренции и (3) пространственную эконометрику.

«Экономные» методы инструментальных переменных

Ebbes, Wedel & Böckenholt (2006) рассматривают три метода оценивания, которые не основываются на наблюдаемых инструментальных переменных для идентификации регрессионных параметров в моделях с эндогенностью: подход «моментов высокого порядка» (HM, *higher moments*) (Erickson & Whited, 2002, Lewbel, 1997), метод «идентификации через гетероскедастичность» (IH, *identification through heteroskedasticity*) (Rigobon, 2003 и Hogan & Rigobon, 2003) и метод «латентных инструментальных переменных» (LIV, *latent instrumental variables*) (Ebbes, Wedel, Böckenholt & Steerneman, 2005). Учитывая проблемы классического инструментального оценивания, в частности проблему слабых инструментов и малую доступность наблюдаемых инструментов, «экономные» методы инструментальных переменных могут быть полезной альтернативой.

В HM-подходе инструменты строятся по имеющимся данным, используя моменты высоких порядков. Состоятельное оценивание требует (среди прочего), чтобы ошибки измерения и ошибки в структурных уравнениях были независимы и имели моменты всех порядков, однако дальнейших предположений о формах распределения не делается. IH-оценка также основана на стратегии идентификации с использованием моментов более высокого порядка и предполагает, что доступна некоторая информация о гетероскедастичности ошибок. Для подсчета IH-оценки требуется наблюдаемая переменная группировки, которая описывает структуру гетероскедастичности в ошибках. LIV-подход аппроксимирует ненаблюдаемые инструменты латентной дискретной переменной. LIV-модель принадлежит классу моделей со смесью нормальных распределений, когда ошибки в регрессионной модели нормально распределены. В отличие от HM- и IH-оценок, этот подход основан на правдоподобии, а идентифицируемость может быть доказана из свойств моделей со смесями вероятностных распределений. Оценки максимального правдоподобия для регрессионных параметров LIV-модели можно получить, используя только зависимые и независимые переменные. Ebbes, Wedel & Böckenholt (2006) докладывают о симуляциях, которые демонстрируют «за» и «против» этих трех методов

при различных условиях. Вывод состоит в том, что при разумном применении эти «экономные» методы ИП могут быть полезными альтернативами в ситуациях, когда есть эндогенный регрессор, но недоступны наблюдаемые инструменты хорошего качества.

Методы моделирования спроса, издержек и конкуренции

Ряд исследований рассматривают эндогенность цены на рынках дифференцированных продуктов. Цена эндогенно определяется спросом и предложением. Berry (1994) и Berry, Levinsohn & Pakes (1995) разработали модель рыночного равновесия с логистической функцией спроса, которая (модель) адаптируется для возможности применения традиционного метода инструментальных переменных. Этот метод применим как для агрегированных, так и для дезагрегированных данных или их комбинации. Итоговая система получается путем объединения модели дискретного выбора для потребительского поведения индивида и функции издержек. Эти две модели заключены в систему установления цен фирмами на рынках дифференцированных товаров. Совместное оценивание ведет к потенциально более эффективным оценкам, чем при использовании только модели спроса, где берется инструмент для цены. Более того, полная система дает детальную информацию о структуре издержек и природе конкуренции. Такие равновесные модели, однако, накладывают больше требований к данным, и некорректная спецификация поведения фирм может привести к смещенным оценкам. Данный подход широко применялся и модифицировался, см., например, Chintagunta, Erdem, Rossi & Wedel (2006), где дан свежий обзор.

Пространственная эконометрика

Недавно опубликованы два исследования в области маркетинга, которые решают проблему эндогенности переменных маркетингового комплекса, используя пространственные зависимости в наблюдаемых на рынке данных, постоянные или несильно варьирующиеся во времени. Эти зависимости вызваны тем, что экономические агенты организованы в пространстве или имеют схожее расположение магазинов. Bronnenberg & Mahajan (2001) идентифицируют корреляцию между переменными маркетингового комплекса и ошибкой, накладывая измеримую пространственную структуру на случайные ошибки в модели. Пространственная карта возникает из ненаблюдаемых действий розничных продавцов в различных местах торговли с несколькими соседствующими рынками. Bronnenberg & Mahajan строят пространственную карту, исходя из географической близости. Можно оценить и протестировать эффект ненаблюдаемого поведения продавцов, включая карту в модель. Результаты авторов применительно к анализу рынка мексиканской еды говорят о том, что ненаблюдаемая компонента в зависимой переменной связана с переменными маркетингового комплекса. van Dijk, van Heerde, Leeflang & Wittink (2004) рассматривают оценку эластичности площадей витрин, отведенных под разные товары, на основе эндогенных данных о площадях витрин для выкладки товаров. Оценивание этих эластичностей затруднено из-за минимальной вариации (во времени) показателей площадях витрин, отведенных под разные товары. Авторы, основываясь на работе Bronnenberg & Mahajan (2001), предлагают моделировать корреляцию между площадью витрин и случайными ошибками, используя пространственную структуру, основанную на схожести характеристик магазинов, потребителей или конкурентов. Их результаты применительно к часто используемым товарам ежедневной гигиены дают оценки эластичностей, превосходящие по качеству оценки из модели с пространственной структурой на основе географической близости в терминах правильности прогнозов. Поскольку продавцы обычно принимают решение о распределении витрин по разным товарам на основе характеристик магазинов, потребителей или конкурентов, ожидается, что схожесть двух географически близких магазинов ниже, чем схожесть двух магазинов с одинаковым профилем в разных регионах.

5 Заключение

Методы инструментальных переменных разработаны для преодоления проблемы зависимости между регрессорами и ошибками. Из настоящего обзора ясно, однако, что поиск подходящих инструментов в каждой конкретной ситуации может быть проблематичным. Традиционные методы инструментальных переменных полагаются на экономическую теорию или интуицию в отыскании инструментов. К сожалению, зачастую инструменты хорошего качества просто недоступны, и инструменты могут быть слабыми или/и эндогенными. Хорошо известно, что качество процедур инференции на основе инструментальной оценки напрямую зависит от качества используемых инструментов. Это явилось темой некоторых недавних исследований в эконометрике, и был сделан ряд предложений для улучшения инференции при наличии слабых инструментов. В целом, меньшее количество инструментов предпочитается большему. Более того, R^2- или F-статистика в регрессии первого шага должны быть всегда рассчитаны и сообщаться как меры силы инструментов. Большинство результатов по проблеме слабых инструментов выводятся при условии экзогенности инструментов, и вопрос о том, что делать в ситуациях с эндогенными инструментами, остается открытым.

Для исследователей-практиков поиск подходящих инструментов затруднителен, а выбор обычно небольшой. Альтернативные «экономные» методы инструментальных переменных, описанные выше, пытаются идентифицировать регрессионные параметры не через наблюдаемые инструменты, а путем использования характеристик распределений эндогенных регрессоров, и могут считаться альтернативой или дополнением к классическому инструментальному оцениванию. Без наличия инструментов хорошего качества техника классического инструментального оценивания ненадежна. Теория, с одной стороны, говорит о том, что наилучшими инструментами являются переменные, сильно коррелированные с эндогенными регрессорами. С другой стороны, однако, чем больше они коррелированны, тем меньше надежда на то, что сами эти инструменты не коррелируют с ошибками (см. Greene, 2000).

Благодарности

Данное эссе основано на второй главе моей диссертации (Ebbes, 2004). Я бы хотел поблагодарить моих научных руководителей, профессоров Михеля Веделя, Ульфа Бёкенхольта и Тона Штернемана за ценные комментарии и предложения по ранним версиям эссе.

Список литературы

Anderson, T.W. & Rubin, H. (1949). Estimators on the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics* 21, 570–582.

Angrist, J.D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* 80, 313–336.

Angrist, J.D., G.W. Imbens & A.B. Krueger (1999). Jacknife instrumental variables estimation. *Journal of Applied Econometrics* 14, 57–67.

Angrist, J.D., G.W. Imbens & D.B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, 444–455.

Angrist, J.D. & A.B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 56, 979–1014.

Arellano, M. (2002). Sargan's instrumental variables estimation and the generalized method of moments. *Journal of Business & Economic Statistics* 20, 450–459.

Bagozzi, R.P., Y. Yi & K.D. Nassen (1999). Representation of measurement error in marketing variables: Review of approaches and extension to three-faced designs. *Journal of Econometrics* 89, 393–421.

Bekker, P.A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, 657–681.

Bekker, P.A. & F. Kleibergen (2003). Finite-sample instrumental variables inference using an asymptotic pivotal statistic. *Econometric Theory* 19, 744–753.

Berry, S., J. Levinsohn & A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica* 63, 841–890.

Berry, S.T. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics* 25, 242–262.

Bound, J., D.A. Jaeger & R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–450.

Bowden, R.J. & D.A. Turkington (1984). Instrumental Variables. New York: Cambridge University Press.

Bronnenberg, B.J. & V. Mahajan (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science* 20, 284–299.

Buse, A. (1992). The bias of instrumental variables estimators. *Econometrica* 60, 173–180.

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics* 3A, 1801–1863. Amsterdam: Elsevier.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69, 1127–1160.

Chintagunta, P., T. Erdem, P. Rossi & M. Wedel (2006). Structural modeling in marketing: Review and assessment. *Marketing Science*, в печати.

Chintagunta, P.K. (2001). Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data. *Marketing Science* 20, 442–456.

Davidson, R. & J.G. MacKinnon (1993). Estimation and Inference in Econometrics. New York: Oxford University Press.

van Dijk, A., H. J. van Heerde, P.S.H. Leeflang, & D.R. Wittink (2004). Similarity-based spatial methods for estimating shelf space elasticities from correlational data. *Quantitative Marketing and Economics* 2, 257–277.

Donald, S.G. & W.K. Newey (2001). Choosing the number of instruments. *Econometrica* 69, 1161–1191.

Ebbes, P. (2004). Latent Instrumental Variables: A New Approach to Solve for Endogeneity. PhD thesis, University of Groningen.

Ebbes, P., U. Böckenholt & M. Wedel (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica* 58, 161–178.

Ebbes, P., M. Wedel & U. Böckenholt (2006). Frugal IV alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics*, в печати.

Ebbes, P., M. Wedel, U. Böckenholt & A.G.M. Steerneman (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics* 3, 365–392.

Erickson, T. & T.M. Whited (2002). Two-step GMM estimation of the errors-in-variables model using high-order moments. *Econometric Theory* 18, 776–799.

Fuller, W. (1977). Some properties of a modification of the limited information estimator. *Econometrica* 45, 939–953.

Gönül, F.F., B.-D. Kim & M. Shi (2000). Mailing smarter to catalog customers. *Journal of Interactive Marketing* 14, 2–16.

Greene, W.H. (2000). Econometric Analysis. New Jersey: Prentice-Hall.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica* 45, 1–22.

Hahn, J. & J. Hausman (2002). A new specification test for the validity of instrumental variables. *Econometrica* 70, 163–189.

Hahn, J. & J. Hausman (2003). Weak instrumens: Diagnosis and cures in empirical econometrics. *Recent Advances in Econometric Methodology* 93, 118–125.

Hamilton, B.H. & J.A. Nickerson (2003). Correcting for endogeneity in strategic management research. *Strategic Organization* 1, 51–78.

Hausman, J.A. (1978). Specification tests for econometrics. *Econometrica* 46, 1251–1271.

Hogan, V. & R. Rigobon (2003). Using unobserved supply shocks to estimate the returns to educations. Technical report, University College Dublin.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl & T.-C. Lee (1985). The Theory and Practice of Econometrics. New York: John Wiley & Sons.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70, 1781–1803.

Kleibergen, F. & E. Zivot (2003). Bayesian and classical approaches to instrumental variables regression. *Journal of Econometrics* 114, 29–72.

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. *Econometrica* 65, 1201–1213.

Nelson, C.R. & R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967–976.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy* 9, 513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69, 307–342.

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics* 85, 777–792.

Ruud, P.A. (2000). An Introduction to Classical Econometric Theory. New York: Oxford University Press.

Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

Sargan, J.D. (1959). The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society, Series B* 21, 91–105.

Shugan, S.M. (2004). Endogeneity in marketing decision models. *Marketing Science* 23, 1–3.

Snijders, T.A.B. & R.J. Bosker (1999). Multilevel Analysis. London: SAGE Publications.

Staiger, D. & J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J.H., J.H. Wright & M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20, 518–529.

Sudhir, K. (2001). Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science* 20, 42–60.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33, 127–169.

Vella, F. & M. Verbeek (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics* 13, 163–183.

Verbeek, M. (2000). A Guide to Modern Econometrics. Chichester: John Wiley & Sons.

Villas-Boas, J.M. & R.S. Winer (1999). Endogeneity in brand choice models. *Management Science* 45, 1324–1338.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 11, 284–300.

Wansbeek, T. & E. Meijer (2000). Measurement Error and Latent Variables in Econometrics. Amsterdam: Elsevier.

Wansbeek, T. & M. Wedel (1999). Marketing and econometrics: editors' introduction. *Journal of Econometrics* 89, 1–14.

White, H. (2001). Asymptotic Theory for Econometricians. New York: Academic Press.

Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT Press.

Yang, S., Y. Chen & G.M. Allenby (2003). Bayesian analysis of simultaneous demand and supply. *Quantitative Marketing and Economics* 1, 251–275.

A non-technical guide to instrumental variables and regressor–error dependencies

Peter Ebbes

Penn State University, University Park, USA

We provide a non-technical summary of most of the recent results that have appeared in the econometric literature on instrumental variables estimation for the linear regression model. Standard inferential methods, such as OLS, are biased and inconsistent when the regressors are correlated with the error term. Instrumental variables methods were developed to overcome this problem, but finding instruments of good quality is cumbersome in any given situation and empirical researchers are often confronted with weak instruments. We review most of the recent studies on weak instruments and point to several methods that have been proposed to deal with such instruments, including "frugal" IV alternatives that do not rely on observed instruments to identify the regression parameters in presence of regressor–error dependencies.

Экскурс в мир инструментальных переменных*

Александр Цыплаков†

Новосибирский государственный университет, Новосибирск

Настоящее эссе посвящено причинам, по которым возникает корреляция объясняющих переменных и ошибки в приложениях регрессионного анализа, последствиям такой корреляции и методу, который призван помочь решить данную проблему, – методу инструментальных переменных.

1 Введение

Основная теоретическая модель, вокруг которой строится данное эссе – это самая обычная множественная линейная регрессия, которая широко используется в эконометрике:

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon. \tag{1}$$

Здесь y – объясняемая («зависимая») переменная, $\mathbf{x}$ – вектор-строка m объясняющих переменных (или «регрессоров»),[1] $\boldsymbol{\beta}$ – коэффициенты регрессии, а ε – случайное возмущение (ошибка). Предполагается, что у нас имеются данные, по которым мы оцениваем неизвестные коэффициенты $\boldsymbol{\beta}$. Это пары $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, подчиняющиеся указанному уравнению:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i,$$

или в матричном виде

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

В курсах эконометрики для начинающих чаще всего исходят из того, что регрессоры $\mathbf{x}$ фиксированы (неслучайные). Это позволяет существенно упростить рассуждения и сделать изложение одновременно формально корректным и технически несложным. Однако такое предположение, очевидно, нереалистично и не отражает многие важные аспекты моделирования при помощи регрессионного анализа. В частности, если некоторые регрессоры случайные, то вполне может быть, что они в вероятностном смысле связаны с ошибкой, что может приводить к различным сложностям.

Пусть, например, $(y_i, \mathbf{x}_i) \sim IID$, т. е. независимы и одинаково распределены в соответствии с некоторым совместным распределением. Пусть, кроме того, существуют первые и вторые моменты совместного распределения ε_i и $\mathbf{x}_i$. Согласно закону больших чисел, при стремлении количества наблюдений к бесконечности $\mathbf{X}^\intercal\mathbf{X}/n \xrightarrow{p} \mathbf{Q}_{xx}$ и $\mathbf{X}^\intercal\boldsymbol{\varepsilon}/n \xrightarrow{p} \mathbf{Q}_{x\varepsilon}$, где $\mathbf{Q}_{xx} = \mathbb{E}[\mathbf{x}^\intercal\mathbf{x}]$ и $\mathbf{Q}_{x\varepsilon} = \mathbb{E}[\mathbf{x}^\intercal\boldsymbol{\varepsilon}]$. При этом для обычной МНК-оценки

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$$

выполнено

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \xrightarrow{p} \boldsymbol{\beta} + \mathbf{Q}_{xx}^{-1}\mathbf{Q}_{x\varepsilon}.$$

*Цитировать как: Цыплаков, Александр (2007) «Экскурс в мир инструментальных переменных», Квантиль, №2, стр. 21–47. Citation: Tsyplakov, Alexander (2007) "A guide to the world of instrumental variables," Quantile, No.2, pp. 21–47.

†Адрес: 630090, г. Новосибирск, Весенний проезд, 6–44. Электронная почта: *tsy@academ.org*

[1] При дальнейшем обсуждении окажется, что уравнение (1) не является регрессией в строгом понимании этого понятия. Тем не менее мы будем ссылаться на (1) как на регрессионное уравнение, а на объясняющие переменные – как на регрессоры.

(Предполагается, что матрица $\mathbf{Q}_{xx}$ невырождена.) Оценки $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ состоятельны[2] в смысле сходимости по вероятности тогда и только тогда, когда $\mathbf{Q}_{x\varepsilon} = \mathbf{0}$, т. е. когда регрессоры и ошибки некоррелированы между собой. Если же ε_i и $\mathbf{x}_i$ коррелированы между собой, то свойство состоятельности оценок теряется.

Ситуацию, когда $\mathbf{Q}_{x\varepsilon} \neq \mathbf{0}$, т. е. когда регрессоры и ошибка коррелированы между собой будем для краткости называть – не совсем формально – *эндогенностью регрессоров*.[3]

Чтобы понять, куда смещаются оценки в случае коррелированности регрессоров и ошибки, удобно ввести понятие линейной проекции. Это понятие нам еще понадобится в дальнейшем. Пусть $\mathbf{w}$ и $\mathbf{v}$ – две векторные случайные величины. Коэффициенты $\mathbf{A}$ линейной проекции $\mathbf{w}$ на $\mathbf{v}$ должны удовлетворять следующим уравнениям (являющимся аналогом нормальных уравнений):

$$\mathbb{E}[\mathbf{v}^\intercal\mathbf{v}]\mathbf{A} = \mathbb{E}[\mathbf{v}^\intercal\mathbf{w}].$$

Если решение единственное, то оно находится как[4]

$$\mathbf{A} = \mathbb{E}[\mathbf{v}^\intercal\mathbf{v}]^{-1}\mathbb{E}[\mathbf{v}^\intercal\mathbf{w}].$$

Определим оператор проекции $\mathcal{P}$ таким образом, что проекция $\mathbf{w}$ на пространство, натянутое на $\mathbf{v}$, равна

$$\mathcal{P}(\mathbf{w}|\mathbf{v}) = \mathbf{v}\mathbf{A}.$$

При этом оказывается верным следующее представление (линейная проекция):

$$\mathbf{w} = \mathbf{v}\mathbf{A} + \mathbf{u},$$

где $\mathbf{u} = \mathbf{w} - \mathcal{P}(\mathbf{w}|\mathbf{v})$ – ошибка линейной проекции, такая что $\mathcal{P}(\mathbf{u}|\mathbf{v}) = \mathbf{0}$. Эта ошибка некоррелирована с регрессорами $\mathbf{v}$, т. е. $\mathbb{E}[\mathbf{u}^\intercal\mathbf{v}] = \mathbf{0}$, и имеет нулевое математическое ожидание ($\mathbb{E}[\mathbf{u}] = \mathbf{0}$), если в $\mathbf{v}$ присутствует константа.

Рассмотрим линейную регрессию (1), где в $\mathbf{x}$ присутствует константа, а ошибка в среднем равна нулю ($\mathbb{E}[\varepsilon] = 0$). Построим линейную проекцию ε на $\mathbf{x}$: $\mathcal{P}(\varepsilon|\mathbf{x}) = \mathbf{x}\boldsymbol{\delta}$. Если регрессоры и ошибка некоррелированы между собой, то $\boldsymbol{\delta} = \mathbf{0}$. В противном случае $\boldsymbol{\delta} \neq \mathbf{0}$, и фактически вместо указанной регрессии мы имеем

$$y = \mathbf{x}\tilde{\boldsymbol{\beta}} + \tilde{\varepsilon},$$

где $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{\delta}$ и $\tilde{\varepsilon} = \varepsilon - \mathbf{x}\boldsymbol{\delta}$.

Если наблюдаемые данные $(\mathbf{y}, \mathbf{X})$ следуют этой модели, то оценка обычным методом наименьших квадратов $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ при стремлении количества наблюдений к бесконечности будет сходиться к вектору $\tilde{\boldsymbol{\beta}}$, а не к вектору $\boldsymbol{\beta}$. В некоторых случаях требуется оценить именно[5] $\tilde{\boldsymbol{\beta}}$. Однако, во многих приложениях интерес представляет зависимость (1) и оценки коэффициентов $\boldsymbol{\beta}$. Данное уравнение может быть содержательно интересным, например, потому что оно несет информацию, которая важна для экономической теории. Часто нам интересны коэффициенты, которые имели бы причинную интерпретацию, а не просто отражали бы корреляцию между переменными. Например, коэффициенты $\boldsymbol{\beta}$ могут нести информацию о том, как именно некоторые мероприятия экономической политики, изменяющие $\mathbf{x}$, скажутся на y.

[2]Несмещенность оценок $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ можно гарантировать при выполнении условий $\mathbb{E}[\varepsilon|\mathbf{x}] = 0$. Однако сама по себе смещенность оценок, если только она исчезает в пределе, при $n \to \infty$, не является столь серьезной проблемой.

[3]В более узком смысле понятие эндогенности относится к ситуации, когда переменные определяются совместно, в рамках системы уравнений (см. пункт 4.1).

[4]Если решение нормальных уравнений не единственно, то можно взять любое решение. Проекция от этого не зависит.

[5]Например, если мы хотим прогнозировать y по $\mathbf{x}$.

2 Почему регрессоры могут быть коррелированы с ошибкой?

Корреляция между регрессорами и ошибкой может быть вызвана разными причинами. Ниже мы рассмотрим наиболее характерные:

- пропущенные переменные, коррелированные с используемыми регрессорами;
- регрессоры, измеренные с ошибкой («ошибки в переменных»);
- одновременные взаимосвязи между переменными (обратная причинность или собственно эндогенность регрессоров);
- лаг зависимой переменной в правой части регрессии с автокоррелированной ошибкой.

2.1 Пропущенные переменные

Пусть интересующее нас уравнение содержит ненаблюдаемую переменную[6] q:

$$y = \mathbf{x}\boldsymbol{\beta} + \gamma q + v, \tag{2}$$

и пусть в этом уравнении ошибка v не коррелирована с $\mathbf{x}$ и q, т. е. $\mathcal{P}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}+\gamma q$. Поскольку q ненаблюдаема, вместо этого мы вынуждены оценивать регрессию

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon,$$

так что ошибка имеет вид $\varepsilon = \gamma q + v$.

К чему приводит пропуск q? Все зависит от того, есть ли взаимосвязь между пропущенной переменной q и остальными регрессорами. Если есть, то оставшиеся регрессоры «возьмут на себя» часть влияния q на y, и из-за этого коэффициенты при них будут смещенными.

Найдем это смещение количественно. Для этого введем линейную проекцию q на $\mathbf{x}$:

$$q = \mathbf{x}\boldsymbol{\delta} + r = \delta_0 + \delta_1 x_1 + \cdots + \delta_m x_m + r,$$

где $\mathcal{P}(r|\mathbf{x}) = 0$. Подставим эту зависимость в исходное уравнение (2):

$$y = \mathbf{x}\boldsymbol{\beta} + \gamma(\mathbf{x}\boldsymbol{\delta} + r) + v,$$

что можно переписать в виде

$$y = \mathbf{x}\tilde{\boldsymbol{\beta}} + \tilde{\varepsilon},$$

где $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \gamma\boldsymbol{\delta}$ и $\tilde{\varepsilon} = v + \gamma r$. В этой регрессии регрессоры $\mathbf{x}$ не коррелированы с $\tilde{\varepsilon}$, так что $\mathcal{P}(y|\mathbf{x}) = \mathbf{x}\tilde{\boldsymbol{\beta}}$. Оценки коэффициентов $\tilde{\boldsymbol{\beta}}$ будут смещенными по отношению к интересующем нас коэффициентам $\boldsymbol{\beta}$, если $\gamma \neq 0$ (ненаблюдаемая переменная q входит в исходное уравнение) и $\boldsymbol{\delta} \neq \mathbf{0}$ (т. е. q коррелирует с используемыми регрессорами $\mathbf{x}$).

Типичный пример подобной проблемы – так называемая «минцеровская» регрессия – влияние образования на заработную плату (см. Mincer, 1958):

$$W = \beta_0 + \beta_1 E + \varepsilon. \tag{3}$$

Минцеровская регрессия (интерпретируемая как регрессия зарплаты по человеческому капиталу) имеет важное значение для нескольких разделов экономической теории. Проблема состоит в том, что в большом количестве эмпирических исследований, в которых оценивалась

[6] Можно несколько: $\gamma_1 q_1 + \gamma_2 q_2 + \cdots$, но в данном случае это не имеет значения, коль скоро все переменные q_j ненаблюдаемы.

такая регрессия, не учитывалось возможное смещение, которое возникает из-за ненаблюдаемых природных способностей, которые, как естественно ожидать, влияют как на длительность обучения, так и на зарплату.

Если обозначить способности через A, интересующее нас соотношение следует записать в виде

$$W = \beta_0 + \beta_1 E + \gamma A + v. \tag{4}$$

В коэффициенте β_1 не будет смещения из-за пропуска A, только если E не связано с A. На самом деле это не так – природные способности положительно влияют на величину образования. Таким образом,

$$\mathcal{P}(A|E) = \delta_0 + \delta_1 E + r,$$

где $\delta_1 > 0$. Вместо β_1 мы получим оценку для $\tilde{\beta}_1 = \beta_1 + \gamma\delta_1$, где $\gamma\delta_1 > 0$ при $\gamma > 0$. Следовательно, оценивая минцеровскую регрессию обычным МНК, мы переоценим влияние образования как такового. Мыслима ситуация, когда образование вообще не влияет на производительность человека и его зарплату ($\beta_1 = 0$ в уравнении (4)), а положительная корреляция между W и E объясняется только пропуском природных способностей. Такая оценка β_1 не будет иметь никакого отношения к человеческому капиталу.

2.2 Одновременность (двусторонняя причинность)

Одновременность имеет место, когда две или более переменные одновременно влияют друг на друга, так что их значения определяются эндогенно из некоторой системы уравнений.

Хрестоматийным примером одновременности является оценивание уравнения спроса (или предложения, см., например, Working, 1927, Wright, 1928, приложение B). Пусть, например, мы хотим оценить уравнение спроса:

$$Q = \mathbf{x}^D \boldsymbol{\beta}^D + \alpha P + \varepsilon^D. \tag{5}$$

Можем пытаться оценить это уравнение с помощью (обычного) МНК, ожидая получить значимо отрицательную оценку для α (убывающая кривая спроса). Однако, скорее всего, такая оценка будет смещенной, если P и Q определяются на рынке одновременно, как точка пересечения спроса и предложения. Одновременно с уравнением спроса следует принять во внимание уравнение предложения:

$$Q = \mathbf{x}^S \boldsymbol{\beta}^S + \gamma P + \varepsilon^S. \tag{6}$$

Два уравнения образуют систему, в которой P и Q определяются эндогенно. Это так называемые *структурные уравнения* (структурная форма системы одновременных уравнений). Решим систему относительно P и Q и получим *приведенную форму* – зависимость P от $\mathbf{x}^D$, $\mathbf{x}^S$, ε^D, ε^S и зависимость Q от тех же переменных:

$$P = \mathbf{x}^D \boldsymbol{\delta}_{DP} + \mathbf{x}^S \boldsymbol{\delta}_{SP} + \theta_{DP}\varepsilon^D + \theta_{SP}\varepsilon^S,$$
$$Q = \mathbf{x}^D \boldsymbol{\delta}_{DQ} + \mathbf{x}^S \boldsymbol{\delta}_{SQ} + \theta_{DQ}\varepsilon^D + \theta_{SQ}\varepsilon^S.$$

Из приведенной формы видно, в частности, что P зависит от ε^D и, следовательно, если рассматривать уравнение спроса как регрессионное соотношение, то в нем регрессор коррелирует с ошибкой. Тот же эффект возникает в уравнении предложения. Проблема одновременных взаимосвязей (которую еще называют проблемой эндогенности или двусторонней причинности) – очень острая в прикладных исследованиях, и она не ограничивается оцениванием спроса и предложения.

Одним из примеров эндогенности объясняющей переменной в регрессии является уравнение, в котором объясняемой переменной служит логарифм цены на авиабилеты, а объясняющей переменной – индекс концентрации Герфиндаля, показывающий, насколько сильно монополизирован локальный рынок авиаперевозок. Можно собрать данные по различным маршрутам – парам городов – и исследовать подобную зависимость. Но такая интерпретация функциональной связи между ценой и индексом концентрации подразумевает экзогенность индекса концентрации, что, вообще говоря, может не выполняться, поскольку причинность может идти в другом направлении. Это обычное направление критики концепции *Structure – Conduct – Performance*, которая некогда была очень популярна в теории отраслевых рынков и породила огромное количество эмпирических исследований. Исходно подразумевалось, что структура рынка, *Structure* (например, индекс концентрации), влияет на поведение фирм, *Conduct*, а оно, в свою очередь, на прибыльность, эффективность и другие аналогичные характеристики функционирования отрасли, *Performance*. Но постепенно стало ясным, что все три аспекта взаимоувязаны и текущее состояние рынка определяется динамическим взаимодействием соответствующих переменных. В частности, в примере с авиаперевозками цены на рынке (маршруте авиаперевозок) могут влиять на решения фирм по поводу того, входить ли на этот рынок или уйти ли с него, и, тем самым, на индекс концентрации (см. Evans, Froeb & Werden, 1993).

Или пусть требуется выяснить, как влияют прямые иностранные инвестиции на развитие регионов. С этой целью можно оценить регрессию некоторого показателя экономического развития на объем инвестиций и прочие (контрольные) переменные. Вполне возможно, что при этом оценки регрессии укажут на значимо положительную связь между инвестициями и развитием. Но не будет ли тут двусторонней причинности? Ведь инвестиции, скорее всего, идут в экономически развитый регион, в котором ожидается хорошая отдача. Таким образом, мы не можем отделить одно влияние от другого и, оценивая уравнение для одностороннего влияния, получим смещенную оценку – некоторую смесь из двух эффектов.

2.3 Ошибки в переменных

Из-за различных причин экономические переменные, как правило, измеряются с ошибками. Бывает, например, что исследователь постулирует модель и вкладывает в переменные модели определенное содержание, а доступные исследователю данные относятся к переменным, которые не вполне соответствуют этому содержанию. В частности, исследователь может создать анкету для проведения обследования группы индивидуумов. Он вкладывает какой-то свой смысл в вопросы анкеты; в то же время, опрашиваемые могут не вполне понять эти вопросы или забыть правильную информацию. Это только один пример. Существует множество других источников ошибок в переменных.

Если есть ошибки измерения, то соотношения, которые оцениваются эконометристами, могут существенно отличаться от реальности. Если переменные в регрессии измерены с ошибкой, то результаты оценивания регрессии с помощью МНК могут оказаться смещенными. Это происходит из-за того, что ошибки измерения переменных регрессии становятся частью ошибки регрессии. Из-за этого ошибки измерения регрессоров входят и в сами регрессоры, и в ошибку регрессии, так что ошибка регрессии и регрессоры будут коррелированными между собой.

Пусть оценивается уравнение

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon,$$

однако на самом деле y порождается в соответствии с уравнением

$$y = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0, \tag{7}$$

и при этом $\mathcal{P}(y|\mathbf{x}_0) = \mathbf{x}_0\boldsymbol{\beta}$ и $\mathcal{P}(\varepsilon_0|\mathbf{x}_0) = \mathbf{0}$. Другими словами, $\mathbf{x}$ – это наблюдаемый аналог ненаблюдаемых «настоящих» регрессоров $\mathbf{x}_0$, измеренный с ошибкой:

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{u},$$

где $\mathbb{E}[\mathbf{u}] = \mathbf{0}$.

Если подставить $\mathbf{x}_0$, выраженные через $\mathbf{x}$, в уравнение (7), то получится

$$y = (\mathbf{x} - \mathbf{u})\boldsymbol{\beta} + \varepsilon_0 = \mathbf{x}\boldsymbol{\beta} + \varepsilon_0 - \mathbf{u}\boldsymbol{\beta}.$$

Таким образом, ошибка оцениваемого уравнения представима следующим образом:

$$\varepsilon = \varepsilon_0 - \mathbf{u}\boldsymbol{\beta}.$$

Эта ошибка может коррелировать с используемыми регрессорами $\mathbf{x}$:

$$\mathbb{C}[\mathbf{x}, \varepsilon] = \mathbb{E}[\mathbf{x}^\intercal\varepsilon] = \mathbb{E}[(\mathbf{x}_0 + \mathbf{u})^\intercal(\varepsilon_0 - \mathbf{u}\boldsymbol{\beta})] = \mathbb{E}[\mathbf{x}_0^\intercal\varepsilon_0] + \mathbb{E}[\mathbf{x}_0^\intercal\mathbf{u}]\boldsymbol{\beta} - \mathbb{E}[\mathbf{u}^\intercal\varepsilon_0] - \mathbb{E}[\mathbf{u}^\intercal\mathbf{u}]\boldsymbol{\beta}.$$

По предположению $\mathbb{E}[\mathbf{x}_0^\intercal\varepsilon_0] = 0$. Если, кроме того, ошибка измерения $\mathbf{u}$ некоррелирована с $\mathbf{x}_0$ и ε_0, то

$$\mathbb{C}[\mathbf{x}, \varepsilon] = -\mathbb{E}[\mathbf{u}^\intercal\mathbf{u}]\boldsymbol{\beta}.$$

Ковариационная матрица ошибки измерения – это некоторая положительно полуопределенная, не равная нулю матрица (если только $\mathbf{u}$ не равна нулю с вероятностью единица, т. е. ошибка измерения отсутствует). Таким образом, кроме вырожденного случая, когда $\mathbb{E}[\mathbf{u}^\intercal\mathbf{u}]\boldsymbol{\beta} = \mathbf{0}$, регрессоры оцениваемого уравнения коррелируют с ошибкой.

Как можно показать, при сделанных предположениях о ковариациях между ε_0, $\mathbf{x}_0$ и $\mathbf{u}$ выполнено $\mathcal{P}(\varepsilon|\mathbf{x}) = \mathbf{x}\boldsymbol{\delta}$, где

$$\boldsymbol{\delta} = -\big(\mathbb{E}[\mathbf{x}_0^\intercal\mathbf{x}_0] + \mathbb{E}[\mathbf{u}^\intercal\mathbf{u}]\big)^{-1}\mathbb{E}[\mathbf{u}^\intercal\mathbf{u}]\boldsymbol{\beta}.$$

Оценки коэффициентов регрессии β будут смещены на величину $\boldsymbol{\delta}$. В случае единственного регрессора и константы, т. е. модели вида

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

данную формулу можно записать в виде

$$\delta_1 = -\frac{\mathbb{V}[u]}{\mathbb{V}[x_0] + \mathbb{V}[u]}\beta_1.$$

Вместо β_1 регрессия будет оценивать $\tilde{\beta}_1$, где

$$\tilde{\beta}_1 = \frac{\mathbb{V}[x_0]}{\mathbb{V}[x_0] + \mathbb{V}[u]}\beta_1.$$

Очевидно, что в этом простейшем случае коэффициент наклона β_1 будет смещен в сторону нуля, и его абсолютная величина будет преуменьшаться.

Рис. 1 иллюстрирует этот простой случай. Сплошной линией показана истинная теоретическая зависимость между x_0 и y. Кружки́ изображают наблюдения (x_0, y). В результате того, что регрессор наблюдается с ошибкой, эти точки сдвигаются на величину ошибки u по горизонтали. Звездочками обозначены соответствующие наблюдения (x, y). Точки могут смещаться и влево, и вправо, но в целом облако наблюдений из-за ошибки расползается по горизонтали, становясь более плоским. В результате линия регрессии, подогнанная с помощью обычного МНК, оказывается слишком пологой (она показана штриховой линией). Если бы наблюдался исходный регрессор x_0, то подогнанная линия регрессии чаще всего не была бы столь пологой (показана пунктиром).

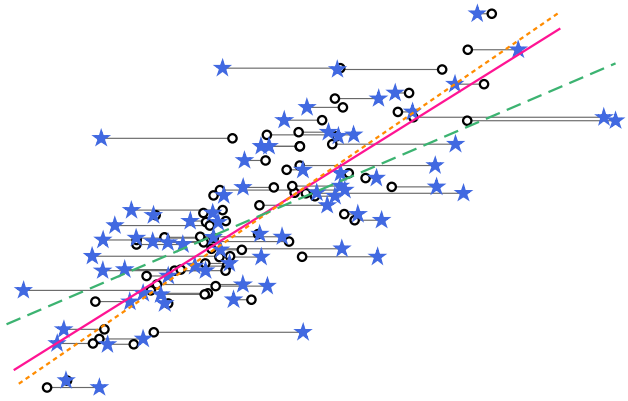

Рис. 1: Ошибки в переменных

2.4 Автокоррелированные ошибки в динамической регрессии

В регрессии с временными рядами с полностью экзогенными регрессорами автокорреляция ошибок не приводит к смещению и несостоятельности оценок.[7] Однако, если среди регрессоров имеются лаги зависимой переменной, то оценки обычного МНК будут несостоятельными. Это объясняется тем, что в этом случае лаги зависимой переменной будут, скорее всего, коррелированы с ошибкой.

Например, если рассматривать модель $\mathsf{ARMA}(1,1)$ как регрессию, в которой в качестве регрессора используется лаг y_t, а ошибка имеет вид $\mathsf{MA}(1)$:

$$y_t = \mu + \varphi y_{t-1} + \varepsilon_t, \tag{8}$$

где $\varepsilon_t = \eta_t - \theta\eta_{t-1}$, то в этой регрессии y_{t-1} будет зависеть от η_{t-1}, поскольку $y_{t-1} = \varphi y_{t-2} + \mu + \eta_{t-1} - \theta\eta_{t-2}$. Это означает, что в (8) регрессор и ошибка в общем случае зависят от η_{t-1} и коррелированы между собой. Таким образом, обычный МНК не подходит для оценивания $\mathsf{ARMA}(1,1)$.

Модель Койка (геометрического распределенного лага) – еще один пример подобной ситуации. Она имеет вид

$$y_t = \mu + \alpha \sum_{j=0}^{\infty} \delta^j x_{t-j} + \varepsilon_t. \tag{9}$$

Применив к этому уравнению преобразование Койка, т. е. умножив его на $1 - \delta\mathsf{L}$, где L – оператор лага, получим

$$y_t = \mu' + \alpha x_t + \delta y_{t-1} + \varepsilon'_t, \tag{10}$$

где $\varepsilon'_t = \varepsilon_t - \delta\varepsilon_{t-1}$ и $\mu' = \mu(1-\delta)$. В этой регрессии опять регрессор y_{t-1} является лагом зависимой переменной, ошибка автокоррелирована (представляет собой процесс $\mathsf{MA}(1)$), и поэтому регрессор будет коррелирован с ошибкой. Конечно, сама по себе модель Койка вряд ли может использоваться в прикладном анализе, но этот простой пример показывает, что проблема корреляции между регрессором и ошибкой легко может проявиться при моделировании динамических взаимосвязей.

[7] Конечно, при этом обычная оценка ковариационной матрицы оценок будет несостоятельной.

3 Метод инструментальных переменных

Предположим, что в регрессии

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon \tag{11}$$

регрессоры $\mathbf{x}$ являются (частично) случайными, и нарушена гипотеза о том, что ошибка ε не зависима от регрессоров $\mathbf{x}$, так что корреляция между регрессорами $\mathbf{x}$ и ошибкой ε может быть не равной нулю. Для дальнейшего удобно разделить регрессоры на две группы – те, про которые известно, что они не коррелируют с ошибкой, $\mathbf{x}^\circ$ (в количестве m° штук), и те, которые находятся под подозрением и могут коррелировать с ошибкой, $\mathbf{x}^\star$ (в количестве $m^\star = m - m^\circ$ штук):

$$y = \mathbf{x}^\circ\boldsymbol{\beta}^\circ + \mathbf{x}^\star\boldsymbol{\beta}^\star + \varepsilon. \tag{12}$$

Какие существуют способы решения проблемы несостоятельности оценок МНК в такой ситуации? Той информации, наличие которой обычно предполагается при рассмотрении модели регрессии, недостаточно, чтобы получить состоятельные оценки коэффициентов $\boldsymbol{\beta}$.

В зависимости от причины, по которой регрессоры коррелируют с ошибкой, может помочь получение информации разного вида. Например, можно собрать данные по пропущенным переменным, измерять регрессоры с большей точностью, и т. д. Универсальным «лекарством» здесь может являться проведение контролируемого эксперимента, когда значения переменных $\mathbf{x}^\star$ выбираются (назначаются) таким образом, чтобы гарантировать их экзогенность. При этом желательно устанавливать значения переменных случайно, чтобы это был так называемый рандомизированный эксперимент.

Еще один классический способ – и мы на нем здесь сосредоточимся – это получить дополнительную информацию за счет сбора данных о переменных, которые связаны с оцениваемым уравнением только косвенно, через эндогенные регрессоры $\mathbf{x}^\star$. Понятно, что применение любого из упомянутых способов ограничивается необходимыми издержками, этическими соображениями, и часто может быть вовсе невозможным.

3.1 Описание метода

Оказывается, что регрессию (11) можно состоятельно оценить, имея набор (вектор-строку) из p вспомогательных переменных $\mathbf{z}$, называемых *инструментальными переменными*. Часто инструментальные переменные называют просто *инструментами*. В английском языке для обозначения инструментальных переменных и соответствующего метода используется аббревиатура IV (от англ. *instrumental variable* – инструментальная переменная).

Для того чтобы переменные $\mathbf{z}$ можно было использовать в качестве инструментальных, нужно, чтобы они удовлетворяли следующим требованиям.

- Инструменты $\mathbf{z}$ некоррелированы с ошибкой ε (в противном случае метод даст несостоятельные оценки, как и МНК). Если это условие не выполнено, то такие переменные называют *негодными инструментами* (англ. *invalid instruments*).

- Инструменты $\mathbf{z}$ достаточно сильно связаны с регрессорами $\mathbf{x}$, т. е. как говорят, являются *релевантными*. Если данное условие не выполнено, то это так называемые *«слабые» инструменты* (англ. *weak instruments*), а то и вовсе нерелевантные. Если инструменты слабые, то, в частности, оценки по методу будут неточными и при малом количестве наблюдений сильно смещенными. Эти и другие проблемы, возникающие в ситуации, когда инструменты являются слабыми, обсуждаются в разделе 6.

Обычно $\mathbf{x}$ и $\mathbf{z}$ содержат общие переменные, т. е. часть регрессоров используется в качестве инструментов. Например, типична ситуация, когда $\mathbf{x}$ содержит константу; тогда в $\mathbf{z}$ тоже следует включить константу, ибо детерминистические регрессоры по своей природе и годные, и релевантные инструменты. В целом следует включить в число инструментов все регрессоры, которые не коррелируют с ошибкой, т. е. $\mathbf{x}^\circ$. Обозначим через $\mathbf{z}^\dagger$ те инструментальные переменные ($p^\dagger = p - m^\circ$ штук), которых нет среди регрессоров, или, другими словами, *внешние инструменты* или *исключенные инструменты*.[8] В этих обозначениях

$$\mathbf{z} = (\mathbf{x}^\circ, \mathbf{z}^\dagger).$$

Пусть имеются n наблюдений, и $\mathbf{y}$, $\mathbf{X}$ и $\mathbf{Z}$ – соответствующие данные в матричном виде. Здесь $\mathbf{y}$ – вектор-столбец длиной n, $\mathbf{X} = [\mathbf{X}^\circ, \mathbf{X}^\star]$ – матрица $n \times m$, $\mathbf{Z} = [\mathbf{X}^\circ, \mathbf{Z}^\dagger]$ – матрица $n \times p$. *Метод инструментальных переменных* состоит в том, что оценки коэффициентов β вычисляются по формуле

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\mathbf{X}^\intercal \mathbf{P}_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{P}_{\mathbf{Z}} \mathbf{y}, \tag{13}$$

где $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}\big(\mathbf{Z}^\intercal \mathbf{Z}\big)^{-1} \mathbf{Z}^\intercal$. В таком виде метод также называют иногда *обобщенным* методом инструментальных переменных (GIVE от *generalized instrumental variables estimator*), поскольку количество инструментальных переменных может быть больше количества регрессоров, в отличие от простого (классического) метода инструментальных переменных, для которого $p = m$.

Может быть полезным выработать геометрическую интуицию, связанную с использованием метода инструментальных переменных. Обычный метод наименьших квадратов сводится к поиску линейной комбинации регрессоров $\hat{\mathbf{y}}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, такой чтобы она ближе всего аппроксимировала зависимую переменную $\mathbf{y}$. Мерой близости служит обычное евклидово расстояние $\|\mathbf{y} - \hat{\mathbf{y}}(\boldsymbol{\beta})\|$ (все рассуждения проводятся в n-мерном евклидовом пространстве). Матрица $\mathbf{P}_{\mathbf{Z}}$ представляет собой *матрицу проекции* на подпространство, натянутое на столбцы матрицы инструментов $\mathbf{Z}$ (коротко будем говорить «подпространство $\mathbf{Z}$»). Метод инструментальных переменных тоже минимизирует расстояние, но это расстояние между проекциями векторов $\hat{\mathbf{y}}(\boldsymbol{\beta})$ и $\mathbf{y}$ на подпространство $\mathbf{Z}$, т. е. это расстояние между $\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{y}}(\boldsymbol{\beta})$ и $\mathbf{P}_{\mathbf{Z}}\mathbf{y}$. На интуитивном уровне, метод инструментальных переменных использует при оценивании регрессии только ту часть изменчивости переменных регрессии, которая остается после проекции их на подпространство $\mathbf{Z}$. Коль скоро инструменты являются годными, то эта часть изменчивости переменных не будет связана с ошибкой, и минимизация расстояния не будет приводить к смещению.

В случае, если количество инструментальных переменных в точности равно количеству регрессоров ($p = m$), получаем собственно классический метод инструментальных переменных. При этом матрица $\mathbf{Z}^\intercal \mathbf{X}$ квадратная и оценки вычисляются как

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\mathbf{Z}^\intercal \mathbf{X})^{-1} \mathbf{Z}^\intercal \mathbf{Z} \, (\mathbf{X}^\intercal \mathbf{Z})^{-1} \mathbf{X}^\intercal \mathbf{Z} \, (\mathbf{Z}^\intercal \mathbf{Z})^{-1} \mathbf{Z}^\intercal \mathbf{y}.$$

Средняя часть формулы сокращается, поэтому

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\mathbf{Z}^\intercal \mathbf{X})^{-1} \mathbf{Z}^\intercal \mathbf{y}. \tag{14}$$

Рассмотрим вывод простого метода инструментальных переменных, т. е. случай точной идентификации. Умножим уравнение регрессии $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon$ слева на инструменты $\mathbf{z}$ (с транспонированием). Получим следующее уравнение:

$$\mathbf{z}^\intercal y = \mathbf{z}^\intercal \mathbf{x} \boldsymbol{\beta} + \mathbf{z}^\intercal \varepsilon.$$

[8] Часто именно такие переменные и называют инструментами. Мы тоже иногда будем использовать термин в этом значении.

Если взять от обеих частей математическое ожидание, то с учетом того, что инструменты некоррелированы с ошибкой ($\mathbb{E}[\mathbf{z}^\intercal\varepsilon] = 0$), получится

$$\mathbf{Q}_{zy} = \mathbf{Q}_{zx}\boldsymbol{\beta},$$

где $\mathbf{Q}_{zx} = \mathbb{E}[\mathbf{z}^\intercal\mathbf{x}]$ и $\mathbf{Q}_{zy} = \mathbb{E}[\mathbf{z}^\intercal y]$.

Заменяя теоретические моменты на выборочные, получим следующие нормальные уравнения, задающие оценки $\hat{\boldsymbol{\beta}}$:

$$\bar{\mathbf{Q}}_{zy} = \bar{\mathbf{Q}}_{zx}\hat{\boldsymbol{\beta}},$$

где $\bar{\mathbf{Q}}_{zy} = \frac{1}{n}\mathbf{Z}^\intercal\mathbf{y}$ и $\bar{\mathbf{Q}}_{zx} = \frac{1}{n}\mathbf{Z}^\intercal\mathbf{X}$, или

$$\mathbf{Z}^\intercal\mathbf{y} = \mathbf{Z}^\intercal\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{15}$$

Очевидно, что эти оценки совпадут с (14). Фактически, мы применяем здесь *метод моментов*.

Метод инструментальных переменных можно рассматривать как так называемый *двухшаговый метод наименьших квадратов* (для него используется аббревиатура 2SLS, от англ. *two-stage least squares*).

1-й шаг. Строим регрессию каждого регрессора $\mathbf{X}_j$ на $\mathbf{Z}$.

$$\mathbf{X}_j = \mathbf{Z}\boldsymbol{\lambda}_j + \mathbf{V}_j. \tag{16}$$

Получим в этой регрессии расчетные значения $\hat{\mathbf{X}}_j$. По формуле расчетных значений в регрессии $\hat{\mathbf{X}}_j = \mathbf{P}_\mathbf{Z}\mathbf{X}_j$. Заметим, что если $\mathbf{X}_j$ входит в число инструментов, то выполнено $\hat{\mathbf{X}}_j = \mathbf{X}_j$, т. е. такая переменная останется без изменений. Значит, данную процедуру достаточно применять только к тем регрессорам, которые не являются инструментами (т. е. могут быть коррелированы с ошибкой). Обобщающее уравнение для регрессий первого шага для таких регрессоров можно записать в виде

$$\mathbf{X}^\star = \mathbf{Z}\boldsymbol{\Lambda} + \mathbf{V} = \mathbf{X}^\circ\boldsymbol{\Lambda}^\circ + \mathbf{Z}^\dagger\boldsymbol{\Lambda}^\dagger + \mathbf{V}. \tag{17}$$

Это уравнение нам понадобится в дальнейшем при обсуждении свойств метода инструментальных переменных. В целом для всей матрицы регрессоров можем записать $\hat{\mathbf{X}} = \mathbf{P}_\mathbf{Z}\mathbf{X}$.

2-й шаг. В исходной регрессии используются $\hat{\mathbf{X}}$ вместо $\mathbf{X}$, т. е. оценивается регрессия вида

$$\mathbf{y} = \hat{\mathbf{X}}\boldsymbol{\beta} + \text{ошибка}. \tag{18}$$

Смысл состоит в том, чтобы использовать регрессоры, «очищенные от ошибок». Получаем следующие оценки:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{2SLS}} &= \left(\hat{\mathbf{X}}^\intercal\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}^\intercal\mathbf{y} = \left(\mathbf{X}^\intercal\mathbf{P}_\mathbf{Z}\mathbf{P}_\mathbf{Z}\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{P}_\mathbf{Z}\mathbf{y} = \\ &= \left(\mathbf{X}^\intercal\mathbf{P}_\mathbf{Z}\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{P}_\mathbf{Z}\mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{IV}}.\end{aligned}$$

Видим, что оценки совпадают.

Конечно, не обязательно непосредственно рассчитывать регрессии, подразумеваемые двухшаговым МНК, чтобы получить оценки (13). Однако такое двухшаговое представление может быть полезным как для теоретического анализа оценок инструментальных переменных, так и для диагностики модели, оцененной по конкретным данным (см. раздел 6).

Если записать оценки в виде

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\hat{\mathbf{X}}^\intercal\mathbf{X}\right)^{-1}\hat{\mathbf{X}}^\intercal\mathbf{y}, \tag{19}$$

то видно, что обобщенный метод инструментальных переменных можно рассматривать как простой метод инструментальных переменных с матрицей инструментов $\hat{\mathbf{X}}$. Такая запись позволяет обосновать обобщенный метод инструментальных переменных. Если исходных инструментов $\mathbf{Z}$ больше, чем регрессоров $\mathbf{X}$ ($p > m$), и мы хотим построить на их основе меньшее количество инструментов m, то имеет смысл сопоставить каждому регрессору $\mathbf{X}_j$ в качестве инструмента такую линейную комбинацию исходных инструментов, которая была бы *наиболее близка* к $\mathbf{X}_j$ (в смысле евклидова расстояния). Этому требованию как раз и удовлетворяют расчетные значения $\hat{\mathbf{X}}_j$.

Другое обоснование обобщенного метода инструментальных переменных состоит, как и выше для классического метода, в использовании уравнений $\mathbb{E}[\mathbf{z}^\intercal y] = \mathbb{E}[\mathbf{z}^\intercal \mathbf{x}]\boldsymbol{\beta}$ или $\mathbf{Q}_{zy} = \mathbf{Q}_{zx}\boldsymbol{\beta}$. Заменой теоретических моментов выборочными получим уравнения $\bar{\mathbf{Q}}_{zy} = \bar{\mathbf{Q}}_{zx}\boldsymbol{\beta}$, число которых больше числа неизвестных. Идея состоит в том, чтобы невязки уравнений $\bar{\mathbf{Q}}_{zy} - \bar{\mathbf{Q}}_{zx}\boldsymbol{\beta}$ были как можно меньшими. Это можно сделать, минимизируя следующую квадратичную форму от невязок:

$$(\bar{\mathbf{Q}}_{zy} - \bar{\mathbf{Q}}_{zx}\boldsymbol{\beta})^\intercal \bar{\mathbf{Q}}_{zz}^{-1} (\bar{\mathbf{Q}}_{zy} - \bar{\mathbf{Q}}_{zx}\boldsymbol{\beta}), \tag{20}$$

где $\bar{\mathbf{Q}}_{zz} = \frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z}$. Минимум достигается при

$$\hat{\boldsymbol{\beta}} = \left(\bar{\mathbf{Q}}_{zx}^\intercal \bar{\mathbf{Q}}_{zz}^{-1} \bar{\mathbf{Q}}_{zx}\right)^{-1} \bar{\mathbf{Q}}_{zx}^\intercal \bar{\mathbf{Q}}_{zz}^{-1} \bar{\mathbf{Q}}_{zy}.$$

Видим, что эта формула совпадает с (13). Эти рассуждения представляют собой применение так называемого *обобщенного метода моментов*, в котором количество условий на моменты может превышать количество неизвестных параметров.[9]

Выбор $\bar{\mathbf{Q}}_{zz}^{-1}$ в качестве взвешивающей матрицы при минимизации невязок в (20) объясняется тем, что она при обычных предположениях дает наименьшую асимптотическую ковариационную матрицу оценок. Другими словами, это оценки так называемого *эффективного метода моментов*.

При соответствующих предположениях оценки метода инструментальных переменных состоятельные[10] и асимптотически нормальные. Чтобы можно было использовать метод инструментальных переменных на практике, нужна оценка ковариационной матрицы, с помощью которой можно было бы вычислить стандартные ошибки коэффициентов и t-статистики и в целом проверять гипотезы по принципу Вальда. Такая оценка имеет вид

$$\mathbb{V}[\hat{\boldsymbol{\beta}}_{\mathrm{IV}}] \approx s^2 \left(\hat{\mathbf{X}}^\intercal \hat{\mathbf{X}}\right)^{-1} = s^2 \left(\mathbf{X}^\intercal \mathbf{P}_{\mathbf{Z}} \mathbf{X}\right)^{-1}. \tag{21}$$

Здесь s^2 — оценка дисперсии ошибок $\sigma^2 = \mathbb{V}[\varepsilon]$, например $s^2 = \mathbf{e}^\intercal\mathbf{e}/n$ или $s^2 = \mathbf{e}^\intercal\mathbf{e}/(n-m)$. Остатки рассчитываются по обычной формуле $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. При этом следует помнить, что остатки, получаемые на втором шаге, тут не годятся, поскольку они равны $\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. Если использовать их для расчета оценки дисперсии, то получим некорректную (обычно завышенную) оценку дисперсии и ковариационной матрицы. Отсюда следует, что из регрессии второго шага можно использовать только оценки коэффициентов. Стандартные ошибки и t-статистики требуется пересчитывать. Подробнее о проверке гипотез речь пойдет ниже, в пункте 5.1.

По-видимому, впервые метод инструментальных переменных был сформулирован в работе Wright (1928) как метод оценивания кривых спроса и предложения. Само название «инструментальные переменные» впервые употреблено в статье Reiersol (1941) при обсуждении модели ошибок в переменных. Метод получил развитие в работах Durbin (1954), Sargan (1958) и др. В контексте систем одновременных уравнений метод развивался параллельно под названием «двухшаговый МНК».

[9] См. Hansen (1982).

[10] Состоятельность доказывается по той же схеме, которая описана выше для обычного МНК.

3.2 Идентификация

Обсудим теперь проблему идентификации.[11] Какие условия должны выполняться, чтобы можно было вычислить оценки (13)?

Во-первых, матрица инструментов должна иметь полный ранг по столбцам ($\operatorname{rank}\mathbf{Z} = p$), иначе $(\mathbf{Z}^\intercal\mathbf{Z})^{-1}$ не существует. Это условие не столь важно, поскольку, на самом деле, значение имеет лишь матрица проекции $\mathbf{P}_\mathbf{Z}$, которую можно определить при любой $\mathbf{Z}$. Мы будем предполагать, что условие выполнено (если это не так, то всегда можно убрать часть инструментов, чтобы избавиться от линейной зависимости). Таким образом, в дальнейшем считаем, что $\operatorname{rank}\mathbf{Z} = p$.

Во-вторых, матрица $\mathbf{X}^\intercal\mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{X} = \hat{\mathbf{X}}^\intercal\hat{\mathbf{X}}$ должна быть невырожденной. Это условие эквивалентно тому, что матрица $\hat{\mathbf{X}}$ имеет полный ранг по столбцам. Для этого необходимо, чтобы $\mathbf{X}$ имела полный ранг по столбцам, так как из $\operatorname{rank}\mathbf{X} < m$ следует $\operatorname{rank}\hat{\mathbf{X}} < m$. Условие $\operatorname{rank}\mathbf{X} = m$ – это стандартное условие идентификации для линейной регрессии, и мы в дальнейшем будем предполагать, что оно выполнено.

Далее, если количество инструментов меньше количества регрессоров ($p < m$), то, поскольку $\hat{\mathbf{X}}$ имеет неполный ранг (даже если $\operatorname{rank}\mathbf{X} = m$), оцениваемое уравнение *неидентифицируемо*, т. е. невозможно вычислить оценки (13). Таким образом, количество инструментов должно быть не меньше m (количество регрессоров). Если $p > m$, то говорят, что уравнение *сверхидентицировано*. Если количество инструментов равно m, то это *точная идентификация*. Если возможен случай сверхидентификации, то это обобщенный метод инструментальных переменных. При точной идентификации ($p = m$) получаем собственно классический метод инструментальных переменных.

Таким образом, необходимое условие идентификации имеет следующий вид:

$$p \geqslant m.$$

Это так называемое *порядковое условие* идентификации, условие на размерность матриц. Словесная формулировка порядкового условия:

> Количество инструментов $\mathbf{Z}$ должно быть не меньше количества регрессоров $\mathbf{X}$.

Заметим, что можно сначала «вычеркнуть» общие переменные в $\mathbf{X}$ и $\mathbf{Z}$ и смотреть только на количество оставшихся. Количество оставшихся инструментов должно быть не меньше количества оставшихся регрессоров.

> Количество внешних инструментов $\mathbf{Z}^\dagger$ должно быть не меньше количества эндогенных регрессоров $\mathbf{X}^\star$,

т. е. $p^\dagger \geqslant m^\star$.

Почему это только необходимое условие? Пусть, например, некоторый регрессор $\mathbf{X}_j$ ортогонален $\mathbf{Z}$. Тогда $\hat{\mathbf{X}}_j = 0$, и невозможно получить оценки $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. Следовательно, данное условие не является достаточным. Необходимое и достаточное условие идентификации при конечном числе наблюдений формулируется следующим образом:

> Матрица $\hat{\mathbf{X}}$ имеет полный ранг по столбцам: $\operatorname{rank}\hat{\mathbf{X}} = m$.

Это так называемое *ранговое условие* идентификации.

Встречаются случаи, когда ранговое условие идентификации соблюдается, но матрица $\hat{\mathbf{X}}$ близка к вырожденности, т. е. в $\hat{\mathbf{X}}$ наблюдается мультиколлинеарность. Например, если инструменты $\mathbf{Z}$ являются слабыми для $\mathbf{X}_j$ в том смысле, что $\mathbf{X}_j$ и $\mathbf{Z}$ почти ортогональны, то $\hat{\mathbf{X}}$ близка к вырожденности.

[11] Идентификация в контексте метода инструментальных переменных тесно связана с идентификацией в контексте систем одновременных уравнений.

3.3 Где взять инструменты?

Самая трудная проблема в методе инструментальных переменных – это поиск подходящих инструментов. Требуется, чтобы инструменты были близко связаны с эндогенными регрессорами (или, другими словами, релевантными), но сами не были эндогенными.

Как мы видели в пункте 2.3, в модели ошибок в переменных ошибка регрессии имеет вид $\varepsilon = \varepsilon_0 - \mathbf{u}\boldsymbol{\beta}$, где ε_0 — ошибка в исходном уравнении, а $\mathbf{u}$ – ошибка измерения регрессоров $\mathbf{x}$. Чтобы переменные $\mathbf{z}$ можно было использовать в качестве инструментов, достаточно, чтобы $\mathbf{z}$ были некоррелированы с ε_0 и $\mathbf{u}$, но были связаны с $\mathbf{x}_0$. Например, это может быть, как и $\mathbf{x}$, некоторый неточный измеритель $\mathbf{x}_0$, но такой, что ошибки измерения $\mathbf{z}$ и $\mathbf{x}$ между собой не связаны.

Один из важных источников инструментов – это наблюдения за переменными из $\mathbf{x}^\star$, но произведенные в более ранний момент времени, другими словами, переменные из $\mathbf{x}^\star$ с лагом. Такое использование лагов для моделей временных рядов рассматривается в пункте 4.2. Для одномоментных (*cross-section*) данных использование запаздывающих значений регрессоров в качестве инструментов тоже актуально. К примеру, Evans, Froeb & Werden (1993) использовали для упоминавшейся выше модели влияния индекса концентрации на цены авиабилетов лаг индекса концентрации в качестве инструмента. При этом требуется иметь наблюдения за одними и теми же экономическими единицами в разные моменты времени, что предполагает использование так называемых панельных данных.

Лучше всего, когда инструментальные переменные до какой-то степени моделируют ситуацию проведения эксперимента, причем значения регрессоров устанавливаются случайным образом (когда, фактически, имеет место рандомизированный эксперимент). Даже если полный эксперимент невозможен, т. е. невозможно устанавливать значения регрессоров на произвольном желаемом уровне, часто возможно оказывать влияние на эндогенные регрессоры. При этом интенсивность влияния на переменную служит естественным инструментом для этой переменной. Если влияние со стороны исследователя в принципе невозможно, как это обычно и бывает в экономике, то может спонтанно возникнуть ситуация, когда внешние силы (природа или действия правительства) создают экзогенные флуктуации в объясняющих переменных. Подобную ситуацию принято называть *естественным экспериментом*.

Интересный случай инструментальных переменных – это инструментальные переменные, некоррелированность которых с ошибкой следует непосредственно из лежащей в основе эконометрической модели экономической теории (см., например, Hansen & Singleton, 1982).

В целом можно сказать, что очень редко когда используемые инструментальные переменные однозначно экзогенны. Часто исследователи выдвигают правдоподобные аргументы в пользу экзогенности переменных, но при более внимательном рассмотрении вполне может оказаться, что они чего-то не учли в своих рассуждениях. В частности, может обнаружиться, что переменная, которая казалась внешним инструментом, на самом деле должна быть одним из регрессоров. В таком случае инструментальная переменная становится пропущенной переменной, «уходит в ошибку», отчего ошибка становится коррелированной с «инструментом», и в результате инструмент является негодным, что приводит к несостоятельности оценок.

Если же инструменты вызывают мало подозрений с точки зрения экзогенности, то вполне вероятно, что они окажутся очень слабыми. В некотором смысле здесь существует обратная зависимость между слабостью инструмента и степенью сомнения в его экзогенности. Одна крайность здесь – это использование в качестве инструмента регрессора, экзогенность которого сомнительна, другая – это использование для получения инструмента датчика случайных чисел, который конечно, породит несомненно экзогенную переменную, но совершенно нерелевантную.

Найти однозначно годные и релевантные инструменты – это большая удача. Как бы то ни было, несмотря на все сложности и подводные камни, концепция инструментальных пере-

менных является очень важной для эконометрического анализа.

4 Инструменты в различных моделях

4.1 Инструментальные переменные и системы одновременных уравнений

Метод инструментальных переменных – это широко используемый метод оценки параметров отдельного структурного уравнения в системе (линейных) одновременных уравнений, т. е. систем регрессионных уравнений, в которых часть переменных определяются внутрисистемно, или, другими словами, являются эндогенными (см., например, Theil, 1953, Basmann, 1957). В этом контексте он известен под названием двухшагового МНК.

Можно предполагать, что в основе метода инструментальных переменных лежит следующая модель:

$$\begin{aligned} y &= \mathbf{x}\boldsymbol{\beta} + \varepsilon = \mathbf{x}^\circ\boldsymbol{\beta}^\circ + \mathbf{x}^\star\boldsymbol{\beta}^\star + \varepsilon, \\ \mathbf{x}^\star &= \mathbf{z}\boldsymbol{\Lambda} + \mathbf{v} = \mathbf{x}^\circ\boldsymbol{\Lambda}^\circ + \mathbf{z}^\dagger\boldsymbol{\Lambda}^\dagger + \mathbf{v}. \end{aligned} \tag{22}$$

В этой модели $\mathbf{z} = [\mathbf{x}^\circ, \mathbf{z}^\dagger]$ – инструменты. Второе уравнение представляет собой линейную проекцию эндогенных регрессоров первого уравнения $\mathbf{x}^\star$ на инструментальные переменные $\mathbf{z}$. Такую систему уравнений называют системой уравнений с ограниченной информацией, поскольку акцент делается на оценивание лишь одного уравнения. Все остальные уравнения сводятся к приведенной форме и любые структурные ограничения, не относящиеся к рассматриваемому уравнению, игнорируются.

Близким к двухшаговому методу наименьших квадратов методом оценивания коэффициентов отдельного уравнения системы одновременных уравнений является метод максимального правдоподобия с ограниченной информацией (англ. *limited information maximum likelihood*, LIML). Метод был предложен в статье Anderson & Rubin (1949). Как можно продемонстрировать, если предположить нормальность ошибок в уравнении (22), отсутствие автокорреляции и условную гомоскедастичность ошибок $(\varepsilon, \mathbf{v})$, задача максимизации функции правдоподобия сводится к следующей задаче наименьшего дисперсионного отношения:

$$\begin{aligned} \kappa &= \frac{\tilde{\mathbf{y}}^\intercal \mathbf{M}_{\mathbf{X}^\circ} \tilde{\mathbf{y}}}{\tilde{\mathbf{y}}^\intercal \mathbf{M}_{\mathbf{Z}} \tilde{\mathbf{y}}} \to \min_{\boldsymbol{\beta}^\star} \\ \text{где } \tilde{\mathbf{y}} &= \tilde{\mathbf{y}}(\boldsymbol{\beta}^\star) = \mathbf{y} - \mathbf{X}^\star\boldsymbol{\beta}^\star, \\ \mathbf{M}_{\mathbf{Z}} &= \mathbf{I} - \mathbf{P}_{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}\big(\mathbf{Z}^\intercal\mathbf{Z}\big)^{-1}\mathbf{Z}^\intercal, \\ \mathbf{M}_{\mathbf{X}^\circ} &= \mathbf{I} - \mathbf{X}^\circ(\mathbf{X}^{\circ\intercal}\mathbf{X}^\circ)^{-1}\mathbf{X}^{\circ\intercal}. \end{aligned}$$

(Заметим попутно, что эту задачу можно записать несколько иначе, в виде задачи минимизации следующей «F-статистики» Андерсона–Рубина:

$$F(\boldsymbol{\beta}^\star) = \frac{\tilde{\mathbf{y}}^\intercal(\mathbf{M}_{\mathbf{X}^\circ} - \mathbf{M}_{\mathbf{Z}})\tilde{\mathbf{y}}/p^\dagger}{\tilde{\mathbf{y}}^\intercal \mathbf{M}_{\mathbf{Z}} \tilde{\mathbf{y}}/(n-p)} \to \min_{\boldsymbol{\beta}^\star}. \tag{23}$$

Эта статистика еще будет упоминаться в дальнейшем.)

В свою очередь, задача наименьшего дисперсионного отношения сводится к поиску минимального собственного значения некоторой матрицы. Если $\hat{\kappa}$ – соответствующее минимальное значение κ (или, что то же самое, наименьшее собственное значение), то оценки параметров $\boldsymbol{\beta}$ первого уравнения в (22) получаются по формуле

$$\hat{\boldsymbol{\beta}}_{\text{LIML}} = (\mathbf{X}^\intercal(\mathbf{I} - \hat{\kappa}\mathbf{M}_{\mathbf{Z}})\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{I} - \hat{\kappa}\mathbf{M}_{\mathbf{Z}})\mathbf{y}.$$

Несложно увидеть, что эти оценки имеют вид

$$\hat{\boldsymbol{\beta}}_{\text{LIML}} = (\hat{\mathbf{X}}^\intercal\mathbf{X})^{-1}\hat{\mathbf{X}}^\intercal\mathbf{y},$$

что совпадает с формулой (19), только в данном случае $\hat{\mathbf{X}} = (\mathbf{I} - \hat{\kappa}\mathbf{M}_{\mathbf{Z}})\mathbf{X}$. Это означает, что в LIML используется несколько другая функция исходных инструментов $\mathbf{Z}$ по сравнению с двухшаговым МНК. С ростом количества наблюдений до бесконечности $\hat{\kappa}$ стремиться к единице и два метода оценивания сближаются.

При тех предположениях, которые лежат в основе LIML, оценки LIML и двухшагового МНК асимптотически эквивалентны и имеют одно и то же асимптотическое нормальное распределение. Однако их распределения в конечных выборках могут различаться, причем какой из методов дает более точные оценки нельзя сказать однозначно. Есть свидетельства, что в случае слабых инструментов или большого числа инструментов LIML предпочтительнее.

Примечательно, что классические методы оценивания систем одновременных уравнений, использующие полную информацию (трехшаговый МНК, FIML), также тесно связаны с методом инструментальных переменных и дают для отдельных уравнений оценки, которые представимы в виде (19) при соответствующем определении матрицы $\hat{\mathbf{X}}$.

4.2 Инструментальные переменные и модели временных рядов

Использование инструментальных переменных в моделях временных рядов можно рассмотреть на примере довольно общей модели ARMAX Бокса—Дженкинса:

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \sum_{j=1}^{p} \varphi_j y_{t-j} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{24}$$

В этой модели $\mathbf{X}_t$ – экзогенные регрессоры. Если рассматривать эту модель как регрессию, то в ней MA-ошибка коррелирована с лагами зависимой переменной. Мы уже видели это на примере частных случаев (8) и (10). Получить состоятельные оценки коэффициентов в этой модели можно с помощью метода инструментальных переменных. В качестве инструментов здесь естественно использовать $\mathbf{X}_t$ и лаги переменной y_t. Действительно, при $j > q$ переменная y_{t-j} не будет коррелировать с $\varepsilon_t, \ldots, \varepsilon_{t-q}$, и, следовательно, в целом с MA-ошибкой $\varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$, поскольку будущие возмущения с точки зрения дальних лагов y_t представляют собой инновации. Полный набор инструментальных переменных для точной идентификации будет иметь вид

$$\mathbf{X}_t, y_{t-q-1}, \ldots, y_{t-q-p}.$$

Такой подход удобно использовать для получения начальных приближений параметров модели ARMA с целью последующего уточнения оценок другими, более эффективными, методами.

Можно использовать в качестве инструментов также лаги экзогенных переменных $\mathbf{X}_t$. В частности, регрессия (10) для модели Койка может оцениваться с помощью инструментов 1 (константы), x_t и x_{t-1}. Умножая уравнение (10) при $t = 2, \ldots, T$ на соответствующие значения инструментов, суммируя и зануляя сумму произведений ошибки и инструментов, получим следующую систему уравнений (аналог формулы (15)):

$$\begin{aligned}
\sum_{t=2}^{T} y_t &= (T-2)\mu' + \alpha \sum_{t=2}^{T} x_t + \delta \sum_{t=2}^{T} y_{t-1}, \\
\sum_{t=2}^{T} x_t y_t &= \mu' \sum_{t=2}^{T} x_t + \alpha \sum_{t=2}^{T} x_t^2 + \delta \sum_{t=2}^{T} x_t y_{t-1}, \\
\sum_{t=2}^{T} x_{t-1} y_t &= \mu' \sum_{t=2}^{T} x_{t-1} + \alpha \sum_{t=2}^{T} x_{t-1} x_t + \delta \sum_{t=2}^{T} x_{t-1} y_{t-1}.
\end{aligned}$$

Решив эти уравнения, найдем оценки метода инструментальных переменных для модели Койка.[12]

4.3 Нелинейный метод инструментальных переменных

Ранее мы предполагали, что оцениваемое уравнение регрессии имеет линейную функциональную форму. Однако, довольно часто отношения между экономическими переменными описываются как нелинейные. Рассмотрим, как следует модифицировать метод инструментальных переменных, чтобы его можно было использовать для оценивания нелинейных регрессий, в которых есть проблема эндогенности. Это *нелинейный метод инструментальных переменных*. Для него используют аббревиатуру NLIV (англ. *nonlinear instrumental variables*).

Инструментальные переменные для нелинейного метода инструментальных переменных, как правило, естественно брать тоже нелинейные. Пусть, к примеру, требуется оценить нелинейную потребительскую функцию:

$$C_i = \alpha + \beta Y_i^{\gamma} + \varepsilon_i.$$

Предположим, что ε_i коррелирована с Y_i, но не с лагами Y_{i-1}, Y_{i-2} или функциями от этих лагов. Тогда можно взять в качестве инструментов Y_{i-1}, Y_{i-2} и, например, их квадраты (или какие-то степени, близкие к вероятному γ). Вопроса оптимального выбора функциональной формы инструментальных переменных мы здесь не будем касаться.

Когда в правой части уравнения регрессии стоят эндогенные переменные, то различие между левой и правой частью уравнения регрессии стирается. Как следствие, можно рассматривать более широкий класс моделей, в котором в левой части стоит некоторая функция от «объясняемой» переменной y, «объясняющих» переменных $\mathbf{x}$ и от параметров. Более того, можно рассматривать модели следующего довольно общего вида:

$$\boldsymbol{\varepsilon} = \mathbf{e}(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}),$$

где $\mathbf{y}$ отождествляется с зависимой переменной, $\mathbf{X}$ с регрессорами, а $\boldsymbol{\theta}$ – вектор неизвестных параметров. Относительно ошибки $\boldsymbol{\varepsilon}$ можно сделать стандартное предположение, что $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ и $\mathbb{E}[\boldsymbol{\varepsilon}^{\intercal}\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$.

Один из примеров такой нелинейной функции – это регрессия Бокса–Кокса:

$$h(y_i, \lambda) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

или

$$h(y_i, \lambda) = \beta_0 + \beta_1 h(x_i, \mu) + \varepsilon_i,$$

где

$$h(u, \lambda) = \begin{cases} (u^{\lambda} - 1)/\lambda, & \lambda \neq 0, \\ \ln(u), & \lambda = 0. \end{cases}$$

В качестве инструментов в такой регрессии (при экзогенности x_i) можно взять x_i или какие-то функции от x_i.

Пусть существует матрица инструментов $\mathbf{Z}$, для которой выполнено $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{Z}] = \mathbf{0}$ и $\mathbb{E}[\boldsymbol{\varepsilon}^{\intercal}\boldsymbol{\varepsilon}|\mathbf{Z}] = \sigma^2\mathbf{I}_n$. Нелинейный метод инструментальных переменных состоит в том, чтобы минимизировать по параметрам $\boldsymbol{\theta}$ следующую функцию:

$$\mathbf{e}(\boldsymbol{\theta})^{\intercal}\mathbf{Z}(\mathbf{Z}^{\intercal}\mathbf{Z})^{-1}\mathbf{Z}^{\intercal}\mathbf{e}(\boldsymbol{\theta}) = \mathbf{e}(\boldsymbol{\theta})^{\intercal}\mathbf{P}_{\mathbf{Z}}\mathbf{e}(\boldsymbol{\theta}), \tag{25}$$

[12] Ср. Liviatan (1963).

где, как и ранее, $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$ – матрица проекции на столбцы матрицы $\mathbf{Z}$. Обозначим через $\mathbf{E}(\boldsymbol{\theta})$ матрицу первых производных (по $\boldsymbol{\theta}$) функции $\mathbf{e}(\cdot)$. В этих обозначениях условие первого порядка минимума имеет вид

$$\mathbf{E}(\boldsymbol{\theta})^\intercal\mathbf{P_Z}\mathbf{e}(\boldsymbol{\theta}) = \mathbf{0}.$$

Минимизировать функцию (25) можно различными способами, но, по-видимому, наиболее удобный способ состоит в том, чтобы использовать следующую вспомогательную регрессию («искусственную регрессию» в терминологии Дэвидсона и Маккиннона, см. Davidson & MacKinnon, 2003):

$$\mathbf{e}(\boldsymbol{\theta}) \text{ на } -\mathbf{P_Z}\mathbf{E}(\boldsymbol{\theta}). \tag{26}$$

Имея некоторую текущую оценку параметров $\boldsymbol{\theta}$, следует построить такую регрессию и получить ней оценку коэффициентов. Эта оценка составит шаг итеративного алгоритма:

$$\Delta\boldsymbol{\theta} = -\big(\mathbf{E}(\boldsymbol{\theta})^\intercal\mathbf{P_Z}\mathbf{E}(\boldsymbol{\theta})\big)^{-1}\mathbf{E}(\boldsymbol{\theta})^\intercal\mathbf{P_Z}\mathbf{e}(\boldsymbol{\theta}).$$

Новая оценка получается как $\boldsymbol{\theta}' = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$. Итеративный алгоритм останавливается, когда R^2 во вспомогательной регрессии оказывается меньше малого положительного числа. Вспомогательная регрессия дает корректную оценку ковариационной матрицы оценок: $s^2\mathbf{E}^\intercal\mathbf{P_Z}\mathbf{E}$, где $s^2 = \mathbf{e}^\intercal\mathbf{e}/(n-m)$ (можно использовать в качестве оценки дисперсии ошибок σ^2 и $s^2 = \mathbf{e}^\intercal\mathbf{e}/n$, поскольку обе оценки одинаково хороши в асимптотическом смысле).

5 Проверка гипотез и диагностика

5.1 Проверка гипотез о коэффициентах

Как обычно, при проверке гипотез о коэффициентах в случае, когда оценки имеют асимптотически нормальное распределение, можно использовать тест Вальда. Статистика Вальда для проверки нулевой гипотезы, о том, что коэффициенты регрессии удовлетворяют k линейным ограничениям $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ (где $\mathbf{R}$ – матрица $k \times m$, $\mathbf{r}$ – вектор длиной k), равна

$$W = (\mathbf{R}\boldsymbol{\beta}_{\mathrm{IV}} - \mathbf{r})^\intercal\big(\mathbf{R}\hat{\mathbb{V}}(\boldsymbol{\beta}_{\mathrm{IV}})\mathbf{R}^\intercal\big)^{-1}(\mathbf{R}\boldsymbol{\beta}_{\mathrm{IV}} - \mathbf{r}).$$

Здесь, как и ранее, $\hat{\mathbb{V}}(\boldsymbol{\beta}_{\mathrm{IV}}) = s^2(\mathbf{X}^\intercal\mathbf{P_Z}\mathbf{X})^{-1}$ – это оценка ковариационной матрицы оценок $\boldsymbol{\beta}_{\mathrm{IV}}$. Эта статистика приближенно распределена как хи-квадрат со степенями свободы, равными числу ограничений, т. е. χ^2_k. Если W оказывается больше выбранного квантиля распределения χ^2_k, то нулевую гипотезу следует отвергнуть. При $k = 1$ более удобно использовать соответствующую z-статистику, т. е. статистику, которая приближенно распределена в соответствии со стандартным нормальным распределением, $N(0, 1)$. Если нулевая гипотеза состоит в том, что $\beta_j = \beta_j^*$, то z-статистика будет равна

$$z = \frac{\beta - \beta_j^*}{\sqrt{\hat{\mathbb{V}}(\boldsymbol{\beta}_{\mathrm{IV}})_{jj}}}.$$

Описанная в пункте 4.3 вспомогательная («искусственная») регрессия является удобным инструментом проверки гипотез в методе инструментальных переменных. Можно использовать стандартные t- и F-статистики из этой регрессии. Распределения этих статистик будут похожи на t- и F-распределения соответственно. Такие тесты являются модификациями описанного теста Вальда и асимптотически ему эквивалентны. В случае линейной регрессии $-\mathbf{P_Z}\mathbf{E}(\boldsymbol{\theta})$ в (26) принимает вид $\hat{\mathbf{X}}$.

5.2 Сверхидентифицирующие ограничения

Как проверить, что инструменты являются годными, т. е. что ошибки ε_i не коррелируют с инструментами $\mathbf{z}_i$? Вообще говоря, провести полную проверку этого условия невозможно, поскольку ошибки ε_i ненаблюдаемы. Но можно для проверки одних инструментов использовать другие инструменты. Для этого требуется иметь достаточно инструментов, необходима сверхидентификация.

Рассмотрим регрессию, которая включает *все* переменные, которые имеются в модели инструментальных переменных:

$$y = \mathbf{x}^{\circ}\boldsymbol{\beta}^{\circ} + \mathbf{x}^{\star}\boldsymbol{\beta}^{\star} + \mathbf{z}^{\dagger}\boldsymbol{\gamma}^{\dagger} + \varepsilon \tag{27}$$

или

$$y = \mathbf{x}^{\star}\boldsymbol{\beta}^{\star} + \mathbf{z}\boldsymbol{\gamma} + \varepsilon, \tag{28}$$

где $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^{\circ\intercal}, \boldsymbol{\gamma}^{\dagger\intercal}\right)^{\intercal}$. Конечно, эта регрессия неидентифицирована, поскольку в ней не остается ни одного свободного инструмента. Чтобы ее оценить, мы по сути накладываем на часть коэффициентов $\boldsymbol{\gamma}$ ограничения, зануляя их. А именно, мы полагаем $\boldsymbol{\gamma}^{\dagger} = \mathbf{0}$. Это условие, что внешние инструменты $\mathbf{z}^{\dagger}$ не нужны для объяснения y и не влияют на него непосредственно, помимо $\mathbf{x}$.

В случае сверхидентификации, т. е. когда количество инструментов больше количества регрессоров ($p > m$), мы фактически накладываем на вектор $\boldsymbol{\gamma}$ больше ограничений, чем требуется для оценивания. Эти «лишние» $p - m$ ограничений принято называть *сверхидентифицирующими ограничениями* (англ. *overidentifying restrictions*). Можно проверить, выполнены ли эти дополнительные ограничения, или, иначе, нужны ли избыточные инструменты в регрессии. Пусть $\mathbf{z}^{\ddagger}$ – это некоторые $p - m$ из $p^{\dagger} = p - m^{\circ}$ внешних инструментов $\mathbf{z}^{\dagger}$. В этих обозначениях можем записать модель, которая точно идентифицирована при использовании $\mathbf{z}$ в качестве инструментальных переменных:

$$y = \mathbf{x}\boldsymbol{\beta} + \mathbf{z}^{\ddagger}\boldsymbol{\gamma}^{\ddagger} + \text{ошибка}.$$

Нулевая гипотеза для *теста на сверхидентифицирующие ограничения* состоит в том, что $\boldsymbol{\gamma}^{\ddagger} = \mathbf{0}$. Ее следует проверять так же, как обычно проверяются ограничения на коэффициенты регрессии с инструментальными переменными (см. пункт 5.1).

Однако, есть более простой способ проверки сверхидентифицирующих ограничений. Он состоит в следующем. Исходная модель (т. е. модель при $\boldsymbol{\gamma}^{\ddagger} = 0$) оценивается методом инструментальных переменных и из нее берутся остатки $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. Далее строится регрессия остатков $\mathbf{e}$ на инструменты $\mathbf{Z}$. При нулевой гипотезе, что ошибки не коррелируют с инструментами, эта регрессия должна иметь незначительную объясняющую силу, что можно измерить с помощью (нецентрального[13]) коэффициента детерминации R_u^2. Таким образом, имеет смысл строить статистику для проверки гипотезы об экзогенности регрессоров на основе R_u^2. Как правило, используют статистику следующего вида:

$$nR_u^2.$$

Эта статистика имеет асимптотически распределение хи-квадрат с числом степеней свободы, равным числу «лишних» инструментов (сверхидентифицирующих ограничений), т. е. χ^2_{p-m}. Если статистика большая (больше критической границы), то следует сделать вывод, что среди инструментов есть неэкзогенные, т. е. негодные. Тест на сверхидентифицирующие ограничения был предложен в статье Basmann (1960).

[13]Т.е. такого, что в качестве знаменателя используется сумма квадратов *нецентрированной* зависимой переменной.

Следует отдавать себе отчет, что невозможно проверить годность каждого отдельного инструмента. Инструменты проверяются взаимно, друг относительно друга и только за счет сверхидентификации. В случае точной идентификации проверка годности инструментов принципиально невозможна, но если имеет место сверхидентификация, то можем подстраховаться. Таким образом, эта статистика является полезным инструментом диагностики. Но в любом случае то, что по крайней мере m инструментов являются экзогенными – это наша априорная гипотеза, которая не поддается проверке.

5.3 Тест Хаусмана

Если в уравнении (12) переменные $\mathbf{x}^\star$ не коррелированы с ошибкой, то не имеет смысла использовать метод инструментальных переменных с $\mathbf{z} = (\mathbf{x}^\circ, \mathbf{z}^\dagger)$ в качестве инструментов, а лучше оценить модель обычным МНК. Оценки обычного МНК $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ при экзогенности регрессоров являются более точными,[14] чем $\hat{\boldsymbol{\beta}}_{\text{IV}}$, поэтому, если нет опасности несостоятельности, то лучше использовать именно их. Таким образом, полезно уметь определять, когда следует использовать метод инструментальных переменных, а когда можно обойтись обычным МНК. Для этого можно использовать *тест Хаусмана*. (Его еще называют тестом Дарбина–Ву–Хаусмана, DWH, или тестом на эндогенность.[15])

Для того чтобы понять идею этого теста, рассмотрим линейную проекцию $\mathbf{x}^\star$ на $\mathbf{z}$, которая уже была введена выше (уравнение (22)):

$$\mathbf{x}^\star = \mathbf{z}\boldsymbol{\Lambda} + \mathbf{v}. \tag{29}$$

Это представление позволяет переформулировать условие применимости обычного МНК: с учетом того, что инструменты $\mathbf{z}$ некоррелированы с ошибкой исходного уравнения ε ($\mathbb{E}[\mathbf{z}\varepsilon] = \mathbf{0}$, инструменты экзогенны), переменные $\mathbf{x}^\star$ некоррелированы с ошибкой тогда и только тогда, когда ошибки $\mathbf{v}$ некоррелированы с ε, т. е. когда $\mathbb{E}[\mathbf{v}\varepsilon] = \mathbf{0}$.

Далее, введем линейную проекцию ε на $\mathbf{v}$:

$$\varepsilon = \mathbf{v}\boldsymbol{\tau} + r \tag{30}$$

По определению линейной проекции условие $\mathbb{E}[\mathbf{v}\varepsilon] = \mathbf{0}$ эквивалентно тому, что $\boldsymbol{\tau} = \mathbf{0}$. Подставив представление ошибки ε из (30) в исходное уравнение (12), получим:

$$y = \mathbf{x}^\circ\boldsymbol{\beta}^\circ + \mathbf{x}^\star\boldsymbol{\beta}^\star + \mathbf{v}\boldsymbol{\tau} + r = \mathbf{x}\boldsymbol{\beta} + \mathbf{v}\boldsymbol{\tau} + r \tag{31}$$

Обычный МНК применим к исходной регрессии тогда и только тогда, когда $\boldsymbol{\tau} = \mathbf{0}$. Ясно, что мы не можем оценить (31) непосредственно, поскольку ошибки уравнения (29) $\mathbf{v}$ ненаблюдаемы. Однако, мы можем получить оценки этих величин из регрессии первого шага двухшагового МНК (17), т. е. из регрессии, соответствующей модели (29).

Эти рассуждения подсказывают следующую процедуру проверки применимости обычного МНК.

1. Строим регрессию $\mathbf{X}^\star$ на инструментальные переменные $\mathbf{Z}$ и берем из этой регрессии остатки $\hat{\mathbf{V}} = \mathbf{X}^\star - \hat{\mathbf{X}}^\star = \mathbf{M}_{\mathbf{Z}}\mathbf{X}^\star$.

2. По аналогии с уравнением (31) строим регрессию $\mathbf{y}$ на $\mathbf{X}$ и $\hat{\mathbf{V}}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{V}}\boldsymbol{\tau} + \text{ошибка}.$$

[14] Это связано с тем, что разность $(\mathbf{X}^\intercal\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1} - (\mathbf{X}^\intercal\mathbf{X})^{-1}$ является положительно полуопределенной.
[15] Имеется в виду эндогенность переменных $\mathbf{x}^\star$.

3. Используем стандартную F-статистику[16] для проверки гипотезы $\boldsymbol{\tau} = \mathbf{0}$ в этой последней регрессии (т. е. гипотезы, что коэффициенты при $\hat{\mathbf{V}}$ равны нулю). Если добавленные переменные $\hat{\mathbf{V}}$ оказываются незначимыми,[17] то следует принять нулевую гипотезу $\mathbb{E}[\mathbf{x}^\star\varepsilon] = \mathbf{0}$ и использовать обычный МНК для оценивания регрессии (12). В противном случае следует использовать метод инструментальных переменных.

Альтернативно, тест Хаусмана (и в этом состояла исходная идея Хаусмана, см. Hausman, 1978) может быть построен на сравнении оценок $\hat{\boldsymbol{\beta}}_{\text{IV}}$ с $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. Статистика Хаусмана основана на разности $\Delta = \hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}}$ и имеет вид квадратичной формы:

$$H = \Delta^\intercal \hat{\mathbb{V}}(\Delta)^{-1}\Delta,$$

где $\hat{\mathbb{V}}(\Delta)$ – оценка ковариационной матрицы разности Δ. Оказывается, что в данном случае $\hat{\mathbb{V}}(\Delta)$ имеет вид

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{IV}}) - \hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{OLS}}),$$

где $\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{IV}})$ – оценка ковариационной матрицы оценок $\hat{\boldsymbol{\beta}}_{\text{IV}}$, а $\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$ – оценка ковариационной матрицы оценок $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. Оценки ковариационных матриц можно рассчитать по формулам

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = s^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$$

и

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = s^2(\mathbf{X}^\intercal\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}.$$

Здесь s^2 – это оценка дисперсии ошибки. В качестве s^2 можно брать остаточную дисперсию, рассчитанную на основе либо оценок обычного МНК (s^2_{OLS}), либо метода инструментальных переменных (s^2_{IV}). Первый вариант (предложенный Дарбином, Durbin, 1954) более предпочтителен, поскольку использует более точную оценку и более устойчив к проблеме слабых инструментов.

Статистика Хаусмана имеет асимптотическое распределение хи-квадрат с $m^\star$ степенями свободы. Если статистика H большая (больше критической границы), то следует сделать вывод, что $\mathbb{E}[\mathbf{v}\varepsilon] \neq \mathbf{0}$. Если мы уверены в том, что инструментальные переменные $\mathbf{z}$ экзогенны, то это означает, что регрессоры $\mathbf{x}^\star$ коррелированы с ошибкой и мы должны использовать для оценивания метод инструментальных переменных. Если H мала, то оправдано использование обычного МНК.

6 Слабые инструменты

Выше было указано, что одно из требований к инструментам состоит в том, чтобы они были релевантными (достаточно сильно связанными с регрессорами) или, другими словами, не являлись слабыми. Желательно дать более точное определение данным понятиям. Прежде всего ясно, что переменные $\mathbf{x}^\circ$ по своей сути релевантны, поскольку они сами являются регрессорами. Таким образом, имеет смысл обсуждать релевантность только внешних регрессоров $\mathbf{z}^\dagger$. При этом имеет смысл говорить о связи внешних инструментов с эндогенными регрессорами $\mathbf{x}^\star$, причем о связи *в чистом виде, не опосредованной экзогенными регрессорами* $\mathbf{x}^\circ$. Так, если имеется единственная эндогенная переменная и единственный внешний инструмент, то речь должна идти о величине частной корреляции между ними относительно $\mathbf{x}^\circ$ (то есть корреляции между ними после устранения общих составляющих, объясняемых взаимодействием с $\mathbf{x}^\circ$).

[16] Если эндогенная переменная только одна, то можно использовать соответствующую t-статистику.

[17] Несложно понять, что те же результаты получатся, если добавлять $\hat{\mathbf{X}}^\star$, а не $\hat{\mathbf{V}}$, и проверять равенство нулю коэффициентов при $\hat{\mathbf{X}}^\star$.

Здесь удобно рассуждать в терминах регрессии первого шага двухшагового МНК (17). Если мы оценили эту регрессию и получили оценку $\hat{\mathbf{\Lambda}}^{\dagger}$ равную нулю, то, хотя внешние инструменты могут быть коррелированы с эндогенными регрессорами, но они ничего нового не дают по сравнению с экзогенными регрессорами, так что матрица проекции для всех инструментов, $\mathbf{P}_{\mathbf{Z}}$, совпадет с матрицей проекции для экзогенных регрессоров, $\mathbf{P}_{\mathbf{X}^\circ} = \mathbf{X}^\circ(\mathbf{X}^{\circ\intercal}\mathbf{X}^\circ)^{-1}\mathbf{X}^{\circ\intercal}$. В этом случае внешние инструменты фактически отсутствуют, и мы получаем неидентифицированное уравнение регрессии. Это случай, когда внешние инструменты являются полностью нерелевантными. На практике матрица $\hat{\mathbf{\Lambda}}^{\dagger}$ хотя и отличается от нуля, но часто оказывается очень малой. В этой ситуации, когда оценка $\hat{\mathbf{\Lambda}}^{\dagger}$ в определенном смысле близка к нулю, инструменты естественно назвать слабыми.

Такое определение слабых инструментов характеризует конкретные данные, с которыми мы имеем дело. Можно рассмотреть также его теоретический аналог. Для этого следует ввести теоретический аналог уравнения (17):

$$\mathbf{x}^\star = \mathbf{z}\mathbf{\Lambda} + \mathbf{v} = \mathbf{x}^\circ\mathbf{\Lambda}^\circ + \mathbf{z}^\dagger\mathbf{\Lambda}^\dagger + \mathbf{v}. \tag{32}$$

С точки зрения этого уравнения инструменты будут слабыми в том случае, когда матрица теоретических коэффициентов $\mathbf{\Lambda}^\dagger$ в определенном смысле близка к нулю.

Известно, что если инструменты слабые, то это влечет много неприятных последствий для метода инструментальных переменных.

- Оценки метода инструментальных переменных оказываются неточными, и это проявляется в больших стандартных ошибках коэффициентов и широких доверительных интервалах, рассчитанных на основе оценки ковариационной матрицы (21). (Эта проблема тесно связана с мультиколлинеарностью в регрессии второго шага (18).)

- Стандартное асимптотическое приближение, согласно которому распределение разности оценок и истинных коэффициентов можно аппроксимировать нормальным распределением с нулевым математическим ожиданием и ковариационной матрицей (21), работает плохо. Во-первых, распределение оценок может очень существенно отличаться от нормального. Во-вторых, оценки метода инструментальных переменных могут быть сильно смещены. В-третьих, как правило, обычные доверительные области оказываются слишком «оптимистическими», преувеличивая точность оценок. Как следствие, обычные асимптотические доверительные области для параметров модели и оценки распределений тестовых статистик могут сильно «врать», т. е. их фактические коэффициенты покрытия (англ. *coverage rates*), могут существенно отличиться от номинального асимптотического уровня доверия.

- Когда инструменты слабые, оценки по методу инструментальных переменных смещены по направлению к (несостоятельным!) оценкам обычного метода наименьших квадратов. Когда инструменты очень слабые, то смещение двух видов оценок становится очень похожим. При этом вполне может возникнуть ситуация, когда смещение оценок по методу инструментальных переменных практически такое же, как и у оценок обычного МНК, а «разброс» оценок существенно больше, так что использование метода инструментальных переменных приводит к существенной потере точности оценок по сравнению с обычным МНК.

- Если инструменты слабые, то даже небольшое нарушение предположения о некоррелированности инструментальных переменных и ошибки регрессии может приводить к очень существенной несостоятельности оценок метода инструментальных переменных.

К этому следует добавить, что указанные проблемы могут усиливаться, когда при оценивании используется большое количество слабых инструментальных переменных, хотя согласно

стандартной асимптотической теории добавление инструментов должно улучшать точность оценок.

Формально слабые в асимптотическом смысле инструменты можно задать как ситуацию, когда в регрессии первого шага матрица коэффициентов при внешних инструментах не является постоянной, а стремится к нулю со скоростью $\sqrt{n}$. Более точно, матрица $\mathbf{\Lambda}^\dagger$ в уравнении (32) моделируется как $\mathbf{\Lambda}^\dagger = \mathbf{C}/\sqrt{n}$, где $\mathbf{C}$ – постоянная матрица (см. Staiger & Stock, 1997). Такой альтернативный подход гарантирует, что инструменты остаются слабыми при стремлении количества наблюдений n к бесконечности. Он позволяет получить более корректную асимптотическую теорию. (Упрощенное изложение соответствующей теории можно найти в Анатольев, 2005, раздел 5.)

Один из способов решения проблемы слабых инструментов – построение доверительных областей, которые не основаны на стандартной асимптотической ковариационной матрице и асимптотической нормальности. Такие доверительные области могут быть, в частности, построены в рамках LIML на основе статистики Андерсона–Рубина (23). Этот и другие подобные подходы пока достаточно сложно использовать на практике.

Основной способ проверки того, являются ли инструменты слабыми, состоит в анализе коэффициентов детерминации и F-статистик в регрессиях (16) на первом шаге двухшагового МНК. Смотрим регрессии первого шага – насколько велики t- и F-статистики для проверки гипотез о равенстве коэффициентов при внешних инструментах нулю, насколько велики частные R^2 – влияние инструментов $\mathbf{Z}^\dagger$ на $\mathbf{X}^\star$ помимо $\mathbf{X}^\circ$ («очищенное» от этого влияния). Эмпирическим измерителем силы инструментов является F-статистика для гипотезы $\Lambda^\dagger = \mathbf{0}$ в регрессии первого шага. При $p = 1$ (единственный внешний инструмент) можно использовать t-статистику. Грубое рабочее правило для единственного эндогенного регрессора ($m^\star = 1$) состоит в том, что F-статистика меньше 10 должна вызывать озабоченность.

К сожалению, если имеется более одного эндогенного регрессора ($m^\star > 1$), то рассмотрения отдельных регрессий (16) может оказаться недостаточно. Cragg & Donald (1993) в этом случае предложили использовать матричный аналог F-статистики для гипотезы $\Lambda^\dagger = \mathbf{0}$:

$$\hat{\Sigma}_{vv}^{-1/2}\mathbf{X}^{\star\intercal}(\mathbf{M}_{\mathbf{X}^\circ} - \mathbf{M}_{\mathbf{Z}})\mathbf{X}^\star\hat{\Sigma}_{vv}^{-1/2}/p^\dagger,$$

где $\mathbf{M}_{\mathbf{X}^\circ}$ и $\mathbf{M}_{\mathbf{Z}}$ имеют то же значение, что и выше, а $\hat{\Sigma}_{vv} = \hat{\mathbf{V}}^\intercal\hat{\mathbf{V}}/(n-p) = \mathbf{X}^{\star\intercal}\mathbf{M}_{\mathbf{Z}}\mathbf{X}^\star/(n-p)$. Нулевую гипотезу о нерелевантности инструментов предлагается проверять на основе статистики, равной минимальному собственному значению этой матричной статистики.

В заключение упомянем характерный (и ставший хрестоматийным) пример исследования, в котором используются слабые инструменты. Это статья Angrist & Krueger (1991), в которой изучалась отдача от образования в духе минцеровской регрессии (3). Ангрист и Крюгер заметили, что из-за особенностей законодательства США, относящегося к обязательному образованию, длительность обучения человека зависит от того, в какое время года он родился. По-видимому, квартал рождения случайно распределен между людьми и не зависит от того, в какой обстановке родился индивидуум, какие у него были родители, и т. д. Если это так, и квартал рождения влияет на заработки только через время обучения, то он является годным инструментом. Таким образом, ситуация, фактически, похожа на естественный эксперимент. Хотя эффект зависимости от квартала не очень сильный, но, поскольку имелось очень большое количество наблюдений, его все же можно уловить. Он хорошо виден из Рис. 2 (слева) – данные там усреднены по людям, родившимся в конкретный квартал конкретного года. В зарплате человека также видна зависимость от времени года, в которое он родился (см. Рис. 2 справа). Таким образом, статистическая значимость связи и, следовательно, релевантность инструментов не вызывает сомнений.

В типичной постановке в качестве внешних инструментов использовались фиктивные переменные для квартала рождения (3 переменных) и фиктивные переменные для сочетания квартала рождения и года рождения ($3 \times 9 = 27$ переменных). Две переменных исключа-

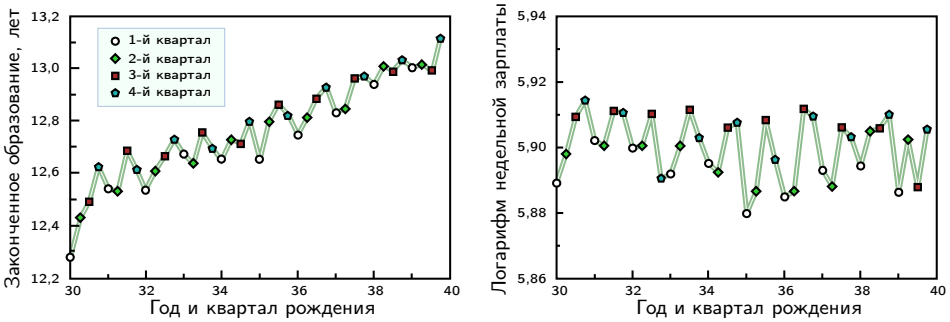

Рис. 2: Иллюстрации к статье Angrist & Krueger (1991) – обоснование использования квартала рождения в качестве инструмента

лись из-за линейной зависимости, когда в модель включались в качестве объясняющих переменных возраст и возраст в квадрате. В этом случае всего было использовано 28 внешних инструментов.

Bound, Jaeger & Baker (1995) обратили внимание на слабость использованных Ангристом и Крюгером инструментов. При использовании данных о 329509 мужчинах 1930–1939 годов рождения в регрессии первого шага (зависимая переменная – образование, регрессоры – все имеющиеся инструментальные переменные) частный R-квадрат для внешних инструментов равен 0,014%, а F-статистика для гипотезы о равенстве нулю коэффициентов при внешних инструментах составила 1,613 при уровне значимости 0,021. Таким образом, F-статистика близка к своему ожидаемому значению 1. Баунд, Джегер и Бейкер указали на то, что если заменить инструменты, использованные в статье, на случайным образом сгенерированные (и не имеющие никакого отношения к данным), то результаты будут похожи те, что представлены в этой статье.

Ангрист и Крюгер в регрессии логарифма зарплаты на образование (измеряемое годами обучения) и другие переменные получили оценку 0,0632 для коэффициента при образовании (аналог коэффициента β_1 в уравнении (3)) со стандартной ошибкой 0,0003 при использовании обычного МНК, и оценку 0,0600 со стандартной ошибкой 0,0299 при использовании метода инструментальных переменных. Баунд, Джегер и Бейкер в серии из 500 экспериментов со случайно сгенерированными «кварталами» получили средний коэффициент 0,061 и среднеквадратическое отклонение коэффициента 0,039. Учитывая низкое значение F-статистики, можно сделать вывод, что не представляет труда сымитировать результаты Ангриста и Крюгера, пользуясь полностью нерелевантными инструментами.

Staiger & Stock (1997) также рассмотрели эти данные и убедились, применив разработанные ими асимптотические методы, что инструменты действительно очень слабые, и что использование обычного асимптотического приближения в данном случае не оправдано. Полученная ими альтернативная 95%-я интервальная оценка на основе статистики Андерсона–Рубина для коэффициента при образовании по тем же данным и той же модели равна $[-0{,}441; 0{,}490]$. Такая неточная оценка вряд ли может представлять какой-то интерес.

По-видимому, эти результаты (помимо демонстрации применения современных методов для слабых инструментов) в основном говорят о том, что не следует включать в регрессию большое количество очень слабых инструментов. Для упомянутых данных (329509 мужчин 1930–1939 годов рождения) зависимость длительности обучения от квартала очевидна. В то же время, авторы первоначальной статьи необоснованно добавили большое количество дополнительных инструментов, наличие связи которых с длительностью обучения никак не подтверждается имеющимися данными – это фиктивные переменные для взаимодействия года рождения и квартала рождения. Включение этих переменных имело бы смысл, если бы сезонность на Рис. 2 очевидным образом менялась бы в зависимости от года рождения. Но

никакого изменения структуры сезонности не заметно. Таким образом, эти дополнительные инструменты не дают ничего, кроме дополнительного шума.

7 Пример использования метода инструментальных переменных

Статья Cutler & Glaeser (1997) представляет собой типичный пример использования инструментальных переменных в исследовании по прикладной микроэкономике. Авторы поставили задачу выяснить, как влияет расовая сегрегация на благосостояние чернокожих американцев. На уровне отдельных индивидуумов наблюдается закономерность, что чернокожие, живущие в окружении чернокожих, менее успешны, чем чернокожие, живущие в преимущественно белом окружении. Но такая связь может объясняться тем, что происходит самоотбор успешных чернокожих, перемещение их в более благополучные районы с преимущественно белым населением. Чтобы избежать этой проблемы, Катлер и Глэсер решили использовать в регрессии данные по отдельным индивидуумам, но в качестве объясняющей переменной взять данные о сегрегации в среднем по городу, в котором живет индивидуум. Другими словами, они изучали вопрос о том, более или менее успешны расовые меньшинства *в целом* в тех городах, где расовая сегрегация более сильна, по сравнению с теми городами, где расовая сегрегация менее сильна.

Базовая модель регрессии в статье имеет вид

$$Succ = \beta_1 Segr + \beta_2 Segr \cdot Black + \textit{прочие факторы} + \varepsilon.$$

Здесь *Succ* – показатель успешности для индивидуума, *Segr* – измеритель сегрегации в городе, *Black* – фиктивная переменная для чернокожих. Коэффициент β_1 измеряет влияние сегрегации на белых, а коэффициент β_2 – разницу между влиянием сегрегации на чернокожих и на белых. Для авторов наибольший интерес представлял коэффициент β_2. Использованные данные[18] относятся к 1990 г.

При указанном подходе к измерению сегрегации все еще остается проблема самоотбора (в данном случае уже из-за перемещения между городами), но проблема становится менее острой. Появляется также проблема обратной причинности: меньшая успешность чернокожих может вести к большей сегрегации. Чтобы решить проблему обратной причинности, авторы использовали инструментальные переменные для сегрегации. Инструменты подбирались так, чтобы они оказывали влияние на сегрегацию, но чтобы показатели успешности чернокожих не оказывали на них влияния. Использованы два разных набора инструментов.

Первый набор инструментальных переменных отражает структуру местных финансов. Используются два таких инструмента: количество местных органов власти, входящих в данный город, и доля доходов местного бюджета, поступающих из бюджетов более высокого уровня (уровня штата и федерального правительства). Авторы исходили из того, что количество местных органов власти может влиять на сегрегацию через механизм Тибу: когда имеется много органов власти, ставки налогов и уровень муниципальных услуг сильнее варьируются в пределах города, что способствует сегрегации. Аналогично, если меньше денег приходит «сверху», то местные налоги должны быть выше, так что расовые меньшинства в большей степени заинтересованы в том, чтобы выиграть на образующейся разнице ставок налогов, что усиливает сегрегацию.

Количество местных органов власти в пределах района незначительно меняется с течением времени, поэтому его можно считать экзогенным для сегрегации. Чтобы снять все подозрения по поводу направления причинности, авторы использовали данные за 1962 г. в качестве инструмента. Доля доходов местных бюджетов, которая приходит от властей штата и федерального правительства, может быть коррелирована с местными условиями. Чтобы удалить местную эндогенную составляющую из этого показателя, авторы берут среднее по штату, а

[18] Данные можно найти на странице *http://trinity.aas.duke.edu/~jvigdor/segregation/index.html*.

не по отдельному городу. При таком выборе показателя он оказывается связанным с политическими характеристиками штата, а не отдельного города. Данные для этого инструмента также взяты за 1962 г., чтобы еще в большей степени ослабить эндогенность. (Этот показатель заметно менялся со временем, но все же корреляция между уровнями 1962 г. и 1987 г. года достаточно высокая: $\rho = 0{,}55$).

В регрессии сегрегации на эти две инструментальные переменные коэффициент детерминации равен $R^2 = 31{,}2\%$, t-статистики равны 8,8 и $-2{,}4$ соответственно[19] (это регрессия по городам). Поскольку в базовую регрессию включена сегрегация и сегрегация, умноженная на фиктивную переменную для расы, то фискальные инструментальные переменные также берутся как сами по себе, так и их произведения с фиктивной переменной для расы.

Второй набор инструментов основан на топографических особенностях города – это количество рек, протекающих внутри округов, и количество рек, протекающих между округами. Использование этих двух показателей в качестве инструментов для сегрегации объясняется тем, что реки служат естественными преградами, разделяя города на части и создавая предпосылки для сегрегации.[20] Были включены также квадраты количества рек, чтобы учесть возможные нелинейные связи между количеством рек и сегрегацией. В регрессии сегрегации на указанные четыре «топографические» инструментальные переменные коэффициент детерминации оказался равным $R^2 = 19{,}8\%$.

Трудно сказать, достаточно ли «сильные» инструменты использованы в этом исследовании, но очевидно, что они значимо связаны с сегрегацией. Поскольку инструментальных переменных много, а эндогенных переменных всего две, то для контроля качества инструментов можно использовать тест на сверхидентифицирующие ограничения (см. пункт 5.2). Действительно, авторы статьи провели такой тест, и оказалось, что инструменты не прошли проверки. Это может объясняться тем, что нарушены предположения, лежащие в основе теста на сверхидентифицирующие ограничения, например, имеет место автокорреляция ошибок регрессии. Но в целом это заставляет с определенной настороженностью относится к результатам исследования.

Используя самые разные переменные в качестве показателя успешности индивидуумов, Катлер и Глэсер обнаружили, что чернокожие существенно менее успешны в расово сегрегированных городах: уменьшение сегрегации на одно среднеквадратическое отклонение уменьшает примерно на одну треть различие между черными и белыми по большинству из показателей.

8 Резюме

- Возможные причины корреляции между регрессорами и ошибкой регрессии – это пропущенные переменные, ошибки в переменных, двусторонняя причинность, использование лага зависимой переменной в условиях автокорреляции.

- Если регрессор и ошибка коррелированы, обычный метод наименьших квадратов дает несостоятельные и асимптотически смещенные оценки. В этом случае метод инструментальных переменных позволяет получить состоятельные оценки.

- Инструментальная переменная не должна быть коррелирована с ошибкой и должна быть коррелирована с эндогенными переменными оцениваемой регрессии.

- Метод инструментальных переменных, фактически, лежит в основе некоторых известных методов статистического оценивания. Метод инструментальных переменных (с мо-

[19] В статье t-статистики не приведены, а приведены стандартные ошибки.
[20] Этот прием первоначально использовала К. Хоксби. Ср. Hoxby (2000).

дификациями) может использоваться для получения состоятельных оценок в самых разных эконометрических моделях.

- Слабые инструменты могут приводить к серьезным проблемам при точечном и интервальном оценивании и при проверке гипотез.
- Поиск годных и релевантных инструментов – это задача, которая требует большой изобретательности. Найти однозначно годные и релевантные инструменты – это большая удача.

9 Дальнейшее чтение

Материал, близкий к тому, что обсуждается в данном эссе, содержится практически во всех эконометрических учебниках. В частности, можно порекомендовать Hayashi (2000, гл. 3), Davidson & MacKinnon (2003, гл. 8), Ruud (2000, гл. 20), Wooldridge (2002, гл. 4, 5 и 6), Cameron & Trivedi (2005, гл. 4). Существует монография Bowden & Turkington (1984), посвященная инструментальным переменным (хотя и несколько устаревшая).

Разнообразные сведения о методе инструментальных переменных можно почерпнуть из статей Angrist & Krueger (2001), Stock (2001), Baum, Schaffer & Stillman (2003). Обзор оценивания в условиях слабости инструментов можно найти в Stock, Wright & Yogo (2002) и Паган (2007).

Список литературы

Анатольев, С.А. (2005). Асимптотические приближения в современной эконометрике. *Экономика и математические методы* 41, 84–94.

Паган, А. (2007). Слабые инструменты. *Квантиль* 2, 71–81.

Anderson, T.W. & H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Angrist, J.D. & A.B. Krueger (1991). Does compulsory school attendance affect schooling and earning? *Quarterly Journal of Economics* 106, 979–1014.

Angrist, J.D. & A.B. Krueger (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15, 69–85.

Basmann, R.L. (1957). A general classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, 77–83.

Basmann, R.L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55, 650–659.

Baum, C.F., M.E. Schaffer & S. Stillman (2003). Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3, 1–31.

Bound, J., D.A. Jaeger & R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–450.

Bowden, R. & D. Turkington (1984). Instrumental Variables. Cambridge: Cambridge University Press.

Cameron, A.C. & P.K. Trivedi (2005). Microeconometrics: Methods and Applications. Cambridge: Cambridge University Press.

Cragg, J.G. & S.G. Donald (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory* 9, 222–240.

Cutler, D.M. & E.L. Glaeser (1997). Are ghettos good or bad? *Quarterly Journal of Economics* 112, 827–872.

Davidson, R. & J.G. MacKinnon (2003). Econometric Theory and Methods. Oxford University Press.

Durbin, J. (1954). Errors in variables. *Review of Institute of International Statistics* 22, 23–54.

Evans, W.N., L.M. Froeb & G.J. Werden (1993). Endogeneity in the concentration–price relationship: Causes, consequences, and cures. *Journal of Industrial Economics* 41, 431–438.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.

Hansen, L.P. & K.J. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–1286.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–1271.

Hayashi, F. (2000). Econometrics. Princeton: Princeton University Press.

Hoxby, C.M. (2000). Does competition among public schools benefit students and taxpayers? *American Economic Review* 90, 1209–1238.

Liviatan, N. (1963). Consistent estimation of distributed lags. *International Economic Review* 4, 44–52.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy* 66, 281–302.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–23.

Ruud, P.A. (2000). An Introduction to Classical Econometric Theory. New York: Oxford University Press.

Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

Staiger, D. & J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J.H. (2001). Instrumental variables in statistics and econometrics. Глава в *International Encyclopedia of the Social & Behavioral Sciences* под редакцией N.J. Smelser & P.B. Baltes, 7577–7582. Amsterdam: Elsevier.

Stock, J.H., J. Wright & M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* 20, 518–529.

Theil, H. (1953). Repeated least-squares applied to complete equation systems. Discussion paper, Central Planing Bureau, Hague.

Wooldridge, J. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT Press.

Working, E.J. (1927). What do statistical "demand curves" show? *Quarterly Journal of Economics* 41, 212–235.

Wright, P.G. (1928). The Tariff on Animal and Vegetable Oils. New York: Macmillan.

A guide to the world of instruments

Alexander Tsyplakov

Novosibirsk State University, Novosibirsk

The essay discusses reasons of correlatedness of explanatory variables and errors in regression applications, consequences of this correlatedness, and the method of instrumental variables aimed to resolve this problem.

Оценивание структурных эконометрических уравнений*

Стефен Поллок†

Колледж королевы Мэри Лондонского университета, Великобритания

Настоящая работа содержит отличающиеся от традиционных выводы ММПОИ- и 2ШМНК-оценок одиночных уравнений классической системы одновременных эконометрических уравнений. Принадлежность обеих к оценкам метода моментов подчеркивает их глубинные сходства. Оценка ММПОИ выводится из критерия наименьших квадратов, использующего интерпретацию структурного уравнения как модели с ошибками в переменных, а 2ШМНК-оценка получена при помощи асимптотически правомерного приближения. Оценку ММПОИ можно вычислить с помощью итерационного алгоритма, стартующего с оценки 2ШМНК. В работе также рассматриваются традиционные выводы 2ШМНК-оценки.

1 Введение

Задача оценивания одиночного уравнения классической эконометрической системы одновременных уравнений была решена исследователями из *Cowles Commission*, разработавшими метод максимального правдоподобия с ограниченной информацией (ММПОИ). Они представили два способа вывода этой оценки.

Первый вывод был произведен Anderson & Rubin (1949) на основе функции правдоподобия для модели одновременных уравнений в приведенной форме. Добавив в нее информацию, связанную с одиночным структурным уравнением, они нашли оценки параметров приведенной формы при ограничениях вместе с оценками параметров структурного уравнения.

В альтернативном выводе Koopmans, Rubin & Leipnick (1950) за отправную точку взяли функцию правдоподобия для структурных параметров полной системы. Они вывели целевую функцию для оценивания интересующего их уравнения путем удаления посторонних параметров через частичную максимизацию.

Эти два способа вывода ММПОИ должны были получить признание среди прикладных эконометристов. Однако обнаружились препятствия, помешавшие всеобщему принятию этой оценки. Первоначальные способы ее получения были слишком длинными и сложными, и немногие были способны овладеть ими. К тому же вычисление оценок подразумевало итерационную процедуру получения скрытых корней, с которой существовавшие на тот момент компьютеры едва справлялись.

Обе задачи были решены несколько лет спустя с помощью оценки двухшагового метода наименьших квадратов (2ШМНК), независимо полученной в Basmann (1957) и Theil (1958) простыми и понятными способами. Выводы обеих оценок производились в рамках всем знакомой классической модели линейной регрессии. Они строились в попытках найти простые пути преодоления той проблемы, из-за которой МНК в обычной линейной регрессии неверно оценивает структурные параметры.

Позже Theil (1961) сумел продемонстрировать сходство 2ШМНК- и ММПОИ-оценок, показав, что они обе являются элементами определенного им «k-класса оценок». Другие авторитетные ученые, включая Malinvaud (1966), также смогли показать это сходство. Несмотря

*Перевод А. Шамгунова и С. Анатольева. Цитировать как: Поллок, Стефен (2007) «Оценивание структурных эконометрических уравнений», Квантиль, №2, стр. 49–59. Citation: Pollock, Stephen (2007) "Estimation of Structural Econometric Equations," Quantile, No.2, pp. 49–59.

†Адрес: Department of Economics, Queen Mary College, University of London, Mile End Road, London E1 4NS, United Kingdom. Электронная почта: *d.s.g.pollock@qmul.ac.uk*

на это, ММПОИ-оценка не была широко признана и часто представлялась без соответствующего вывода.

Недавно в работе, касающейся происхождения оценок 2ШМНК и ММПОИ, Anderson (2005) поведал, как Anderson & Rubin (1950) получили асимптотическое распределение ММПОИ-оценки путем нахождения асимптотического распределения статистики, по своей сути являвшейся оценкой 2ШМНК. Несмотря на то, что эта работа дает краткий отчет о первом выводе Anderson & Rubin (1949), она не проводит прямой параллели между обеими оценками.

Остается желание продемонстрировать близкую сущность этих оценок, что подразумевает прямой вывод обеих. Это и является целью данной работы.

2 Структурные уравнения

Классическая модель линейных одновременных эконометрических уравнений – это стохастическая система, связывающая M выходных эндогенных переменных с K входными экзогенными переменными. Особенностью этой модели является то, что каждая зависимая переменная вектора-строки $y_{t\bullet} = [y_{t1}, y_{t2}, \ldots, y_{tM}]$ является функцией не только K экзогенных переменных вектора $x_{t\bullet} = [x_{t1}, x_{t2}, \ldots, x_{tK}]$, но и некоторых других переменных из $y_{t\bullet}$.

Эту особенность можно интерпретировать как наличие постоянной обратной связи между выходом системы и входными данными. Поэтому j-е структурное уравнение, выражающее y_{tj} через элементы из $x_{t\bullet}$ и остальные переменные из $y_{t\bullet}$, можно записать как

$$y_{tj} = y_{t\bullet}c_{\bullet j} + x_{t\bullet}\beta_{\bullet j} + \varepsilon_{tj}, \tag{1}$$

где $c_{\bullet j}$ и $\beta_{\bullet j}$ являются векторами параметров этой системы. Также подразумевается, что $c_{jj} = 0$ с целью предотвратить появление y_{tj} и в левой, и в правой части. Уравнение также содержит случайное возмущение ε_{tj}.

Другим способом записи структурного уравнения, который ставит y_{tj} по соседству с другими экзогенными переменными расширенной системы, является выражение

$$y_{t\bullet}\gamma_{\bullet j} + x_{t\bullet}\beta_{\bullet j} + \varepsilon_{tj} = 0. \tag{2}$$

Таким образом, $\gamma_{\bullet j}$ и $c_{\bullet j}$ отличаются только своими j-ми элементами, равными $\gamma_{jj} = -1$ и $c_{jj} = 0$ соответственно. Условие $\gamma_{jj} = -1$, идентифицирующее зависимую переменную структурного уравнения, называется нормирующим правилом.

M структурных уравнений, собранные воедино, составляют следующую систему:

$$[y_{t1}, y_{t2}, \ldots, y_{tM}] = y_{t\bullet}[c_{\bullet 1}, c_{\bullet 2}, \ldots, c_{\bullet M}] + x_{t\bullet}[\beta_{\bullet 1}, \beta_{\bullet 2}, \ldots, \beta_{\bullet M}] + [\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{tM}], \tag{3}$$

которая в компактном виде выглядит следующим образом:

$$y_{t\bullet} = y_{t\bullet}C + x_{t\bullet}B + \varepsilon_{t\bullet} \tag{4}$$

Также ее можно записать как

$$y_{t\bullet}\Gamma + x_{t\bullet}B + \varepsilon_{t\bullet} = 0, \tag{5}$$

где $\Gamma = [\gamma_{\bullet 1}, \gamma_{\bullet 2}, \ldots, \gamma_{\bullet j}]$.

3 Приведенная форма

Если мы готовы пренебречь деталями структуры эконометрической модели, то можно выразить каждую переменную выхода из $y_{t\bullet} = [y_{t1}, y_{t2}, \ldots, y_{tM}]$, используя только экзогенные переменные. Такое представление называется приведенной формой модели. Приведенная форма получена из (5) умножением справа на матрицу, обратную к Γ, описывающую мгновенную обратную связь между переменными. Таким образом, получаем

$$y_{t\bullet} = x_{t\bullet}\Pi + \eta_{t\bullet}, \quad \text{где } \Pi = -B\Gamma^{-1} \text{ и } \eta_{t\bullet} = -\varepsilon_{t\bullet}\Gamma^{-1}. \tag{6}$$

Теперь следует сделать предположения о случайных элементах модели. Предположим, что элементы вектора $\varepsilon_{t\bullet} = [\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{tM}]$, являющиеся M структурными ошибками, распределены независимо по времени так, что для любого t выполняется

$$\mathbb{E}[\varepsilon_{t\bullet}] = 0 \quad \text{и} \quad \mathbb{V}[\varepsilon_{t\bullet}] = \mathbb{E}[\varepsilon'_{t\bullet}\varepsilon_{t\bullet}] = \Sigma_{\varepsilon\varepsilon}. \tag{7}$$

Также предполагается, что структурные ошибки распределены независимо от экзогенных переменных, так что $\mathbb{C}[\varepsilon_{t\bullet}, x_{s\bullet}] = 0$ для любых t и s.

Отсюда следует, что для вектора шоков приведенной формы $\eta_{t\bullet} = -\varepsilon_{t\bullet}\Gamma^{-1}$ выполнено

$$\mathbb{E}[\eta_{t\bullet}] = 0 \quad \text{и} \quad \mathbb{V}[\eta_{t\bullet}] = \Gamma'^{-1}\mathbb{V}[\varepsilon_{t\bullet}]\Gamma^{-1} = \Gamma'^{-1}\Sigma_{\varepsilon\varepsilon}\Gamma^{-1} = \Omega. \tag{8}$$

Преобразованные ошибки также независимы от $x_{t\bullet}$, откуда следует условие $\mathbb{C}[\eta_{t\bullet}, x_{s\bullet}] = 0$ для любых t и s.

4 Проблема идентификации и структурная модель

Структурная модель одновременных уравнений подвержена так называемой проблеме идентификации, ограничивающей возможности оценивания структурных параметров. Имея в наличии достаточный набор наблюдений, мы всегда можем оценить параметры статистической связи между эндогенными переменными из $y_{t\bullet}$ и экзогенными из $x_{t\bullet}$ в приведенной форме. Однако если мы собираемся оценить параметры структурных связей, то необходимо иметь априорную информацию о структуре модели.

Предположим, что статистические свойства данных можно полностью описать с помощью первых и вторых моментов. Обозначим дисперсионные матрицы $x_{t\bullet}$ и $y_{t\bullet}$ через $\mathbb{V}[x_{t\bullet}] = \Sigma_{xx}$ и $\mathbb{V}[y_{t\bullet}] = \Sigma_{yy}$, а матрицу их ковариаций через $\mathbb{C}[x_{t\bullet}, y_{t\bullet}] = \Sigma_{xy}$. Совмещая приведенную форму регрессионного отношения (6) с тривиальным равенством по $x_{t\bullet}$, получим следующую систему:

$$\begin{bmatrix} y_{t\bullet} & x_{t\bullet} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Pi & I \end{bmatrix} = \begin{bmatrix} \eta_{t\bullet} & x_{t\bullet} \end{bmatrix}. \tag{9}$$

Полагая $\mathbb{V}[\eta_{t\bullet}] = \Omega$ и $\mathbb{C}[\eta_{t\bullet}, x_{t\bullet}] = 0$, получим

$$\begin{bmatrix} I & -\Pi' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Pi & I \end{bmatrix} = \begin{bmatrix} \Omega & 0 \\ 0 & \Sigma_{xx} \end{bmatrix}. \tag{10}$$

Умножая слева эту систему на матрицу, обратную к самой левой, получаем эквивалентное уравнение в виде

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Pi & I \end{bmatrix} = \begin{bmatrix} I & \Pi' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Omega & 0 \\ 0 & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} \Omega & \Pi'\Sigma_{xx} \\ 0 & \Sigma_{xx} \end{bmatrix}. \tag{11}$$

Из этой системы можно выделить уравнения $\Sigma_{yy} - \Sigma_{yx}\Pi = \Omega$ и $\Sigma_{xy} - \Sigma_{xx}\Pi = 0$, из которых получаем параметры, характеризующие приведенную форму связей:

$$\Pi = \Sigma_{xx}^{-1}\Sigma_{xy} \quad \text{и} \quad \Omega = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{12}$$

Эти параметры могут быть оценены с помощью эмпирических аналогов матриц моментов Σ_{xx}, Σ_{yy} и Σ_{xy}, доступных в виде $M_{xx} = T^{-1}\sum_t x'_{t\bullet}x_{t\bullet}$, $M_{yy} = T^{-1}\sum_t y'_{t\bullet}y_{t\bullet}$ и $M_{xy} = T^{-1}\sum_t x'_{t\bullet}y_{t\bullet}$.

Теперь скомбинируем структурное уравнение (5) с тривиальным равенством, получая при этом аналог соотношения (9):

$$\begin{bmatrix} y_{t\bullet} & x_{t\bullet} \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ B & I \end{bmatrix} = \begin{bmatrix} \varepsilon_{t\bullet} & x_{t\bullet} \end{bmatrix}. \tag{13}$$

Используя $\mathbb{V}[\varepsilon] = \Sigma_{\varepsilon\varepsilon}$ и $\mathbb{C}[\varepsilon, x] = 0$, получаем

$$\begin{bmatrix} \Gamma' & B' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ B & I \end{bmatrix} = \begin{bmatrix} \Sigma_{\varepsilon\varepsilon} & 0 \\ 0 & \Sigma_{xx} \end{bmatrix}, \tag{14}$$

и, действуя по аналогии с (10)–(11), приходим к эквивалентному выражению

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ B & I \end{bmatrix} = \begin{bmatrix} \Gamma'^{-1} & \Pi' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{\varepsilon\varepsilon} & 0 \\ 0 & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} \Omega\Gamma & \Pi'\Sigma_{xx} \\ 0 & \Sigma_{xx} \end{bmatrix}. \tag{15}$$

Из этого равенства получаем фундаментальные уравнения, связывающие структурные параметры Γ и B с матрицами моментов переменных. Это уравнения можно записать в двух альтернативных формах:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{yy} - \Omega & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \Gamma \\ B \end{bmatrix} = \begin{bmatrix} \Pi'\Sigma_{xy} & \Pi'\Sigma_{xx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \Gamma \\ B \end{bmatrix}. \tag{16}$$

Первое уравнение прямо следует из (15). Второе следует из равенств $\Sigma_{yy} = \Pi'\Sigma_{xx}\Pi + \Omega$ и $\Sigma_{xy} = \Sigma_{xx}\Pi$, полученных из (12). Действительно, заменяя Π' на $\Sigma_{yx}\Sigma_{xx}^{-1}$, мы можем выразить матрицу второго уравнения, используя только моменты переменных.

Примем уравнение (16) за основу, на базе которой мы можем оценить значения структурных параметров Γ и B. Очевидно, что в таком случае система содержит недостаточно информации для оценивания. В частности, составляющее ее уравнение $\Pi'\Sigma_{xy}\Gamma + \Pi'\Sigma_{xx}B = 0$ получено преобразованием сопутствующего уравнения $\Sigma_{xy}\Gamma + \Sigma_{xx}B = 0$, и поэтому не содержит дополнительной информации. На самом деле, если на матрицы моментов не наложены ограничения, кроме естественных условий симметричности и положительной определенности, то количество неизвестных параметров, которые можно вывести из уравнения (16), не может превышать MK, что равно количеству параметров матрицы Π в приведенной форме.

Теоретически, априорная информация о Γ и B может принимать разные формы. На практике обычно рассматриваются только линейные ограничения на параметры, часто являющиеся правилами нормализации, устанавливающими диагональные элементы Γ равными -1, и исключающими ограничениями, приравнивающими некоторые элементы Γ и B нулю. Если ни одно из ограничений не задействует более одного уравнения, то есть возможность рассматривать каждое уравнение по отдельности.

Если ограничения на параметры j-го уравнения принимают форму исключающих ограничений или правил нормализации, то их можно представить в виде уравнения

$$\begin{bmatrix} R'_\diamond & 0 \\ 0 & R'_* \end{bmatrix} \begin{bmatrix} \gamma_{\bullet j} \\ \beta_{\bullet j} \end{bmatrix} = \begin{bmatrix} r_j \\ 0 \end{bmatrix} \quad \text{или} \quad \begin{bmatrix} R'_\diamond & 0 \\ 0 & R'_* \end{bmatrix} \begin{bmatrix} \gamma_{\bullet j} + e_j \\ \beta_{\bullet j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{17}$$

где R_* содержит набор столбцов единичной матрицы I_K порядка K, $R_\diamond$, аналогично, состоит из набора столбцов единичной матрицы I_M порядка M, а r_j является вектором, состоящим из нулей и -1, согласно правилам нормализации. Вектор e_j есть j-й столбец I_M, единица из которого сокращает нормированный элемент $\gamma_{\bullet j}$.

Мы можем представить все эти ограничения в компактной форме:

$$\begin{bmatrix} \gamma_{\bullet j} \\ \beta_{\bullet j} \end{bmatrix} = \begin{bmatrix} S_\diamond & 0 \\ 0 & S_* \end{bmatrix} \begin{bmatrix} \gamma_{\diamond j} \\ \beta_{* j} \end{bmatrix} - \begin{bmatrix} e_j \\ 0 \end{bmatrix}, \tag{18}$$

где $\gamma_{\diamond j}$ и $\beta_{* j}$ состоят из M_j и K_j не связанных ограничениями элементов $\gamma_{\bullet j}$ и $\beta_{\bullet j}$, и где $S_\diamond$ и S_* – аналоги $R_\diamond$ и R_*, состоящие из столбцов I_M и I_K соответственно.

При подстановке решения (18) в уравнение $\Sigma_{xy}\gamma_{\bullet j} + \Sigma_{xx}\beta_{\bullet j} = 0$, являющееся j-м уравнением системы (16), получим

$$\Sigma_{xy}S_\diamond\gamma_{\diamond j} + \Sigma_{xx}S_*\beta_{* j} = \Sigma_{xy}e_j. \tag{19}$$

Это набор K уравнений с M_j+K_j неизвестными; и, при условии, что матрица $[\Sigma_{xy}, \Sigma_{xx}]$ имеет полный ранг, необходимым и достаточным условием для идентифицируемости параметров j-го уравнения является $K \geq M_j + K_j$.

Если это условие выполнено, то любого подмножества уравнений (19) размера $M_j + K_j$ будет достаточно, чтобы определить $\gamma_{\diamond j}$ и β_{*j}. Однако нам особо интересен набор из $M_j + K_j$ независимых уравнений в форме

$$\begin{bmatrix} S_\diamond'\Pi'\Sigma_{xy}S_\diamond & S_\diamond'\Pi'\Sigma_{xx}S_* \\ S_*'\Sigma_{xy}S_\diamond & S_*'\Sigma_{xx}S_* \end{bmatrix} \begin{bmatrix} \gamma_{\diamond j} \\ \beta_{*j} \end{bmatrix} = \begin{bmatrix} S_\diamond'\Pi'\Sigma_{xy}e_j \\ S_*'\Sigma_{xy}e_j \end{bmatrix}, \tag{20}$$

полученной умножением слева уравнения (19) на матрицу $[\Pi S_\diamond, S_*]'$. Эти уравнения, полученные при использовании лишь связей между параметрами нашей модели и моментами векторов данных x и y, должны быть основой любой приемлемой оценки параметров индивидуальных структурных уравнений, независимо от предпосылок, из которых она получена.

5 Оценивание одиночного структурного уравнения методом наименьших квадратов

Рассмотрим тождество $\eta_{t\bullet}\Gamma = -\varepsilon_{t\bullet}$, описывающее связь между структурной и приведенной ошибками. Оно содержит равенство $\eta_{t\bullet}\gamma_{\bullet j} = -\varepsilon_{tj}$, которое можно использовать для придания j-му структурному уравнению вида

$$(y_{t\bullet} - \eta_{t\bullet})\gamma_{\bullet j} + x_{t\bullet}\beta_{\bullet j} = 0. \tag{21}$$

Это уравнение модели ошибок в переменных, в которой ошибки распространяются только на подмножество переменных.

Проигнорируем индекс, указывающий на расположение j-го структурного уравнения в системе M уравнений. Если бы дисперсионная матрица $\mathbb{V}[\eta_{t\bullet}] = \Omega$ была известна, то оценки параметров γ и β получались бы путем нахождения допустимых значений переменных, минимизирующих функцию

$$\sum_{t=1}^{T} \eta_{t\bullet}\Omega^{-1}\eta_{t\bullet}' = \sum_{t=1}^{T}(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})', \quad \text{где} \quad \mu_{t\bullet} = y_{t\bullet} - \eta_{t\bullet} = x_{t\bullet}\Pi, \tag{22}$$

при ограничении

$$\mu_{t\bullet}\gamma + x_{t\bullet}\beta = 0. \tag{23}$$

Последнее ограничение следует из соотношения $\Pi\Gamma + B = 0$, связывающего параметры приведенной и структурной моделей. Вместе уравнения (21)–(23) составляют постановку задачи, для которой Pollock (1979) получил оценки параметров структурной модели.

Минимизация выражения (22) происходит в два шага. Для начала можно заметить, что при фиксированных значениях γ и β уравнение (23) определяет гиперплоскость в пространстве размерности $K + M$, содержащую векторы $[y_{t\bullet}, x_{t\bullet}]'$, состоящие из наблюдений переменных системы. Точка $[\mu_{t\bullet}, x_{t\bullet}]'$ содержится в данной гиперплоскости, и в метрике, определенной матрицей Ω^{-1}, квадрат расстояния от нее до соответствующего вектора наблюдений равен

$$\|y_{t\bullet} - \mu_{t\bullet}\|^2_{\Omega^{-1}} = (y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})'. \tag{24}$$

Сначала минимизируем это расстояние для любых заданных γ и β. Затем найдем те значения γ и β, которые минимизируют сумму из квадратов расстояний, отраженную в формуле (22). Поэтому рассмотрим следующую функцию Лагранжа:

$$L = (y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})' + 2\lambda(\mu_{t\bullet}\gamma + x_{t\bullet}\beta). \tag{25}$$

Приравнивая производную этой функции по $\mu'_{t\bullet}$ к нулю, получим условие первого порядка на минимум расстояния

$$-(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1} + \lambda\gamma' = 0, \tag{26}$$

из которого следует, что

$$(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1} = \lambda\gamma', \tag{27}$$

или

$$(y_{t\bullet} - \mu_{t\bullet}) = \lambda\gamma'\Omega. \tag{28}$$

Вместе эти два уравнения дают

$$(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})' = \lambda^2\gamma'\Omega\gamma. \tag{29}$$

Но, умножая (28) справа на γ и пользуясь соотношением $-\mu_{t\bullet}\gamma = x_{t\bullet}\beta$, получим:

$$y_{t\bullet}\gamma - \mu_{t\bullet}\gamma = y_{t\bullet}\gamma + x_{t\bullet}\beta = \lambda\gamma'\Omega\gamma, \tag{30}$$

откуда следует, что

$$\lambda = \frac{y_{t\bullet}\gamma + x_{t\bullet}\beta}{\gamma'\Omega\gamma}.$$

Таким образом, (29) можно представить в следующем виде:

$$(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})' = \frac{(y_{t\bullet}\gamma + x_{t\bullet}\beta)^2}{\gamma'\Omega\gamma}. \tag{31}$$

Введем матрицы $Y' = [y'_{1\bullet}, y'_{2\bullet}, \ldots, y'_{t\bullet}]$ и $X' = [x'_{1\bullet}, x'_{2\bullet}, \ldots, x'_{t\bullet}]$, содержащие в себе весь набор наблюдений в системе за T периодов. Тогда выражение для суммы квадратов отклонений наблюдений от регрессионной гиперплоскости принимает вид

$$\sum_{t=1}^{T}(y_{t\bullet} - \mu_{t\bullet})\Omega^{-1}(y_{t\bullet} - \mu_{t\bullet})' = \frac{(Y\gamma + X\beta)'(Y\gamma + X\beta)}{\gamma'\Omega\gamma}. \tag{32}$$

Целевая функция (32) должна быть минимизирована при ограничениях, учитывающих всю априорную информацию о γ и β. В сочетании с информацией, содержащейся в выборке Y и X, этой априорной информации должно быть достаточно для идентификации параметров структурной модели. Априорная информация о γ и β обычно имеет форму *исключающих ограничений*, указывающих на то, что некоторые переменные, присутствующие в расширенной системе, отсутствуют в конкретном структурном уравнении. Также следует учесть *правило нормировки*, указывающее на то, что один из элементов $y_{t\bullet}$ является зависимой переменой в конкретном уравнении.

Общепринятым способом выражения априорной информации о параметрах является запись уравнения в виде

$$R'_1\gamma + R'_2\beta = r. \tag{33}$$

В случае, если все ограничения исключающие, то, пока элементы вектора r равны нулю, соответствующие элементы матриц R'_1 и R'_2 будут равны нулям и единицам, кроме соответствующего нормирующему правилу случая, когда элемент r равен -1. Возвращаясь к уравнению (17), получим, что $[R_\diamond, 0] = R_1$ и $[0, R_*] = R_2$.

Выпишем лагранжиан оптимизационной задачи оценивания параметров структурной модели при ограничениях:

$$L = \frac{(Y\gamma + X\beta)'(Y\gamma + X\beta)}{\gamma'\Omega\gamma} + 2\kappa'(R_1'\gamma + R_2'\beta - r). \tag{34}$$

Дифференцируя его по γ, используя формулу производной произведения и приравнивая результат к нулю, получим

$$\frac{(Y\gamma + X\beta)'Y}{\gamma'\Omega\gamma} - \frac{(Y\gamma + X\beta)'(Y\gamma + X\beta)\gamma'\Omega}{(\gamma'\Omega\gamma)^2} + \kappa' R_1' = 0. \tag{35}$$

Определив новый множитель $\mu = \kappa\gamma'\Omega\gamma$ и умножив обе части (35) на $\gamma'\Omega\gamma$, получим уравнение, транспонированное к которому имеет вид

$$Y'Y\gamma + Y'X\beta - \left\{\frac{(Y\gamma + X\beta)'(Y\gamma + X\beta)}{\gamma'\Omega\gamma}\right\}\Omega\gamma + R_1\mu = 0. \tag{36}$$

Далее, дифференцируя функцию Лагранжа (34) по β и приравнивая результат к нулю, получаем уравнение

$$\frac{(Y\gamma + X\beta)'X}{\gamma'\Omega\gamma} + \kappa' R_2' = 0, \tag{37}$$

которое при умножении на $\gamma'\Omega\gamma$ и транспонировании дает

$$X'Y\gamma + X'Y\beta + R_2\mu = 0. \tag{38}$$

Комбинируя уравнения (36) и (38) вместе с уравнением ограничений (33), получим следующую систему:

$$\begin{bmatrix} Y'Y - \lambda\Omega & Y'X & R_1 \\ X'Y & X'X & R_2 \\ R_1' & R_2' & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ r \end{bmatrix}, \tag{39}$$

где

$$\lambda = \left\{\frac{(Y\gamma + X\beta)'(Y\gamma + X\beta)}{\gamma'\Omega\gamma}\right\}. \tag{40}$$

Чтобы использовать эти уравнения для оценивания параметров структурного уравнения, следует оценить значение Ω, дисперсионной матрицы шоков приведенной формы модели

$$\Omega = \mathbb{V}[\eta_{t\bullet}] = \mathbb{E}\left[(y_{t\bullet} - x_{t\bullet}\Pi)'(y_{t\bullet} - x_{t\bullet}\Pi)\right]. \tag{41}$$

Прямое применение метода моментов дает подходящую оценку

$$\hat{\Omega} = \frac{1}{T}(Y - X\hat{\Pi})'(Y - X\hat{\Pi}) = \frac{1}{T}Y'\{I - X(X'X)^{-1}X'\}Y, \tag{42}$$

где $\hat{\Pi} = (X'X)^{-1}X'Y$ – оценка коэффициентов приведенной модели из уравнения (6).

Вместе уравнения (39) и (40) составляют нелинейную систему, которую следует решать итерационными методами, чтобы оценить структурные параметры. Можно привести итерационный алгоритм решения системы. На первом шаге производится присвоение $\lambda = \lambda_{(0)} = T$ в уравнении (39). В результате решения уравнения получаем оценки первого приближения $\gamma_{(1)}$ и $\beta_{(1)}$. Подстановка этих оценок в (40) дает обновленный коэффициент $\lambda_{(1)}$, которым заменим $\lambda_{(0)} = T$. При вторичном решении уравнения (39) с $\lambda = \lambda_{(1)}$ получаем оценки второго порядка $\gamma_{(2)}$ и $\beta_{(2)}$.

Несложно понять, как этот алгоритм обобщается на любое количество итераций. Процедуру можно остановить, когда значения γ и β, полученные из последовательных итераций, приблизительно равны. На практике четырех-пяти циклов бывает достаточно.

Значения $\gamma_{(1)}$ и $\beta_{(1)}$, возникающие на первом шаге алгоритма, на самом деле являются оценками 2ШМНК. Значения, к которым сходится алгоритм, являются оценками ММПОИ. Еще две итерационные схемы, начинающиеся с 2ШМНК-оценок и сходящиеся к ММПОИ-оценкам, были представлены в Pollock (1983). Первая из них производит итерации с учетом последующих оценок П, что позволяет учитывать информацию об ограничениях на структурные параметры, тогда как второй метод производит итерации с учетом оценки Ω при ограничениях.

6 Стандартные формы оценок

Полезно вывести стандартные формы уравнений для 2ШМНК- и ММПОИ-оценивания, используя систему из уравнений (39) и (40). Для начала примем обычное допущение, что, не считая правила нормализации, которое мы проигнорируем, априорные ограничения на γ и β принимают вид исключающих ограничений в форме отсутствия в интересующем нас структурном уравнении некоторых переменных, присутствующих в расширенной системе.

Примем форму записи $Y = [Y_*, Y_{**}]$ и $X = [X_*, X_{**}]$, где Y_{**} и X_{**} – матрицы исключенных переменных, и пусть первый столбец Y_* равен вектору наблюдений зависимой переменной структурного уравнения. В таком случае решаемые уравнения принимают вид

$$\begin{bmatrix} Y_*'Y_* - \lambda\hat{\Omega}_* & Y_*'X_* \\ X_*'Y_* & X_*'X_* \end{bmatrix} \begin{bmatrix} \gamma_* \\ \beta_* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{43}$$

$$\lambda = \left\{ \frac{(Y_*\gamma_* + X_*\beta_*)'(Y_*\gamma_* + X_*\beta_*)}{\gamma_*'\Omega_*\gamma_*} \right\}, \tag{44}$$

$$\hat{\Omega}_* = \frac{1}{T} Y_*' \{ I - X(X'X)^{-1}X' \} Y_*. \tag{45}$$

При решении второго уравнения системы (43), представляющего из себя $X_*'Y_*\gamma_* + X_*'X_*\beta_* = 0$, получаем

$$\beta_* = -(X_*'X_*)^{-1}X_*'Y_*\gamma_*. \tag{46}$$

При подстановке обратно в первое уравнение системы (43) получаем

$$(Y_*'Y_* - \lambda\hat{\Omega}_*)\gamma_* + Y_*'X_*(X_*'X_*)^{-1}X_*'Y_*\gamma_* = 0, \tag{47}$$

что можно переписать как

$$\{ Y_*'(I - P_*)Y_* - \lambda\hat{\Omega}_* \}\gamma_* = 0, \quad \text{где } \hat{\Omega}_* = \frac{1}{T} Y_*'(I - P)Y_*, \tag{48}$$

$P = X(X'X)^{-1}X'$ и $P_* = X_*(X_*'X_*)^{-1}X_*'$.

Заметим, что первое уравнение в (48) имеет форму уравнения для оценки модели ошибок в переменных. Стоит осознать этот факт, и альтернативные способы нахождения оценок γ_* и β_* напрашиваются сами собой. Сначала уравнение (48) решается путем нахождения характеристического корня λ и соответствующего характеристического вектора γ_*, используя любую стандартную технику, например, степенной метод.

Такая методология позволяет находить значение γ_* при произвольных ограничениях, например, при нормирующем ограничении, присваивающем первому элементу вектора γ_* значение -1. Путем подстановки правильно нормированной γ_* обратно в уравнение (46) можно найти оценку β_*.

Численные результаты работы этой схемы в точности те же, что получились бы при исполнении описанного ранее итерационного алгоритма до полной сходимости. На самом деле этот метод нахождения ММПОИ-оценок был общим результатом двух достаточно непохожих друг на друга выводов Anderson & Rubin (1949) и Koopmans, Rubin & Leipnik (1950).

Теперь рассмотрим случай, когда определяющие оценки уравнения нормируются с самого начала. Определим вектора $[-1, \gamma_\diamond'] = \gamma_*'$ и $[y_0, Y_\diamond] = Y_*$ для удобства умножения. В таком случае структурное уравнение, ранее записывавшееся как $Y_*\gamma_* + X_*\beta_* + \varepsilon = 0$, преобразуется в $y_0 = Y_\diamond\gamma_\diamond + X_*\beta_* + \varepsilon$, и соответствующая система уравнений, определяющая оценки параметров, выглядит как

$$\begin{bmatrix} y_0'y_0 - \lambda\hat{\omega}_{00} & y_0'Y_\diamond - \lambda\hat{\omega}_{0\diamond} & y_0'X_* \\ Y_\diamond'y_0 - \lambda\hat{\omega}_{\diamond 0} & Y_\diamond'Y_\diamond - \lambda\hat{\Omega}_{\diamond\diamond} & Y_\diamond'X_* \\ X_*'y_0 & X_*'Y_\diamond & X_*'X_* \end{bmatrix} \begin{bmatrix} -1 \\ \gamma_\diamond \\ \beta_* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{49}$$

где

$$\hat{\Omega}_* = \begin{bmatrix} \hat{\omega}_{00} & \hat{\omega}_{0\diamond} \\ \hat{\omega}_{\diamond 0} & \hat{\Omega}_{\diamond\diamond} \end{bmatrix}. \tag{50}$$

Игнорируя первое уравнение системы, служащее для выражения λ через $\gamma_\diamond$ и β_*, а также преобразуя оставшиеся уравнения, получаем систему

$$\begin{bmatrix} Y_\diamond'Y_\diamond - \lambda\hat{\Omega}_{\diamond\diamond} & Y_\diamond'X_* \\ X_*'Y_\diamond & X_*'X_* \end{bmatrix} \begin{bmatrix} \gamma_\diamond \\ \beta_* \end{bmatrix} = \begin{bmatrix} Y_\diamond'y_0 - \lambda\hat{\omega}_{\diamond 0} \\ X_*'y_0 \end{bmatrix}. \tag{51}$$

Положим $\lambda = T$. Это не только то значение, которое λ принимает на первом шаге итерационного алгоритма, описанного выше в разделе 5, но и значение, приводящее уравнения в соответствие с условиями на моменты (16) (по этой причине $\lambda = T$ можно было бы описать как асимптотическое значение, в свете анализа Anderson, 2005). Решением полученной системы будет 2ШМНК-оценка.

Перепишем уравнения, определяющие 2ШМНК-оценку, в более общей форме при помощи определения (42). Получаем

$$Y'Y - T\hat{\Omega} = Y'Y - (Y - X\hat{\Pi})'(Y - X\hat{\Pi}) = Y'Y - Y'(I - P)Y = Y'PY = \hat{\Pi}'X'X\hat{\Pi}, \tag{52}$$

где $P = X(X'X)^{-1}X'$. Также отметим, что $X'Y = X'\{X(X'X)^{-1}X'\}Y = X'X\hat{\Pi}$. Используя эти результаты, оценочные уравнения (51) можно преобразовать в

$$\begin{bmatrix} \hat{\Pi}_{X\diamond}'X'X\hat{\Pi}_{X\diamond} & \hat{\Pi}_{X\diamond}'X'X_* \\ X_*'X\hat{\Pi}_{X\diamond} & X_*'X_* \end{bmatrix} \begin{bmatrix} \gamma_\diamond \\ \beta_* \end{bmatrix} = \begin{bmatrix} \Pi_{X\diamond}'X'X\hat{\Pi}_{X0} \\ X_*'X\hat{\Pi}_{X0} \end{bmatrix}, \tag{53}$$

где $X\hat{\Pi}_{X\diamond} = X(X'X)^{-1}X'Y_\diamond$ и $X\hat{\Pi}_{X0} = X(X'X)^{-1}X'y_0$. Пользуясь последним равенством, перепишем уравнения в виде

$$\begin{bmatrix} \hat{\Pi}_{X\diamond}'X'Y_\diamond & \hat{\Pi}_{X\diamond}'X'X_* \\ X_*'XY_\diamond & X_*'X_* \end{bmatrix} \begin{bmatrix} \gamma_\diamond \\ \beta_* \end{bmatrix} = \begin{bmatrix} \Pi_{X\diamond}'X'y_0 \\ X_*'Xy_0 \end{bmatrix}, \tag{54}$$

являющимся прямым аналогом уравнений (20), описывающих связи между популяционными моментами и параметрами структурного уравнения.

7 Двухшаговый метод наименьших квадратов и метод инструментальных переменных

Уравнения на оценки 2ШМНК были независимо получены в Basmann (1957) и Theil (1958) с помощью подходов, сильно отличающихся от использованного в данной работе. Подход авторов заключался в подробном изучении причин того, почему невозможно получить состоятельные оценки параметров структурного уравнения методом наименьших квадратов.

Причина провала МНК-оценивания заключается в нарушении ключевого положения регрессионного анализа о некоррелированности ошибок с объясняющими переменными правой части уравнения. В уравнении $y_0 = Y_\diamond\gamma_\diamond + X_*\beta_* + \varepsilon$ присутствует прямая зависимость $Y_\diamond$ от структурных ошибок ε. Ошибки, однако, независимы от экзогенных переменных X_*.

Первоначальные выводы 2ШМНК-оценок основывались на идее, что можно очистить переменные $Y_\diamond$ от их зависимости от ε, после чего стандартная регрессия методом наименьших квадратов становится приемлемым способом оценивания. Таким образом, если бы $X\Pi_{X\diamond}$ были доступны, то ими можно было бы заменить $Y_\diamond$, и проблема зависимости была бы преодолена.

Несмотря на то, что значение $X\Pi_{X\diamond}$ неизвестно, его состоятельная оценка доступна в виде $\hat{Y}_\diamond = X\hat{\Pi}_{X\diamond}$. Нахождение $\hat{\Pi}_{X\diamond}$ представляет собой первый шаг процедуры 2ШМНК. Применение МНК к уравнению $y_o = \hat{Y}_\diamond\gamma_\diamond + X_*\beta_* + e$ является вторым шагом. Совместно эти этапы представляют собой подход Theil (1958), давшего имя оценке двухшагового метода наименьших квадратов.

Альтернативный подход, приводящий к той же оценке 2ШМНК, осуществляется через метод инструментальных переменных. Суть этого метода состоит в нахождении ряда переменных, коррелированных с регрессорами, но некоррелированных с ошибками.

В случае структурного уравнения подходящими инструментальными переменными будут переменные, экзогенные ко всей системе, содержащиеся в матрице X. Умножая слева структурное уравнение на X', получим

$$X'y_0 = X'Y_\diamond\gamma_\diamond + X'X_*\beta_* + X'\varepsilon. \tag{55}$$

В этой системе перекрестные члены соответствуют матрицам моментов, имеющим следующие пределы:

$$\operatorname{plim}(T^{-1}X'y_0) = \Sigma_{xy}e_0, \quad \operatorname{plim}(T^{-1}X'Y_\diamond) = \Sigma_{xy}S_\diamond, \quad \operatorname{plim}(T^{-1}X'X_*) = \Sigma_{xx}S_*, \tag{56}$$
$$\operatorname{plim}(T^{-1}X'\varepsilon) = 0.$$

При замене матриц моментов на их предельные значения получим уравнение

$$\Sigma_{xy}e_0 = \Sigma_{xy}S_\diamond\gamma_\diamond + \Sigma_{xx}S_*\beta_*, \tag{57}$$

ранее уже встречавшееся, см. (19). В этой системе K уравнений и $M_\diamond + K_*$ параметров. Можно предположить, что $[\Sigma_{xy}, \Sigma_{xy}]$ имеет полный ранг, в этом случае необходимым условием идентифицируемости параметров $\gamma_\diamond$ и β_* является $K \geq M_\diamond + K_*$, смысл которого в том, что число оцениваемых структурных параметров системы не должно превышать количества экзогенных переменных.

Эмпирическим аналогом (57) является уравнение

$$X'y_0 = X'Y_\diamond\gamma_\diamond + X'X_*\beta_*. \tag{58}$$

В случае, если $K = M_\diamond + K_*$, это уравнение можно явно разрешить и получить оценки. Однако если $K > M_\diamond + K_*$, то уравнения алгебраически несовместны. В этом случае говорят, что его параметры сверхидентифицированы. Для разрешения этой несовместности можно применить к (55) обобщенный метод наименьших квадратов. Ошибка в (55), $X'\varepsilon$, имеет

дисперсионную матрицу $\mathbb{V}[X'\varepsilon] = \sigma^2 X'X$. При использовании ее в ОМНК-оценивании мы снова получим оценку 2ШМНК.

Basmann (1957) в своей работе получил оценку 2ШМНК, следуя описанному ОМНК-подходу. Эту оценку также можно рассматривать в рамках метода инструментальных переменных, как это было сделано в Sargan (1958).

Список литературы

Anderson, T.W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics* 127, 1–16.

Anderson, T.W. & H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.

Anderson, T.W. & H. Rubin (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 21, 570–82.

Basmann, R.L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, 77–84.

Koopmans, T.C., H. Rubin & R.B. Leipnik (1950). Measuring the equation systems of dynamic economics. Глава 2 в *Statistical Inference in Dynamic Economic Models* под редакцией T.C. Koopmans. New York: John Wiley & Sons.

Malinvaud, E. (1966). Statistical Methods of Econometrics. Amsterdam: North-Holland.

Pollock, D.S.G. (1979). The Algebra of Econometrics. Chichester: John Wiley & Sons.

Pollock, D.S.G. (1983). Varieties of the LIML estimator. *Australian Economic Papers* 1983, 499–506.

Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

Theil, H. (1958, 1961). Economic Forecasts and Economic Policy. Amsterdam: North-Holland.

Estimation of Structural Econometric Equations

Stephen Pollock

Queen Mary College, University of London, United Kingdom

Derivations are offered for the LIML and the 2SLS estimators of single equations of the classical econometric simultaneous-equation system that differ from the usual ones. By assimilating both estimators to the method of moments, their essential similarities are highlighted. The LIML estimator is derived from a least-squares criterion that exploits the interpretation of the structural equation as an error-in-variables model, and the 2SLS estimator is obtained by an approximation that is asymptotically valid. The LIML estimator may be calculated via an iterative procedure that begins with the 2SLS estimator. The conventional derivations of the 2SLS estimator are also reviewed.

Оптимальные инструменты*

Станислав Анатольев†

Российская экономическая школа, Москва

Настоящее эссе содержит краткий обзор оптимального инструментирования в линейных и нелинейных моделях, как кросс-секционных, так и на стационарных временных рядах. Разобраны примеры разумного построения инструментов.

1 Безусловные кросс-секционные модели

Рассмотрим привычное линейное уравнение с инструментальными переменными

$$y = x'\beta + e, \quad \mathbb{E}[ze] = 0,$$

где x является $k \times 1$ вектором объясняющих переменных («регрессоров»), β – оцениваемым $k \times 1$ вектором параметров, z – $l \times 1$ вектором *базовых* экзогенных инструментов, $l \geqslant k$. Инструменты используются из-за потенциальной эндогенности объясняющих переменных. В частном случае линейной регрессии имеем $z = x$ и $l = k$. Пусть имеется случайная выборка $\{(x_i, y_i, z_i)\}_{i=1}^n$, т.е. рассматриваются кросс-секции.

Вспомним необходимые условия качества базовых инструментов z.

- Условие «годности» инструментов, т.е. их некоррелированности с ошибками:

$$\mathbb{E}[ze] = 0.$$

- Условие «релевантности» инструментов, т.е. их коррелированности с регрессорами:

$$\underset{k\times l}{Q_{xz}} = \mathbb{E}[xz'] \text{ – матрица полного ранга } k.$$

- Условие линейной «неповторяемости» инструментов:

$$\underset{l\times l}{Q_{zz}} = \mathbb{E}[zz'] \text{ – невырожденная матрица.}$$

В случае точной идентификации ($l = k$) используется простая оценка инструментальных переменных (ИП)

$$\widehat{\beta}_{IV} = \left(\sum_{i=1}^n z_i x_i'\right)^{-1} \sum_{i=1}^n z_i y_i.$$

В случае сверхидентификации ($l > k$) мы из определенных проекционных соображений строим оценку двухшагового метода наименьших квадратов (2ШМНК):

$$\widehat{\beta}_{2SLS} = \left(\sum_{i=1}^n x_i z_i' \left(\sum_{i=1}^n z_i z_i'\right)^{-1} \sum_{i=1}^n z_i x_i'\right)^{-1} \sum_{i=1}^n x_i z_i' \left(\sum_{i=1}^n z_i z_i'\right)^{-1} \sum_{i=1}^n z_i y_i.$$

*Цитировать как: Анатольев, Станислав (2007) «Оптимальные инструменты», Квантиль, №2, стр. 61–69. Citation: Anatolyev, Stanislav (2007) "Optimal instruments," Quantile, No.2, pp. 61–69.

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

Заметим, что эту оценку можно переписать как

$$\widehat{\beta}_{2SLS} = \left(\sum_{i=1}^{n} \widehat{\zeta}_i x_i'\right)^{-1} \sum_{i=1}^{n} \widehat{\zeta}_i y_i,$$

где $\widehat{\zeta}_i$ представляет собой i-е «наблюдение» вектора

$$\underset{k\times 1}{\widehat{\zeta}} = \frac{1}{n}\sum_{j=1}^{n} x_j z_j' \left(\frac{1}{n}\sum_{j=1}^{n} z_j z_j'\right)^{-1} z = \underset{k\times l}{\widehat{Q}_{xz}} \; \underset{l\times l}{\widehat{Q}_{zz}^{-1}} \; \underset{l\times 1}{z},$$

являющегося определенным линейным преобразованием базового инструмента z. Сама же 2ШМНК-оценка выглядит как обычная (т.е. как будто бы при точной идентификации) ИП-оценка с инструментом $\widehat{\zeta}_i$ для i-го наблюдения.

Известно, что при условной гетероскедастичности асимптотически эффективной является не 2ШМНК, а эффективная ОММ-оценка

$$\begin{aligned}\widehat{\beta}_{GMM} &= \left(\sum_{i=1}^{n} x_i z_i' \left(\sum_{i=1}^{n} z_i z_i' \widehat{e}_i^2\right)^{-1} \sum_{i=1}^{n} z_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i z_i' \left(\sum_{i=1}^{n} z_i z_i' \widehat{e}_i^2\right)^{-1} \sum_{i=1}^{n} z_i y_i \\ &= \left(\sum_{i=1}^{n} \widehat{\zeta}_i x_i'\right)^{-1} \sum_{i=1}^{n} \widehat{\zeta}_i y_i,\end{aligned}$$

которая также выглядит как обычная ИП-оценка с инструментом $\widehat{\zeta}$, который является линейной комбинацией базового инструмента:

$$\underset{k\times 1}{\widehat{\zeta}} = \frac{1}{n}\sum_{j=1}^{n} x_j z_j' \left(\frac{1}{n}\sum_{j=1}^{n} z_j z_j' \widehat{e}_j^2\right)^{-1} z = \underset{k\times l}{\widehat{Q}_{xz}} \; \underset{l\times l}{\widehat{Q}_{zze^2}^{-1}} \; \underset{l\times 1}{z}.$$

Отсюда следует естественная интерпретация: эффективная ОММ-оценка (или 2ШМНК-оценка в случае условной гомоскедастичности) берет первоначальные инструменты z (их l штук) и умножает их слева на выписанную выше последовательность матриц, имеющую в совокупности размерность $k \times l$. Таким образом берутся, пусть и сложные, линейные комбинации от базовых инструментов так, чтобы получился точно идентифицирующий инструмент. Используемое взвешивание является оптимальным, ибо мы рассматриваем эффективный ОММ. Таким образом, эффективный ОММ неявно «ужимает» ячейки с информацией до необходимого размера оптимальным образом. Приведенный выше $\widehat{\zeta}$ называется *оптимальным инструментом*. Используя популяционные аналоги, можно получить (конечно же, недоступный) *идеальный оптимальный инструмент*, который асимптотически эквивалентен $\widehat{\zeta}$:

$$\zeta = Q_{xz} Q_{zze^2}^{-1} z,$$

где $Q_{zze^2} = \mathbb{E}[zz'e^2]$.

2 Условные кросс-секционные модели

Большинство линейных моделей в современной эконометрике имеют вид

$$y = x'\beta + e, \quad \mathbb{E}[e|z] = 0.$$

Здесь те же обозначения, что использовались ранее; единственное отличие – в условности ограничения $\mathbb{E}[e|z] = 0$. Действительно, в той же линейной регрессии (т.е. при $z = x$) мы

имеем ограничение $\mathbb{E}[e|x] = 0$, а не $\mathbb{E}[xe] = 0$; последнее условие определяет всего лишь линейную проекцию и моделью не является. Мы продолжаем рассматривать случай кросс-секций.

В такой модели не только сами базовые инструменты z и их линейные комбинации могут использоваться в качестве инструментов, но и любые нелинейные функции от z. Действительно, для любой функции $f : \mathbb{R}^l \to \mathbb{R}^k$ инструмент $f(z)$ является годным: $\mathbb{E}[f(z)e] = 0$. Естественно, подразумеваются существование всех используемых моментов и релевантность $f(z)$. Какую функцию f выбрать наиболее выгодно с точки зрения максимизации асимптотической эффективности? Тождественную, логарифмическую, синусоидальную?

Конечно, не все так просто. Как минимум, форма функции должна зависеть от свойств рассматриваемой задачи. Обратимся вновь к случаю $z = x$. При условной гомоскедастичности мы, естественно, применяем МНК, что соответствует выбору тождественной функции в качестве f; при условной же гетероскедастичности мы применили бы ОМНК (отвлекаясь пока от проблематичности его реализации), что соответствует выбору $f(x) = x/\mathbb{E}[e^2|x]$. Оказывается, что и в общем случае оптимальный инструмент имеет похожую простую форму (Chamberlain, 1987)

$$\zeta = \frac{\mathbb{E}\left[x|z\right]}{\mathbb{E}\left[e^2|z\right]}.$$

Несмотря на простоту формы оптимального инструмента, его практическая реализация проблематична: в формуле фигурируют условные матожидания, о которых в данной задаче наверняка ничего не известно, поэтому их надо оценивать, причем если это делать честно, следует применять непараметрические методы. В данном контексте было предложено применять метод ближайших соседей (Robinson, 1987) и разложения в ряды (Newey, 1990), помимо прочего. Конечно, это не очень приятно. К тому же у исследователя должна быть солидного размера выборка для получения терпимой точности непараметрического оценивания, иначе асимптотика будет «работать» плохо.

3 Оптимальные инструменты в моделях временных рядов

В моделях временных рядов теория и практика оптимальных инструментов сильно усложняются, но и становятся более интересными. Обычно модель имеет вид

$$y_t = x_t'\beta + e_t, \quad \mathbb{E}\left[e_t|\Im_t\right] = 0,$$

где x_t is $k \times 1$, z_t – $\ell \times 1$ базовый инструмент, а $\Im_t$ содержит информацию в $z_t, z_{t-1}, \ldots$. Все ряды предполагаются стационарными и эргодическими. Максимальное множество годных инструментов «дважды бесконечно»: во-первых, как в предыдущем разделе, оно включает все нелинейные функции от z_t, а во-вторых, и все лаги (также вместе с нелинейными функциями от них) z_{t-j}, $j > 0$. Скомбинировать оптимальным образом имеющуюся в инструментах и их предыстории информацию является в общем случае сверхзадачей. Мы кратко опишем имеющиеся теоретические результаты в разной сложности задачах; подробности можно узнать в Anatolyev (2007) и в цитируемых там источниках.

Если задача относится к категории «одноперiодных», т.е. ошибка e_t представляет собой мартингальное приращение по отношению к своему прошлому (грубо говоря, серийно некоррелирована), то форма оптимального инструмента следующая:

$$\zeta_t = \frac{\mathbb{E}\left[x_t|\Im_t\right]}{\mathbb{E}\left[e_t^2|\Im_t\right]}.$$

Заметим, что эта форма в точности та же, какую мы наблюдали в предыдущем разделе. Например, рассмотрим условно гомоскедастичную авторегрессию $\mathsf{AR}(k)$

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_k y_{t-k} + \varepsilon_t,$$

где ε_t – строгий белый шум. Очевидно, оптимальный инструмент ζ_t пропорционален вектору $x_t = (y_{t-1}\ y_{t-2}\ \cdots\ y_{t-k})'$, что сводит дело к оцениванию с помощью МНК.

Если задача относится к категории «многопериодных», т.е. ошибка e_t серийно коррелирована до ненулевого конечного порядка q, и наличествует условная гомоскедастичность, то форма оптимального инструмента (точнее, рекурсивное соотношение для него) следующая (Hansen, 1985):

$$\Theta(L)\zeta_t = \mathbb{E}\left[\Theta(L^{-1})^{-1}x_t|\Im_t\right],$$

где $\Theta(L) = \theta_0+\theta_1 L+\cdots+\theta_q L^q$ – лаговый полином из разложения Вольда ошибки: $e_t = \Theta(L)\varepsilon_t$. Например, в «двухпериодной» (т.е. $q = 1$) условно гомоскедастичной ARMA(1,1)-модели

$$y_t = \rho y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1},$$

где ε_t – строгий белый шум, оптимальный инструмент представим в виде

$$(1-\theta L)\,\zeta_t = \frac{1}{1-\theta L}y_{t-2},$$

или в конечном счете

$$\zeta_t = \sum_{i=1}^{\infty} i\theta^{i-1}y_{t-1-i}.$$

Интересно, что оптимальный инструмент в этом случае задействует всю предысторию «объясняющей» переменной, причем взвешивает ее старые значения с затуханием, чуть более медленным, чем экспоненциальное. Задействование всей предыстории – это последствие многопериодности задачи. Можно сравнить данный случай с рассмотренным выше случаем AR-модели, где оптимальный инструмент игнорирует предысторию вовсе.

Наконец, в самой общей категории задач – «многопериодных», т.е. с серийно коррелированной до ненулевого конечного порядка q ошибкой e_t, в присутствии условной гетероскедастичности – форма оптимального инструмента следующая (Anatolyev, 2003):

$$\Phi_t(L)\zeta_t = \mathbb{E}\left[\Psi_t(L^{-1})x_t|\Im_t\right],$$

где лаговые полиномы $\Phi_t(L)$ и $\Psi_t(L)$ на этот раз зависят от момента времени t. Определяются последние из сложной нелинейной системы (фактически функциональных) уравнений, имеющей множественные решения. Дополнительное условие стабильности рекурсии, приведенной выше, выделяет из множества решений искомое, определяющее оптимальный инструмент. Более того, увы, искомое решение не выражаемо в явном виде. Чтобы взглянуть, как эта система выглядит, рассмотрим случай «двухпериодной» задачи (т.е. $q = 1$), когда $\omega_t = \mathbb{E}[e_t^2|\Im_t]$, $\gamma_t = \mathbb{E}\left[e_t e_{t-1}|\Im_t\right]$, $\mathbb{E}\left[e_t e_{t-j}|\Im_t\right] = 0$ при $j > 1$. Тогда рекурсия для оптимального инструмента выглядит как

$$\zeta_t = \phi_t(\zeta_{t-1} - \gamma_t^{-1}\delta_t),$$

а система имеет следующий вид:

$$\gamma_t + \phi_t\left(\omega_t + \mathbb{E}_t\left[\phi_{t+1}\gamma_{t+1}|\Im_t\right]\right) = 0,$$

$$\delta_t = \mathbb{E}\left[x_t + \phi_{t+1}\delta_{t+1}|\Im_t\right].$$

Наконец, условие стабильности имеет вид

$$\mathbb{E}[\log|\phi_t|] < 0.$$

Систему можно разрешить в явном виде лишь в случае условной гомоскедастичности, и оптимальный инструмент тогда сводится к рассмотренному в предыдущей категории задач.

Очевидно, что оптимальные инструменты, имеющие такую сложную форму, почти невозможно реализовать на практике, по крайней мере в чистом виде (не сильно успешные попытки описаны в Anatolyev, 2007). Поэтому большее распространение получила идея сократить множество, в котором ищется оптимальный инструмент, таким образом жертвуя какой-то долей асимптотической эффективности, зато приобретая надежду, что результирующий инструмент будет иметь не такую сложную форму, и его будет не так трудно воплотить в жизнь на практике.

Наиболее интуитивным подмножеством всего множества годных инструментов является пространство линейных комбинаций базовых инструментов. То есть жертвуются все нелинейные инструменты, зато используется вся предыстория. Очевидно, что жертва может быть значительной, если, например, вновь вспомнить МНК и ОМНК в линейной регрессии. Ведь МНК как раз использует линейным образом регрессоры, а вот использование ОМНК требует нелинейного взвешивания регрессора. С другой стороны, мы знаем, что реализовать МНК гораздо проще, чем ОМНК (ведь форму условной гетероскедастичности мы, конечно же, не знаем). А если условная гетероскедастичность несильная, то, возможно, мы не очень-то много и теряем.

Теория так называемых *линейных оптимальных инструментов* в наиболее общем виде разработана в West, Wong & Anatolyev (2002). Например, рассмотрим условно гетероскедастичную авторегрессию $\mathsf{AR}(1)$

$$y_t = \rho y_{t-1} + \varepsilon_t,$$

где ε_t – симметрично распределенное мартингальное приращение. Тогда оптимальный инструмент равен

$$\zeta_t = \sum_{i=1}^{\infty} \frac{\rho^{r-1}}{\tau_r} \varepsilon_{t-i},$$

где $\tau_r = \mathbb{E}[\varepsilon_t^2 \varepsilon_{t-r}^2]\mathbb{E}[\varepsilon_t^2]^{-2}$. В условно гетероскедастичной же $\mathsf{ARMA}(1,1)$-модели

$$y_t = \rho y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1},$$

где ε_t – строгий белый шум, оптимальный инструмент выглядит еще более громоздко:

$$\zeta_t = \sum_{i=1}^{\infty} \phi_i \varepsilon_{t-1-i},$$

где

$$\phi_1 = \left(1 + \sum_{i=1}^{\infty} \theta^{2i} \frac{\tau_1}{\tau_{i+1}} \right)^{-1} \sum_{i=1}^{\infty} \theta^{i-1} \frac{\rho^i - \theta^i}{\tau_{i+1}}, \quad \phi_r = \sum_{i=0}^{\infty} \theta^i \frac{\rho^{j+i} - \theta^{j+i} (1 + \theta\phi_1\tau_1)}{\tau_{r+i+1}}.$$

Теперь в обоих примерах оба оптимальных инструмента задействует всю предысторию регрессора. К сожалению, правда заключается в том, что такой «пассивный» способ эксплуатировать условную гетероскедастичность не дает большого выигрыша по сравнению с ее неэксплуатацией. То есть, например, в $\mathsf{AR}(1)$-модели оптимальная ИП-оценка по асимптотическим свойствам гораздо ближе к МНК, чем к ОМНК (см. Anatolyev, 2007). С другой стороны, вся система весов зависит от сравнительно легко оцениваемого набора параметров (по крайней мере, если ввести параметризацию для ε_t^2 типа ARCH), не содержащих условных матожиданий.

4 Нелинейные модели

Всю представленную выше теорию с определенными модификациями можно применять и к нелинейным моделям регрессионного типа

$$y = m(x, \beta) + e, \quad \mathbb{E}[e|z] = 0,$$

а также к еще более общим моделям

$$\mathbb{E}[m(x, \beta)|z] = 0.$$

Разумеется, при этом после расчета оптимальных или «почти оптимальных» инструментов необходимо применять численную минимизацию ОММ-критерия, что в общем-то приходится делать независимо от выбора инструментов.

Более того, на этапе построения и расчета оптимальных инструментов необходим дополнительный этап получения предварительных состоятельных оценок параметров β. Причина в том, что объект, выполняющий в нелинейной модели функцию регрессора x – это так называемый *квазирегрессор* $\partial m(x, b)/\partial b$, оцененный в точке $b = \beta$ (в линейном случае $m(x, \beta) = x'\beta$, и квазирегрессором является сам регрессор). Дополнительный этап получения предварительных оценок обуславливается как раз зависимостью квазирегрессора от неизвестного параметра. Но состоятельную оценку обычно получить относительно легко (например, прогнать ОММ с неоптимально подобранными инструментами, скажем, с базовыми). Получив состоятельную оценку для β и подставив ее в выражение для квазирегрессора, следует оцененный квазирегрессор и использовать вместо x во всех формулах предыдущих разделов. В разделах 6–7 приводится последовательность действий для двух актуальных нелинейных моделей.

5 Выводы для практической работы

Что полезного может вынести эконометрист-практик из представленной выше теории оптимальных инструментов? Очевидно, в сложных задачах использование теории «по полной программе» затруднительно ввиду сложности конструируемых объектов даже на популяционном уровне, а тем более из-за трудностей в реализации в конкретной выборке. Тем не менее, некоторые упрощенные версии оптимальных или «почти оптимальных» инструментов заслуживают пристального внимания. В некоторых ситуациях промежуточной целью может служить построение «хороших», то есть сильных инструментов, с целью избежать проблем, связанных с использованием слабых инструментов (см. Цыплаков, 2007, раздел 6). В любом случае теорию полезно знать для лучшей ориентации во многих практических ситуациях.

Например, пусть есть основания считать, что в конкретной одноперiодной (или кросс-секционной) задаче очень сильная условная гетероскедастичность. Значит, в такой ситуации использование оптимального инструмента должно дать большую отдачу, чем в случае слабой гетероскедастичности, при которой идея оптимального инструментирования не стоит выделки. Другая ситуация – известно, что ошибка в задаче сильно автокоррелирована, а условная гетероскедастичность, наоборот, слабая. В таком случае мудрым решением будет построить инструмент, который был бы оптимальным в отсутствии гетероскедастичности вообще, ибо пренебрежение гетероскедастичностью существенно упрощает процесс реализации оптимального инструмента. Конечно, такой инструмент не будет исконно оптимальным, ибо мы используем неверную посылку при его построении, но он несильно, скорей всего, будет отличаться от исконно оптимального, если условная гетероскедастичность и впрямь слабая.

Далее мы разбираем две актуальные модели, для которых применяем теорию оптимальных инструментов и высказанные выше соображения для построения если не совсем оптимальных, то по крайней мере удачных инструментов. Один из примеров, попроще, – кросс-секционный; другой, посложнее, – на временных рядах.

6 Пример: нелинейная функция потребления

Обратимся к задаче оценивания нелинейной функции потребления:

$$C = \mu + \delta Y^\gamma + \varepsilon,$$

где C – потребление, Y – доход, а ε – ошибка со свойством $\mathbb{E}[\varepsilon|Y] = 0$. Подразумевается наличие случайной выборки. Вектором истинных параметров является $\beta = (\mu, \delta, \gamma)'$. В данной ситуации входящие в регрессионную функцию переменные экзогенны, поэтому, скажем, НЛМНК, который неявно подразумевает в качестве инструмента квазирегрессор $(1, Y^\gamma, \delta Y^\gamma \ln(Y))'$, дает состоятельные и асимптотически нормальные оценки. Их-то и можно использовать в качестве предварительных; назовем их $\hat\beta_0 = (\hat\mu_0, \hat\delta_0, \hat\gamma_0)'$.

В оцениваемом уравнении наверняка имеется сильная гетероскедастичность, например, из-за эффекта размера. Оптимальный инструмент согласно разделу 2 равен отношению квазирегрессора к скедастичной функции. Поэтому напрашивается параметризация скедастичной регрессии, например, как

$$(C - \hat\mu_0 - \hat\delta_0 Y^{\hat\gamma_0})^2 = \kappa + \lambda Y^\theta + \eta,$$

и оценивание ее с помощью НЛМНК. Расчетные значения этой скедастичной регрессии можно считать хорошим приближением для значений скедастичной функции. Возможно, что вышеуказанная параметризация избыточна, и простая квадратичная функция хорошо бы приблизила истинную скедастичную функцию; тогда бы для оценивания скедастичной регрессии хватило бы обычного МНК. Впрочем, о качестве параметрической спецификации лучше судить по обычной регрессионной диагностике. Если же данных много, то скедастичную регрессию лучше оценить непараметрически, тем более что это несложно сделать ввиду наличия всего одного регрессора; в таком случае конструируемые инструменты претендуют на звание оптимальных. Мы же, параметризуя скедастичную регрессию, получим в конечном счете «почти оптимальные» инструменты.

Итак, оценив квазирегрессор, используя предварительные НЛМНК-оценки, и оценив скедастичную функцию, мы делим первое на второе и получаем серию оптимальных или «почти оптимальных» инструментов, после чего применяем обычный нелинейный метод инструментальных переменных (см. Цыплаков, 2007, стр. 36). Внимательный читатель должен заметить, что описанные в этом примере действия – это фактически алгоритм обобщенного нелинейного метода наименьших квадратов. Это не совпадение, а следствие кросс-секционного контекста задачи и отсутствия эндогенности в регрессорах.

7 Пример: высокочастотные межтрансакционные доходности

Второй пример касается моделирования высокочастотных доходностей на финансовом рынке. Пусть $t_0 < t_1 < ... < t_i < ...$ – моменты осуществления финансовых трансакций, t_i для i-ой трансакции. Величина $d_i = t_i - t_{i-1}$ называется i-ой дюрацией. Далее, если за p_{t_i} обозначить цену финансового актива в момент i-ой трансакции, то доходность этой трансакции будет $R_i = \ln(p_{t_i}) - \ln(p_{t_{i-1}})$. Введем понятие нормированной (относительно дюрации) доходности $r_i = R_i/\sqrt{d_i}$. Рассмотрим недавно предложенную модель Meddahi, Renault & Werker (2006):

$$\mathbb{E}\left[(r_i^2 - \lambda) - (r_{i-1}^2 - \lambda)\exp(-\kappa d_{i-1})\frac{v(\kappa d_i)}{v(\kappa d_{i-1})}|\Im_{i-2}\right] = 0,$$

где $v(x) = (1 - \exp(-x))/x$ и $\Im_i = \{d_j, r_j, j \le i\}$. Вектором истинных параметров является $\beta = (\lambda, \kappa)'$.

Понятно, что здесь можно найти уйму годных инструментов, но вопрос в том, как построить если не оптимальные, то хотя бы очень хорошие инструменты. В данной задаче мы

имеем серийную корреляцию в ошибках (задача двухпериодная по построению) и наверняка сильную условную гетероскедастичность (присущую любым моделям для высокочастотных финансовых данных). Как мы знаем из раздела 3, явная формула для оптимального инструмента не выведена. Поэтому мы не рассчитываем на оптимальное инструментирование, но попробуем приблизиться к нему, привлекая соображения из раздела 5. Разумной выглядит следующая последовательность действий.

- Выбираем небольшой вектор инструментов, измеримых относительно $\Im_{i-2}$, например $(1, \ln(d_{i-2}), \ln(d_{i-3}), r_{i-2}, r_{i-3})'$ (здесь и далее дюрации прологарифмированы с целью симметризации распределения инструментов). Применяем ОММ и получаем предварительные оценки для λ и κ. Используя эти оценки, оцениваем квазирегрессоры $\partial m_i(\beta)/\partial\lambda$ и $\partial m_i(\beta)/\partial\kappa$, где $m_i(\beta)$ – выражение в уравнении модели под знаком условного матожидания. Кроме того, оценим дисперсию $\mathbb{E}[m_i(\beta)^2]$ и автоковариацию первого порядка $\mathbb{E}[m_i(\beta)m_{i-1}(\beta)]$, откуда получим состоятельную оценку для θ, коэффициента в разложении Вольда $m_i(\beta) = \varepsilon_i - \theta\varepsilon_{i-1}$. Последняя оценка пригодится позже.

- Линейно регрессируем оцененные квазирегрессоры на предикторах из $\Im_{i-2}$, а именно, на всяческих функциях от недавних d_{i-2} и r_{i-2}, например, на

  $$(1, \ln(d_{i-2}), r_{i-2}, r_{i-2}\ln(d_{i-2}), ..., \ln(d_{i-4}), r_{i-4}, r_{i-4}\ln(d_{i-4}))',$$

  руководствуясь обычными критериями качества регрессионной подгонки (такими как значения F-статистики), при этом ограничивая количество регрессоров; здесь можно проявить незаурядную фантазию. Расчетные значения этих двух регрессий можно использовать как приближения для $\mathbb{E}\left[\partial m_i(\beta)/\partial\lambda|\Im_{i-2}\right]$ и $\mathbb{E}\left[\partial m_i(\beta)/\partial\kappa|\Im_{i-2}\right]$. Аналогично, используя предварительные оценки, оцениваем условную дисперсию $\mathbb{E}[m_i(\beta)^2|\Im_{i-2}]$ и условную автоковариацию первого порядка $\mathbb{E}[m_i(\beta)m_{i-1}(\beta)|\Im_{i-2}]$, линейно регрессируя оцененные $m_i(\beta)^2$ и $m_i(\beta)m_{i-1}(\beta)$ на тех же (или других, в зависимости от качества подгонки) предикторах и получая расчетные значения.

- Довольно простой, но сильный инструмент можно теперь получить, как в предыдущем примере, разделив оценки проекции квазирегрессоров на оценку условной дисперсии. Этот инструмент статический, и его построение фактически повторяет алгоритм в предыдущем примере, игнорируя серийную корреляцию ошибок.

- Напротив, можно игнорировать условную гетероскедастичность, но учесть серийную корреляцию, что немного более трудоемко. Для этого мы строим проекции не только самого квазирегрессора, но и его J (скажем, 10) будущих значений, и модифицируем формулу, приведенную в разделе 3, следующим образом:

  $$\zeta_i = \theta\zeta_{i-1} + \sum_{j=0}^{J}\theta^j\mathbb{E}\left[\frac{\partial m_{i+j}(\beta)}{\partial\beta}|\Im_{i-2}\right] + \frac{\theta^{J+1}}{1-\theta}\mathbb{E}\left[\frac{\partial m_i(\beta)}{\partial\beta}\right].$$

  Имея оценки для всех задействованных компонент, остается только раскрутить эту рекурсию, начав с первого наблюдения и полагая довыборочные значения всех компонент равными нулю.

- Итак, мы построили два хороших, хотя и не оптимальных инструмента. Заметим, что каждый из них состоит из двух компонент, т.к. квазирегрессор состоит из двух компонент. Для максимально выгодного использования построенных инструментов необходимо теперь прогнать ОММ со сверхидентифицирующим инструментом, состоящим, таким образом, из четырех компонент.

Результирующий инструмент, строго говоря, не являемся оптимальным. Настоящий оптимальный инструмент, как говорилось в разделе 3, имеет очень сложную структуру и задан в неявном виде. В Anatolyev (2007) описано, как построить так называемый «приблизительно оптимальный» инструмент, который будет наверняка лучше построенного по вышеприведенной схеме, хотя вряд ли, по мнению автора, намного.

Список литературы

Цыплаков, А. (2007). Экскурс в мир инструментальных переменных. *Квантиль* 2, 21–47.

Anatolyev, S. (2003). The form of the optimal nonlinear instrument for multiperiod conditional moment restrictions. *Econometric Theory* 19, 602–609.

Anatolyev, S. (2007). Optimal instruments in time series: a survey. *Journal of Economic Surveys* 21, 143–173.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–334.

Hansen, L.P. (1985). A method for calculating bounds on the asymptotic variance-covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30, 203–228.

Meddahi, N., E. Renault & B. Werker (2006). GARCH and irregularly spaced data. *Economics Letters* 90, 200–204.

Newey, W.K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58, 809–837.

Robinson, P. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 55, 875–891.

West, K. D., K.-f. Wong & S. Anatolyev (2002). Instrumental variables estimation of heteroskedastic linear models using all lags of instruments. Working Paper, University of Wisconsin–Madison.

Optimal instruments

Stanislav Anatolyev

New Economic School, Moscow

This essay briefly surveys optimal instrumentation in linear and nonlinear models, both cross-sectional and stationary time series. Examples of judicious construction of instruments are given.

Слабые инструменты*

Адриан Паган†

Технологический университет Квинсленда, Брисбен, Австралия

Настоящее эссе представляет собой обзор избранной литературы по слабым инструментам. В нем рассматриваются простые модели для иллюстрации проблем, возникающих из-за слабости инструментов, и методы, предложенные для их решения. Поскольку литература по данному вопросу находится в процессе развития, в эссе кратко отражены лишь наиболее свежие результаты.

1 Проблемы, связанные с распределением инструментальной оценки при слабых инструментах

Рассмотрим простую модель

$$y_t = x_t\theta + u_t,$$

где $u_t \sim i.i.d.(0, \sigma_u^2)$, $\mathbb{E}[x_t u_t] \neq 0$. Также имеется набор инструментальных переменных z_t, удовлетворяющих $\mathbb{E}[z_t u_t] = 0$. Предположим, что

$$\begin{pmatrix} z_t \\ x_t \end{pmatrix} \sim i.i.d. \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 & \sigma_{zx} \\ \sigma_{zx} & \sigma_x^2 \end{pmatrix} \right).$$

Таким образом, имеются один регрессор и один инструмент. Тогда обычная инструментальная оценка имеет следующий вид:

$$\begin{aligned} \hat{\theta} - \theta_0 &= \left(\sum_t z_t x_t \right)^{-1} \left(\sum_t z_t u_t \right) \\ &= \left(\frac{T^{-1} \sum_t z_t x_t}{\hat{\sigma}_z \hat{\sigma}_x} \right)^{-1} \left(\frac{T^{-1} \sum_t z_t u_t}{\hat{\sigma}_z \sigma_u} \right) \frac{\sigma_u}{\hat{\sigma}_x} \\ &= \hat{\rho}_{zx}^{-1} \hat{\rho}_{zu} \cdot \frac{\sigma_u}{\hat{\sigma}_x}, \end{aligned} \tag{1}$$

где $\hat{\sigma}_z$ и $\hat{\sigma}_x$ – оценки стандартных отклонений z_t и x_t, а $\hat{\rho}_{zx}$ и $\hat{\rho}_{zu}$ – оценки коэффициентов корреляции (оценка $\hat{\rho}_{zu}$ включает истинное стандартное отклонение u_t, σ_u). Из этого выражения видно, что распределение $\hat{\theta} - \theta_0$ зависит от произведения трех случайных величин, причем считается, что $\sigma_u/\hat{\sigma}_x$ быстро сходится к константе, так что для изучения остается величина $\hat{\rho}_{zu}/\hat{\rho}_{zx}$. Любые проблемы с распределением $\hat{\theta}$ связаны с $\hat{\rho}_{zx}$, а именно с тем, насколько случайной является величина $\hat{\rho}_{zx}$, и насколько вероятны малые по абсолютному значению реализации этой случайной величины. Поскольку $\hat{\rho}_{zx}$ стоит в знаменателе, для таких реализаций мы получим большие значения $\hat{\theta} - \theta_0$, что приведет к скошенной плотности распределения $\hat{\theta} - \theta_0$.

Асимптотическая теория применяется в предположении, что размер выборки достаточно большой, чтобы воспринимать $\hat{\rho}_{zx}$ как константу и рассматривать

$$T^{1/2}(\hat{\theta} - \theta_0) = \hat{\rho}_{zx}^{-1}(T^{1/2}\hat{\rho}_{zu}) \cdot \frac{\sigma_u}{\hat{\sigma}_x},$$

*Перевод Б. Гершмана и С. Анатольева. Материал подготовлен на основе лекций, прочитанных автором на курсах Economics 4202 в университете Нового Южного Уэльса в 2002–2004 гг. и Economics 607 в университете Джона Гопкинса в 2004 г. Цитировать как: Паган, Адриан (2007) «Слабые инструменты», Квантиль, №2, стр. 71–81. Citation: Pagan, Adrian (2007) "Weak instruments," Quantile, No.2, pp. 71–81.

†Адрес: School of Economics and Finance, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia. Электронная почта: *a.pagan@qut.edu.au*

и, поскольку $T^{1/2}(\hat{\rho}_{zu} - \rho_{zu}) = T^{1/2}\hat{\rho}_{zu}$, ожидается, что $T^{1/2}\hat{\rho}_{zu}$ будет асимптотически иметь стандартное нормальное распределение $N(0,1)$ (если $\rho_{zu} = 0$). Кроме того, все другие величины сходятся к своим истинным значениям. Тогда при условии $\rho_{zx} \neq 0$ получаем, что $T^{1/2}(\hat{\theta} - \theta_0)$ будет иметь асимптотическое распределение $N(0, \rho_{zx}^{-2}\sigma_u^2/\sigma_x^2)$.

Ситуация осложняется, если $\rho_{zx} = 0$, поскольку тогда при применении того же подхода в больших выборках происходит деление на нуль. В случае, когда $\rho_{zx} = 0$, инструменты называют *нерелевантными*. Они являются *годными*, так как $\rho_{zu} = 0$, но представляют малый интерес. Однако можно сказать нечто большее. В частности, предположим, что асимптотически оценка $\hat{\rho}_{zx}$ также нормально распределена вокруг своего истинного значения (нуля в нашем случае), то есть $T^{1/2}\hat{\rho}_{zx}$ сходится по распределению к $N(0, v)$ (если $\rho_{zx} = 0$, то $v = 1$). Тогда имеем

$$\hat{\theta} - \theta_0 = (T^{1/2}\hat{\rho}_{zx})^{-1}(T^{1/2}\hat{\rho}_{zu}) \cdot \frac{\sigma_u}{\hat{\sigma}_x},$$

что в больших выборках реализуется в

$$\frac{N(0,1)}{N(0,1)} \cdot \frac{\sigma_u}{\hat{\sigma}_x}.$$

Отсюда ясно, что оценка $\hat{\theta}$ несостоятельна, то есть всегда существует случайный разрыв между $\hat{\theta}$ и θ_0. Иными словами, в отличие от стандартного случая, когда $\mathbb{V}[\hat{\theta} - \theta_0]$ стремится к нулю при $T \to \infty$, здесь дисперсия $\hat{\theta} - \theta_0$ не уменьшается с увеличением размера выборки. Таким образом, в случае нерелевантности инструментов «хорошие» свойства инструментальной оценки не сохраняются. Более того, если две нормальные случайные величины в приведенном выражении независимы, $\hat{\theta} - \theta_0$ будет иметь распределение Коши, у которого отсутствуют моменты.

Может показаться неправдоподобным, что ρ_{zx} в точности равняется нулю. Более вероятно, что эта величина просто мала. Рассмотрим подробнее, как изменится анализ в этом случае. Для этого предположим, что

$$x_t = z_t\pi + \xi_t,$$

где $\mathbb{E}[z_t\xi_t] = 0$. Это не что иное, как уравнение в приведенной форме, в котором z_t предполагается экзогенным. Полагая для удобства, что $\mathbb{E}[z_t] = \mathbb{E}[x_t] = 0$, получаем

$$\mathbb{E}[z_t x_t] = \mathbb{E}[z_t^2]\pi,$$

так что

$$\rho_{zx} = \frac{\mathbb{E}[z_t x_t]}{\sigma_x\sigma_z} = \frac{\mathbb{E}[z_t^2]\pi}{\sigma_x\sigma_z} = \frac{\pi\sigma_z}{\sigma_x}.$$

Будем называть *слабым инструментом* такой, для которого величина ρ_{zx} мала. Это означает, что π мало.

На первый взгляд кажется, что асимптотическая теория все еще применима в ситуации слабых, но релевантных инструментов, поскольку $\hat{\rho}_{zx}$ сходится к ненулевому значению. На самом деле это верно, но не следует спешить с выводами. Все-таки при $\rho_{zx} = 0$ теория не работает, и интуиция подсказывает, что если ρ_{zx} отличается от нуля на крайне малую величину, то свойства оценки будут ближе к ситуации $\rho_{zx} = 0$, чем к случаю $\rho_{zx} \neq 0$. Конечно, это, в сущности, размышления о свойствах оценки в конечных выборках, то есть о том, насколько большим должно быть T, чтобы действовала асимптотическая теория.

Возможно, полезно поразмышлять над этим эвристически, записав $\hat{\rho}_{zx}$ как $\rho_{zx} + \eta$, где $\eta \sim N(0, v/T)$, так что

$$T^{1/2}(\hat{\theta} - \theta_0) = \left\{ \frac{T^{1/2}\hat{\rho}_{zu}}{\rho_{xz} + N(0, v/T)} \right\} \frac{\sigma_u}{\hat{\sigma}_x}.$$

Ясно, что когда T становится большим, член $N(0, v/T)$ исчезает, делая реализации выражения в знаменателе, $\rho_{zx} + N(0, v/T)$, более близкими к ненулевому значению ρ_{zx}. Но в малой выборке и при достаточно большом значении v возможно, что реализация $\hat{\rho}_{zx}$ близка к нулю, что приводит к большому значению для $T^{1/2}(\hat{\theta} - \theta_0)$. В этом случае асимптотическая теория не дает хорошей аппроксимации поведения оценки θ в малых выборках. Поэтому понятно, что очень маленькое значение ρ_{zx} означает, что свойства оценки выглядят скорей как в случае $\rho_{zx} = 0$, чем как в случае $\rho_{zx} \neq 0$.

Хотелось бы иметь представление о том, что происходит по мере приближения ρ_{zx} к нулю. Данная ситуация аналогична той, которая возникает с тестовыми статистиками, которые всегда отвергают неверную нулевую гипотезу в больших выборках, то есть являются состоятельными. Чтобы сравнивать такие тесты, используется идея локальной альтернативы, то есть находятся распределения тестовых статистик при приближении альтернативной гипотезы к нулевой по мере увеличения размера выборки. В данной ситуации поступают аналогично, а именно, полагают $\pi = \phi/\sqrt{T}$. Тогда

$$\begin{aligned} T^{1/2}\hat{\rho}_{zx} &= T^{1/2}\rho_{zx} + T^{1/2}\eta \\ &= \frac{\phi\sigma_z}{\sigma_x} + T^{1/2}\eta \\ &= \frac{\phi\sigma_z}{\sigma_x} + \varepsilon, \end{aligned}$$

где $\varepsilon \sim N(0, v)$. Проанализируем, что происходит в случае слабых инструментов. Получаем

$$\begin{aligned} \hat{\theta} - \theta_0 &= (T^{1/2}\hat{\rho}_{zx})^{-1}(T^{1/2}\hat{\rho}_{zu}) \cdot \frac{\sigma_u}{\hat{\sigma}_x} \\ &\to^d \frac{N(0,1)}{N(\phi\sigma_z/\sigma_x, v)}\frac{\sigma_u}{\sigma_x}. \end{aligned}$$

Плохая новость в том, что теперь $\hat{\theta} - \theta_0$ – это отношение двух случайных величин, которое не сходится к нулю, то есть $\hat{\theta}$ не является состоятельной оценкой θ_0. Более того, $\hat{\theta} - \theta_0$ не может иметь нормальное распределение. Следовательно, когда ρ_{zx} мало, подобный анализ дает лучшее описание поведения $\hat{\theta} - \theta_0$, чем стандартная асимптотика, что было подтверждено с помощью симуляций. Вероятно, распределение $\hat{\theta} - \theta_0$ в конечных выборках будет сильно отличаться от нормального, когда присутствуют слабые инструменты, а оценки коэффициентов будут существенно смещены.

2 Как обнаружить слабые инструменты

Как обнаружить слабые инструменты? Поскольку проблема возникает, когда ρ_{zx} близко к нулю, логично тестировать гипотезу $\rho_{zx} = 0$, используя оценку $\hat{\rho}_{zx}$. Но поскольку $\hat{\rho}_{zx}$ кратно $\hat{\pi}$, МНК-оценке π в регрессии x_t на z_t, можно вместо этого тестировать гипотезу $\pi = 0$. Для этого подходит либо t-тест, либо F-тест на значимость регрессоров z_t в регрессии для x_t.

Последняя интерпретация становится еще полезней, если отойти от простой модели. Важной модификацией является модель, в которой в оцениваемом уравнении появляются дополнительные регрессоры w_t, наличествующие и в приведенной форме:

$$\begin{aligned} y_t &= x_t\theta + w_t\alpha + u_t, \\ x_t &= z_t\pi + w_t\gamma + \xi_t. \end{aligned}$$

Для анализа этой ситуации запишем уравнения в матричной форме:

$$\begin{aligned} y &= X\theta + W\alpha + u, \\ X &= Z\pi + W\gamma + \xi. \end{aligned}$$

Домножая слева оба уравнения на $M_W = I - W(W'W)^{-1}W'$, получим, что

$$\begin{aligned} M_W y &= M_W X\theta + u^*, \\ M_W X &= M_W Z\pi + \xi^*, \end{aligned}$$

или

$$\begin{aligned} y^* &= X^*\theta + u^*, \\ X^* &= Z^*\pi + \xi^*, \end{aligned}$$

так что вместо y, X и Z остаются y^*, X^*, Z^*. Величины вроде y^* являются остатками регрессии на W, а $\hat{\pi}$ теперь является оценкой π в регрессии X^* на Z^*, что совпадает с оценкой в регрессии X на Z и W. Тогда F-тест на равенство π нулю должен предусматривать присутствие W в регрессии, то есть правильной мерой наличия проблем со слабыми инструментами является корреляция между x_t и z_t *после устранения влияния* w_t. Поскольку F-статистика для $\pi = 0$ в модели без регрессоров равна $(1 - R^2)/R^2$, а при наличии w_t в уравнении в этой формуле R^2 просто заменяется на частный R^2, ясно, что даже когда R^2 большой – x_t хорошо объясняется с помощью z_t и w_t, – частный R^2 может быть очень мал, то есть большая часть x_t объясняется w_t, а не z_t. В этом случае z_t фактически является слабым инструментом. Проблема, конечно, в том, что переменная w_t недоступна в качестве инструмента для x_t, поскольку ее уже «использовали» при оценке α, то есть w_t нужна в качестве инструмента для себя самой. Действительно, большую часть объяснения x_t дает w_t, а не z_t, означая, что R^2 является плохим показателем того, насколько z_t полезны в качестве инструментов.

Staiger & Stock (1997) рекомендуют классифицировать инструменты как слабые, если F-статистика при тестировании $\pi = 0$ в регрессии x_t на z_t и w_t меньше 10.[1] В терминах R^2 это означает, что частный R^2 должен быть больше, чем 0,1. Это разумное практическое правило, которое широко используется. Что именно следует тестировать, когда $\dim(X) > 1$, менее понятно, поскольку в этом случае вектор переменных x_t регрессируется на набор регрессоров z_t. Одно из возможных предложений – использовать канонические корреляции – разработано в Hall, Rudebusch & Wilcox (1996). Stock & Yogo (2005) предложили использовать многомерный аналог коэффициента концентрации и выбирать его наименьшее собственное значение. Авторы приводят таблицу консервативных критических значений для такого теста, поскольку точное распределение найти сложно.

3 Инференция на основе инструментальной оценки

Обратимся теперь к инференции на основе инструментальной оценки. Интерес вызывают два вопроса: во-первых, вывод тестовой статистики для гипотезы $H_0 : \theta = \theta_0$, и во-вторых, способ построения доверительных интервалов. Последний вопрос слишком сложен для рассмотрения в настоящем эссе. Рассмотрим проблему нахождения тестовой статистики, которая давала бы вероятностное значение для тестирования нулевой гипотезы. Обычно для этого используется t-статистика.

Поскольку для оценивания используются моменты

$$\mathbb{E}[m_t] = \mathbb{E}[z_t u_t] = \mathbb{E}[z_t(y_t - x_t\theta)] = 0,$$

из теории метода моментов следует, что дисперсия $T^{1/2}(\hat{\theta} - \theta_0)$ имеет вид

$$V_{\hat{\theta}} = \mathbb{E}\left[\frac{\partial m_t}{\partial \theta}\right]^{-2} \mathbb{V}[m_t] = \sigma_{zx}^{-2}(\sigma_u^2\sigma_z^2),$$

[1] Как говорится в их статье, это оценка параметра концентрации, основного показателя, который влияет на распределение 2ШМНК-оценки в конечных выборках. Shea (1995) предлагает расширение теста на случай, когда X – многомерная величина, что оказалось полезным, например, в Pagan & Robertson (1996).

поскольку

$$\frac{\partial m_t}{\partial \theta} = -x_t z_t,\ \mathbb{V}[m_t] = \sigma_u^2 \mathbb{E}[z_t^2].$$

При стандартном выводе t-статистики $V_{\hat{\theta}}$ оценивается с помощью $\hat{\sigma}_{zx}^{-2}(\hat{\sigma}_u^2 \hat{\sigma}_z^2)$, то есть

$$\begin{aligned} t_{\hat{\theta}} &= \frac{T^{1/2}(\hat{\theta} - \theta_0)}{\sqrt{\hat{V}_{\hat{\theta}}}} = \frac{T^{1/2}\hat{\sigma}_{zx}^{-1}\hat{\sigma}_{zu}}{\hat{\sigma}_{zx}^{-1}(\hat{\sigma}_u \hat{\sigma}_z)} \\ &= \frac{T^{1/2}\hat{\sigma}_{zu}}{\hat{\sigma}_u \hat{\sigma}_z} \\ &= (T^{1/2}\hat{\rho}_{zu})\frac{\sigma_u}{\hat{\sigma}_u}. \end{aligned}$$

Обычно $T^{1/2}\hat{\rho}_{zu}$ асимптотически $N(0,1)$, так что возможные проблемы с распределением t-статистики возникают из-за множителя $\sigma_u/\hat{\sigma}_u$. В стандартных условиях этот множитель сходится к 1, что приводит к известному результату об асимптотической нормальности $t_{\hat{\theta}}$. Как меняется этот вывод при наличии слабых инструментов? Ответ кроется в равенстве

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{1}{T}\sum_t (y_t - x_t\hat{\theta})^2 = \frac{1}{T}\sum_t (y_t - x_t\theta_0 - x_t(\hat{\theta} - \theta_0))^2 \\ &= \frac{1}{T}\sum_t (y_t - x_t\theta_0)^2 + (\hat{\theta} - \theta_0)^2 \left(\frac{1}{T}\sum_t x_t^2\right) - 2(\hat{\theta} - \theta_0)\left(\frac{1}{T}\sum_t u_t x_t\right). \end{aligned}$$

Асимптотически, первый член – это σ_u^2, а два других члена обычно исчезают, поскольку $\hat{\theta}$ является состоятельной оценкой для θ_0. Но при слабых инструментах («локально нулевой асимптотике») $\hat{\theta}$ несостоятельна, так что $\hat{\sigma}_u$ асимптотически является случайной величиной, что делает распределение t-статистики нестандартным, так как это отношение нормальной случайной величины к распределению $\hat{\sigma}_u$. Конечно, когда вид этого распределения неизвестен, построение доверительного интервала затруднительно.

4 Построение полезных тестовых статистик

Для случая $\dim(X) = 1$ и $\dim(Z) \geq \dim(X)$ предложено много решений. Наиболее старый метод предлагает обходной путь для тестирования гипотезы $H_0 : \theta = \theta^*$ – с помощью решения другой задачи.[2] А именно, при оценивании θ предполагалось, что $\mathbb{E}[z_t u_t] = 0$. Теперь это условие выглядит как $\mathbb{E}[z_t(y_t - x_t\theta_0)] = 0$, где θ_0 – истинное значение θ, так что

$$\mathbb{E}[z_t(y_t - x_t'\theta^*)] = \mathbb{E}[z_t x_t'(\theta_0 - \theta^*)]. \tag{2}$$

Предполагая, что инструменты не являются нерелевантными, можно проверять нулевую гипотезу $\theta_0 = \theta^*$, тестируя $\mathbb{E}[z_t(y_t - x_t\theta^*)] = 0$.

Равенство (2) можно переформулировать как тест на условные моменты в виде

$$\mathbb{E}[z_t(y_t - x_t'\theta^*) - \gamma] = 0,$$

а $\gamma = 0$ можно тестировать, используя $\hat{\gamma} = T^{-1}\sum_{t=1}^T z_t(y_t - x_t'\theta_0)$. Записывая в матричной форме (и отбрасывая T), получаем $Z'(y - X\theta_0)$. Если нулевая гипотеза $H_0 : \theta = \theta_0$ верна и $u_t \sim i.i.d.(0, \sigma_0^2)$, то $\mathbb{V}[Z'(y - X\theta_0)] = \sigma_0^2 Z'Z$, что дает статистику

$$AR = \frac{(y - X\theta_0)'Z(Z'Z)^{-1}Z(y - X\theta_0)}{\tilde{\sigma}_0^2},$$

[2] Важно отметить, что под θ подразумеваются коэффициенты при эндогенных переменных. Если в исходном соотношении имеются экзогенные переменные, они устраняются, как показано в разделе 2, так что y_t, z_t и x_t будут остатками регрессий для исходных переменных.

где $\tilde{\sigma}_0^2 = T^{-1}\sum_t (y_t - x_t'\theta_0)^2$. Это – *тестовая статистика Андерсона–Рубина* (AR). Она имеет асимптотическое распределение $\chi^2(\dim(Z))$ и вполне хорошо ведет себя в конечных выборках. Тот факт, что она имеет распределение $\chi^2(\dim(Z))$, а не $\chi^2(\dim(\theta))$, печален, поскольку $\dim(Z)$ может значительно превосходить $\dim(\theta)$. Если тестируются *все* параметры θ, тест естественным образом расширяется на случай $\dim(\theta) > 1$, но это скорее редкость.

Есть и другие способы тестирования, более непосредственные, чем AR-тест. Выше было описано не что иное, как тест Вальда. Как насчет LM-теста? Тест Вальда и LM-тест идентичны, когда $\dim(Z) = 1$, просто в LM-тесте $\hat{\sigma}^2$ заменяется на $\tilde{\sigma}_0^2$. Отсюда следует, что распределение t-статистики из LM-теста асимптотически стандартное нормальное. Этот изящный результат впервые получили Wang & Zivot (1996). К сожалению, он не выполнен для более интересных моделей, в частности, для случая, когда инструментов больше, чем регрессоров, требующих инструментирования, то есть $\dim(z_t) > \dim(x_t)$. Чтобы понять, почему так происходит, заметим, что 2ШМНК-оценка для θ имеет вид

$$\hat{\theta} - \theta_0 = (\hat{\pi}'Z'Z\hat{\pi})^{-1}(\hat{\pi}'Z'u), \quad \mathrm{sd}(\hat{\theta}) = \hat{\sigma}_u(\hat{\pi}'Z'Z\hat{\pi})^{-1/2},$$

так что

$$t_{\hat{\theta}} = \frac{(\hat{\pi}'Z'Z\hat{\pi})^{-1}(\hat{\pi}'Z'u)}{\hat{\sigma}_u(\hat{\pi}'Z'Z\hat{\pi})^{-1/2}}.$$

В случае $\dim(Z) = \dim(X) = 1$, $\hat{\pi}$ – скалярная величина, которая сокращается в числителе и знаменателе, оставляя нас с

$$t_{\hat{\theta}} = \frac{Z'u}{\hat{\sigma}_u(Z'Z)^{1/2}},$$

и тогда достаточно просто заменить $\hat{\sigma}_u$ на $\tilde{\sigma}_0$, чтобы обеспечить асимптотическую нормальность $t_{\hat{\theta}}$. Но если $\dim(Z) > \dim(X)$, $\hat{\pi}$ не сокращается, и возникает вопрос о распределении $\hat{\pi}$ при локально нулевой асимптотике в случае слабых инструментов

$$\hat{\pi} = \frac{\phi}{\sqrt{T}} + (Z'Z)^{-1}Z'\xi.$$

Чтобы $\hat{\pi}$ не ушло в нуль, необходимо, чтобы $T^{1/2}\hat{\pi}$ было асимптотически

$$N(\phi, \sigma_\xi^2\sigma_z^{-2}).$$

Для анализа последствий таких действий заметим, что

$$t_{\hat{\theta}} = \frac{[(T^{1/2}\hat{\pi}')(T^{-1}Z'Z)(T^{1/2}\hat{\pi})]^{-1}(T^{1/2}\hat{\pi}')(T^{-1/2}Z'u)}{\hat{\sigma}_u[(T^{1/2}\hat{\pi}')(T^{-1}Z'Z)(T^{1/2}\hat{\pi})]^{-1/2}},$$

и, учитывая, что $\hat{\pi}$ и $T^{-1/2}Z'u$ имеют предельные распределения, t-статистика – это произведение (и отношение) множества (асимптотически) нормальных случайных величин. Следовательно, даже при замене $\hat{\theta}$ на θ_0 при оценивании σ_u проблемы, связанные со слабыми инструментами, остаются.

Есть и другие случаи, когда $\hat{\pi}$ исчезает из тестовой статистики. Например, если $\dim(\theta) = \dim(Z)$, $\hat{\pi}$ будет квадратной матрицей и исчезнет из квадратичной формы, приводя к обычной χ^2-статистике. Следовательно, в этом случае, при условии, что $\tilde{\sigma}_0^2$ используется в качестве оценки для σ_u^2, тестовая статистика действительно является $\chi^2(\dim(\theta))$-распределенной случайной величиной. Этот случай интересен, поскольку он возникает в структурных векторных авторегрессиях с долгосрочными ограничениями (см. Pagan & Robertson, 1998), хотя там слабые инструменты возникают из-за близкого к свойствам рядов с единичным корнем поведения x_t, и локально нулевая асимптотика уже неприменима, так как тогда π должно быть порядка δ/T, а не $\delta/\sqrt{T}$.

Возвращаясь к случаю $\dim(\theta) = 1$ и $\dim(Z) > 1$, предположим, что берется другая оценка π. Для ее получения рассмотрим систему

$$\begin{aligned} y_t &= x_t\theta + u_t, \\ x_t &= z_t'\pi + \xi_t. \end{aligned}$$

Если

$$\begin{pmatrix} u_t \\ \xi_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u\xi} \\ \sigma_{u\xi} & \sigma_\xi^2 \end{pmatrix}\right),$$

то можно записать

$$\xi_t = \frac{\sigma_{u\xi}}{\sigma_u^2} u_t + \eta_t,$$

где η_t не зависит от u_t. Полагая $\theta = \theta_0$, можно оценить u_t, а также $\sigma_{u\xi}/\sigma_u^2$, так что регрессия $x_t - (\hat{\sigma}_{u\xi}/\hat{\sigma}_u^2)\hat{u}_t$ на z_t даст альтернативную оценку для π, скажем, $\tilde{\pi}$. Оказывается, это оценка ММПОИ при $\theta = \theta_0$. Ясно, что $\tilde{\pi}$ является функцией от η_t, а значит, не зависит от u_t, откуда следует, что (условно на Z) $Z'u$ не зависит от $\tilde{\pi}$. Эта независимость означает, что теперь при выводе распределения t-статистики, использующей $\tilde{\pi}$ вместо $\hat{\pi}$, его можно рассматривать условно на $\tilde{\pi}$. Следовательно, t-статистика будет асимптотически нормальна. Замена $\hat{\sigma}^2$ на σ_0^2 и $\hat{\pi}$ на $\tilde{\pi}$ означает применение LM-теста на $\theta = \theta_0$, что и предложил Kleibergen (2002). Эта статистика напрямую обобщается на случай $\dim(\theta) > 1$. Как и в большей части литературы по этой тематике, критическим предположением является экзогенность Z (или возможность преобразований условно на Z), что не выполняется в случае, когда слабость инструментов возникает из-за поведения наподобие рядов с единичным корнем.

Проблемы, вызванные свойствами $\hat{\pi}$, можно решить другими способами. Одно из предложений – рассматривать распределение условно на какой-нибудь функции от $\hat{\pi}$, в частности параметре концентрации $\phi = \sigma_\xi^{-2}\pi' Z'Z\pi$. Его оценка $\hat{\phi}$ – это F-статистика для гипотезы $\pi = 0$. Hillier & Forchini (2004) обсуждают логику таких действий, утверждая, что естественно брать распределение $\hat{\theta}$ условно на результате теста $\pi = 0$, поскольку он содержит информацию о θ из выборки. Авторы получают моменты этого условного распределения и сравнивают метод с МНК-оцениванием. Если параметр концентрации очень мал, результаты говорят о предпочтительности МНК-оценки. Hillier & Forchini (2004) также выводят распределения для случая $\dim(\theta) > 1$, используя минимальное собственное значение многомерного аналога параметра концентрации. Они отмечают, однако, что эти выражения настолько сложны для расчетов, что лучше выбрать какую-нибудь альтернативную функция от π.

Moreira (2003) рассматривает тест отношения правдоподобия (LR) для гипотезы $\theta = \theta_0$. Если F-статистику в тесте на равенство нулю коэффициентов при z_t в регрессии $x_t - (\hat{\sigma}_{u\xi}/\hat{\sigma}_u^2)\hat{u}_t$ на z_t обозначить за $\tilde{r}$ (что оценивает модифицированный «параметр концентрации» $r = \sigma_\eta^{-2}\pi' Z'Z\pi$), то, как показывает Kleibergen (2005),

$$LR = \frac{1}{2}[AR - \tilde{r} + \sqrt{(AR + \tilde{r})^2 - 4\tilde{r}(AR - LM)}].$$

Как было отмечено ранее, AR и $\tilde{r}$ должны быть независимы, а значит, в этом случае логично искать распределение условно на $\tilde{r}$. Распределение этого условного LR-теста (CLR) можно найти численно по алгоритму, недавно предложенному в Andrews, Moreira & Stock (2005).

Все внимание в описанной выше литературе уделялось тому, чтобы отыскать тестовую статистику, при определенных условиях независимую от π. Хотелось бы также, чтобы тест был мощным, и этим вопросом в последнее время занимались Poskitt & Skeels (2005) (используя асимптотику малого параметра концентрации) и Andrews, Moreira & Stock (2006). Последние

рассматривают построение точечно оптимальных инвариантных тестов. Они также отмечают, что функция мощности для LM-теста немонотонна, и CLR-тест является лучшим из трех перечисленных тестов.

Близкая проблема, которую только недавно начали изучать, – тестирование ограничений на подмножества θ, например, $\theta' = (\theta_1', \theta_2')$, и требуется проверить гипотезу $H_0 : \theta_1 = \theta_{10}$. Если к параметрам θ_2 не относятся проблемы со слабыми инструментами, то результаты, перечисленные выше, выполнены, поскольку инструментальная оценка θ_2 имеет «хорошие» свойства. Если же они также подвержены смещению из-за слабости инструментов, следует внести некоторые поправки. В недавних статьях Dufour & Taamouti (2005) и Kleibergen (2007) исследуется, каким образом это можно сделать; в первой статье используются проекционные методы для получения стандартных тестов, тогда как во второй устанавливаются некоторые ограничения на распределение $\hat{\theta}_1$.

Множество и других вопросов рассматривается в литературе. Один из них касается ОММ-оценивания. ОММ-оценка имеет вид

$$\hat{\theta} - \theta_0 = \left(-\sum_t \frac{\partial m_t}{\partial \theta} \right)^{-1} \sum_t m_t(\theta_0),$$

где обычно предполагается, что $T^{-1} \sum_t \partial m_t / \partial \theta$ сходится к константе. Но может случиться, что $\mathbb{E}[\partial m_t / \partial \theta]$ либо равняется нулю, либо очень мало, и тогда возникают те же проблемы, что и со слабыми инструментами (в простейшем случае с инструментальными переменными $-\partial m_t / \partial \theta = z_t x_t$). Таким образом, может оказаться, что асимптотика для ОММ-оценки не работает. Используя обобщенное информационное неравенство, имеем $\mathbb{E}[-\partial m_t / \partial \theta] = \mathbb{E}[m_t L_{\theta t}]$, где $L_{\theta t}$ – скор-функции для θ, и тогда возникают весьма реальные проблемы с ОММ-оценкой, если выбраны моменты m_t, не коррелированные со скор-функциями, так что следует немного поэкспериментировать, чтобы определить, какие моменты являются подходящими. В такой ситуации полезно иметь теоретическую модель, поскольку для нее можно запустить симуляции при выбранных значениях параметров и изучить вопрос о выборе моментов с помощью численных методов. Несовпадение распределения ОММ-оценки с тем, которое предполагает асимптотическая теория, хорошо изучено, например, в Kocherlakota (1990).

5 Примеры слабых инструментов

Оценивание кривой Филлипса в новокейнсианской модели

Новокейнсианская модель, приобретающая все большую популярность в макроэкономических исследованиях, имеет вид

$$\begin{aligned} \pi_t &= \delta_1 \mathbb{E}_t[\pi_{t+1}] + \delta_2 \pi_{t-1} + \lambda x_t + \varepsilon_{AS,t}, \\ x_t &= \mu \mathbb{E}_t[x_{t+1}] + (1-\mu) x_{t-1} - \phi(r_t - \mathbb{E}_t[\pi_{t+1}]) + \varepsilon_{IS,t}, \\ r_t &= \rho r_{t-1} + (1-\rho)(\beta \mathbb{E}_t[\pi_{t+1}] + \gamma x_t) + \varepsilon_{MP,t}, \end{aligned}$$

где π_t – темп инфляции, y_t – разрыв между текущим и потенциальным уровнями выпуска, а r_t – ставка процента, определяемая правилом денежной политики. В модели присутствуют три типа шоков: шок совокупного предложения $\varepsilon_{AS,t}$, шок со стороны кривой IS (шок спроса) $\varepsilon_{IS,t}$, и шок денежной политики $\varepsilon_{MP,t}$. Пусть кривая Филлипса оценивается стандартным образом, то есть предполагаются рациональные ожидания, и $\mathbb{E}_t[\pi_{t+1}]$ заменяется на π_{t+1}. Тогда требуется инструмент для π_{t+1}. Но инструмент также нужен и для x_t, так как это эндогенная переменная. Какие инструменты доступны? Известно, что в решении этой системы (при отсутствии серийной корреляции шоков) π_t, x_t и r_t являются функциями только своих первых лагов. Следовательно, среди доступных инструментов только x_{t-1}, r_{t-1} и π_{t-1}.

Но π_{t-1} уже присутствует в выражении для кривой Филлипса, так что в итоге в качестве инструментов для π_{t+1} остаются только x_{t-1} и r_{t-1}. Как покажет любая регрессия, эти инструменты являются слабыми, то есть фактически оценить δ_1 и δ_2 невозможно. На практике часто предполагают ограничение $\delta_2 = 1 - \delta_1$ («гибридная модель»), что преобразует кривую Филлипса в

$$\Delta\pi_t = \delta_1 \mathbb{E}_t[\pi_{t+1} - \pi_{t-1}] + \lambda x_t + \varepsilon_{AS,t},$$

и тогда π_{t-1} оказывается хорошим инструментом для $\pi_{t+1} - \pi_{t-1}$. Таким образом, иногда проблему слабых инструментов можно обойти, адаптируя модель из некоторых теоретических соображений.

Уравнения Эйлера в задачах управления запасами

Слабые инструменты возникают в различных ситуациях. Иногда выбору инструментов не уделяется должного внимания, о чем свидетельствуют, например, дискуссии об отдаче от образования, когда выводы очень чувствительны к выбору инструментов. В других ситуациях сложности могут возникать либо из-за природы используемой модели, либо из-за взаимодействия свойств модели с данными. Возможным примером последней причины может служить «тест на общие факторы» Vahid & Engle (1993), когда необходимы инструменты для темпов роста выпуска и потребления, а найти корреляцию между этими темпами роста удается редко.[3] В некоторых случаях возможно определить источник слабого инструмента. Рассмотрим такой пример.

Параметры, возникающие в системе условий первого порядка, связанной с уравнениями Эйлера, определяющими оптимальный выбор переменных управления, часто оценивают с помощью ОММ, что может привести к проблеме слабых инструментов. Во многих ситуациях эта возможность становится действительностью. Gregory, Pagan & Smith (1993) показали, что при использовании уравнений Эйлера из линейно-квадратичной оптимизационной модели инструменты нерелевантны при оценки коэффициента дисконтирования, если переменные роста являются интегрированными первого порядка. Похожий пример приводится в статье Fuhrer, Moore & Schuh (1995). Авторы обнаружили крайне плохие результаты применения ОММ-оценки в определенных условиях, даже при числе наблюдений, измеряемом тысячами.

Fuhrer, Moore & Schuh (1995) получают условия первого порядка, описывающие оптимальный выбор запасов в линейно-квадратичной задаче минимизации

$$\sum_{j=0}^{\infty} \beta^j \mathbb{E}_t[C_Y(Y_{t+j}) + C_N(t_{t+j}, S_{t+j})]$$

при ограничении $N_{t+j} = N_{t+j-1} + Y_{t+j} - S_{t+j}$, где Y_t – выпуск, S_t – продажи, N_t – уровень запасов, и

$$\begin{aligned} C_Y(Y_{t+j}) &= (\delta/2)Y_{t+j}^2 + (\alpha/2)(\Delta Y_{t+j})^2, \\ C_N(N_{t+j}, S_{t+j}) &= (\phi/2)(N_{t+j} - \omega S_{t+j})^2. \end{aligned}$$

Показатели ОММ настолько плохи, что авторы рекомендуют использовать метод максимального правдоподобия, даже если используется неверная плотность (эта оговорка возникает из-за необходимости специфицировать форму процесса для продаж при выписывании

[3] Это, по существу, тест на серийную корреляцию после инструментирования, и, следовательно, распределение статистики зависит от других оцениваемых параметров. Поскольку инструментальная оценка параметров линейной модели подвержена влиянию слабых инструментов, это справедливо и для распределения соответствующей статистики. Этот вопрос подробно не рассматривается в литературе, но получено, что тесты на серийную корреляцию могут давать странные результаты, если они основаны на слабых инструментах.

функции правдоподобия). Имеет смысл задуматься об источнике плохих показателей ОММ-оценки в модели Fuhrer, Moore & Schuh (1995), так как понимание этих причин важно для оценки подобных рекомендаций.

Уравнение Эйлера для приведенной оптимизационной задачи выглядит как

$$\mathbb{E}_t \left[\delta \left(Y_t - \beta Y_{t+1}\right) + \alpha(\Delta Y_t - 2\beta \Delta Y_{t+1} + \beta \Delta Y_{t+2}) + \phi \left(N_t - \omega S_t\right)\right] = 0, \tag{3}$$

что дает систему условий на моменты

$$\mathbb{E} \left[z_t(\delta \left(Y_t - \beta Y_{t+1}\right) + \alpha(\Delta Y_t - 2\beta \Delta Y_{t+1} + \beta \Delta Y_{t+2}) + \phi \left(N_t - \omega S_t\right))\right] = 0, \tag{4}$$

где z_t – инструменты из информационного множества, используемые в условном ожидании. Из (4) ясно, что невозможно идентифицировать все входящие параметры, и исследователи прибегают к различным вариантам нормализации. Fuhrer, Moore & Schuh (1995) работают с пятью типами нормализации. Из них два, обозначим их A и B, имеют следующий вид: $A : \delta = 1$, $B : \phi = 1$. Для анализа этих различных условий используем два упрощающих предположения. Во-первых, предположим, что $S_t = S_{t-1} + e_t$, где $e_t \sim i.i.d.(0, \sigma^2)$, то есть продажи являются случайным блужданием. Во-вторых, пусть $\beta = 1$.[4] Когда S_t является $I(1)$-рядом, можно показать, что N_t также является $I(1)$, а ω – параметр коинтеграции. При этих предположениях нормализация типа A предполагает линейную модель, в которой ΔY_{t+1} является зависимой переменной, а $N_t - \omega S_t$ и $\Delta^2 Y_{t+2}$ – регрессорами, для последнего из которых нужны инструменты. При нормализации типа B, $N_t - \omega S_t$ является независимой переменной, а ΔY_{t+1} и $\Delta^2 Y_{t+2}$ – регрессоры, и для обоих из них требуются инструменты. В качестве инструментов обычно применяются приращения Y_t и S_t, поскольку переменные в уровнях являются плохими инструментами для ΔY_{t+1}, так как в этом случае $I(1)$-переменные берутся как инструменты для $I(0)$ переменных. $\Delta Y_{t+1} = \Delta S_{t+1} + \Delta^2 N_{t+1}$ и, следовательно, чтобы лаги ΔY_t были хорошими инструментами, необходимо наличие серийной корреляции в $\Delta^2 N_t$. Если вывести выражение для процесса N_t при предложенной спецификации для S_t, можно установить, что серийная корреляция в $\Delta^2 N_t$ очень мала. Это означает, что такие нормализации, как B, приводят к проблеме слабых инструментов. Следовательно, ясен источник вывода авторов о том, что (стр. 143) «в случае сглаживающей модели, нормализация B требует 30000 наблюдений для сходимости к истинному значению».

Этот пример показывает, почему так сложно выдвинуть общий подход к проблеме слабых инструментов. Если бы ряд S_t не был близок к ряду с единичным корнем, инструменты могли бы быть очень эффективными, так что исход сильно зависит от природы соответствующих процессов.[5]

Список литературы

Andrews, D.W.K., M.J. Moreira & J.H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74, 715–752.

Andrews, D.W.K., M.J. Moreira & J.H. Stock (2005). Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics*, в печати.

Dufour, J.-M. & M. Taamouti (2005). Projection-based statistical inference in linear statistical models with possibly weak instruments. *Econometrica* 73, 1351–1365.

Forchini, G. & G. Hillier (2003). Conditional inference for possibly unidentified structural equations. *Econometric Theory* 19, 707–743.

[4] В исходном анализе данных и симуляциях $\beta = 0{,}995$, и процесс для продаж по существу содержит единичный корень при слабой автокорреляции. Fuhrer, Moore & Schuh (1995) отмечают, что если процесс для продаж неверно специфицирован как AR(1)-процесс вместо AR(3), «оцененный лаговый коэффициент в AR(1) будет приблизительно равен сумме коэффициентов для AR(3)-процесса» (стр. 143). Это дает корень, равный 0,956.

[5] Тот же вывод имеет место для примера из статьи Gregory, Pagan & Smith (1993).

Fuhrer, J., G. Moore & S. Schuh (1995). Estimating the linear-quadratic inventory model: Maximum likelihood versus generalized method of moments. *Journal of Monetary Economics* 35, 115–158.

Gregory, A.W., A.R. Pagan & G.W. Smith (1993). Estimating linear quadratic models with integrated processes. Глава в *Models, Methods and Applications of Econometrics: Essays in Honor of A.R. Bergstrom* под редакцией P.C.B. Phillips. Cambridge: Basil Blackwell.

Hall, A.R., G.D. Rudebusch & D.W. Wilcox (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review* 37, 283–298.

Kleibergen, F. (2002). Pivotal statistics for testing parameters in instrumental variables regression. *Econometrica* 70, 1781–1803.

Kleibergen, F. (2006). Testing. Глава в *The New Palgrave Dictionary of Economics*, под редакцией S. Durlauf & L. Blume. Palgrave Macmillan, в печати.

Kleibergen, F. (2007). Subset statistics in the linear IV regression model. Manuscript, Brown University.

Kocherlakota, N.R. (1990). On tests of representative consumer asset pricing models. *Journal of Monetary Economics* 26, 285–304.

Moreira, M.J. (2003). A conditional likelihood ratio test for structural models. *Econometrica* 71, 1027–1048.

Nelson, C.R. & R. Startz (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63, 125–140.

Pagan, A.R. & J.C. Robertson (1998). Structural models of the liquidity effect. *Review of Economics and Statistics* 80, 202–217.

Poskitt, D.S. & C.L. Skeels (2005). Small concentration asymptotics and instrumental variables inference. Research Paper No.948, University of Melbourne.

Shea, J. (1997). Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics* 79, 348–352.

Staiger, D. & J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J.H. & J. Wright (2000). GMM with weak identification. *Econometrica* 68, 1055–1096.

Stock, J.H. & M. Yogo (2005). Testing for weak instruments in linear IV regression. Глава в *Identification and Inference for Econometric Models: A Festshrift in Honor of Thomas Rothenberg* под редакцией J.H. Stock & D.W.K. Andrews. Cambridge: Cambridge University Press.

Stock, J.H., J.H. Wright & M. Yogo (2002). A survey of weak instruments and weak identification in GMM. *Journal of Business & Economic Statistics* 20, 518–529.

Vahid, F. & R. Engle (1993). Common trends and common cycles. *Journal of Applied Econometrics* 8, 341–360.

Wang, J. & E. Zivot (1996). Inference on a structural parameter in instrumental variables regression with weakly correlated instruments. *Econometrica* 66, 1389–1404.

West, K.D. & D.W. Wilcox (1996). A comparison of alternative instrumental variable estimators of a dynamic linear model. *Journal of Business & Economic Statistics* 14, 281–293.

Weak instruments

Adrian Pagan

Queensland University of Technology, Brisbane, Australia

This essay is a selective guide to the literature on weak instruments. We use simple models to illustrate the problems raised by weak instruments and the suggestions that have been made to solve them. Because the literature is one that is still developing we only briefly touch on some recent developments.

Размышления об инструментальных переменных*

Кристофер Симс†

Принстонский Университет, Принстон, США

Рассматривается вопрос об использовании метода инструментальных переменных и метода моментов с точки зрения теории принятия решений. Априорные веры играют чрезвычайно важную роль, поэтому в случаях, когда инструменты, возможно, «слабые», или их количество велико относительно количества наблюдений, важно описать особенности правдоподобия вне рамок ИП- или ММП-оценивания и соответствующих асимптотических (т.е. локально квадратичных) стандартных ошибок. ИП и ОММ привлекательны из-за удобства вычислений в большинстве случаев и (ложного) убеждения, что они применимы при малом числе «предположений». Здесь обсуждаются подходы, позволяющие сделать последнее утверждение более обоснованным.

Ключевые слова: байесовский подход, обобщенный метод моментов, инструментальные переменные, слабые инструменты, выбор инструментов, энтропия
Классификация JEL: C11, C13, C44

1 Введение

Ситуации, где применяются ИП и ОММ, хорошо иллюстрируют ценность байесовского подхода, поскольку в этих случаях, в отличие от большинства ситуаций, при больших выборках трактовка байесовским анализом правдоподобия как апостериорной функции распределения (при плоском априорном распределении) оказывается не всегда эквивалентна классической (небайесовской) трактовке вероятных спецификаций до получения выборки, как если бы они были полноценными вероятными спецификациями после получения выборки.

Байесовский подход к выдаче научных результатов рассматривает весь анализ данных как отчет о форме функции правдоподобия в виде, наиболее удобном для читателя. В этом эссе мы рассматриваем инференцию, основанную на методе моментов, с этой точки зрения.

2 Литература

Описание механики байесовского анализа инструментальных переменных недавно пополнилось новыми результатам, в частности, хорошо отраженными в Geweke (1996) и Kleibergen & Zivot (2000). В ряде недавних работ, куда входят наряду с прочими Kitamura & Stutzer (1997), Zellner, Tobias & Ryu (1997) и Kim (2002), авторы пытаются «вывести» ИП, в некоторых случаях вместе с послевыборочными апостериорными распределениями, используя «автоматические» методы информационной теории с целью избежать использования явных априорных распределений. В одной из работ Филлипс и Чао рекомендуют использовать априорные распределения Джеффриза – другой класс «автоматически» генерируемых априорных распределений.

Исследования с использованием байесовской техники развивают понимание необычных особенностей функций правдоподобия, возникающих в моделях с одновременностью из-за

*Перевод С. Анатольева и А. Белякова. Цитировать как: Симс, Кристофер (2007) «Размышления об инструментальных переменных», Квантиль, №2, стр. 83–94. Citation: Sims, Christopher A. (2007) "Thinking About Instrumental Variables," Quantile, No.2, pp. 83–94.

†Адрес: Department of Economics, Princeton University, 104 Fisher Hall, Princeton, NJ 08544-1021, USA. Электронная почта: *sims@princeton.edu*

полюсов и нулей у отображения между приведенной формой (ПФ) и структурными параметрами. Часть этого эссе посвящена примерам, показывающим важность таких особенностей на практике. Исследования «автоматических» априорных и апостериорных распределений мотивированы привлекательностью утверждения, что ОММ и ИП требуют немного предположений, и одновременно неудовлетворенностью этим постулатом на практике. Данное эссе начинается с обсуждения слабых предположений.

3 Уклонение от предположений

Какие бы предположения не делались, все они не бесспорные, поэтому эконометристу всегда удобно иметь возможность заявить на критику профессиональной аудитории, что его выводы практически не зависят от каких-либо предположений. Модели, в которых сделано мало предположений, часто называют «непараметрическими», в отличие от перегруженных предположениями параметрических моделей. Некоторое время назад, когда спектральные методы впервые появились в эконометрике, вначале считалось, что те позволяют анализировать временные ряды, используя намного более слабые предположения, чем параметрические ARIMA-модели, которые были и все еще остаются общепринятыми в эконометрике. В обзорной статье в 1974 году (Sims, 1974) мной было показано, что это впечатление обманчиво. Спектральный метод приводит к *асимптотической* теории с малым количеством предположений, но при этом статистики частотной области сглаживаются посредством окон или ядер, сужающихся по мере роста выборки, а также предполагается гладкость представляющих интерес объектов частотных областей. Понятно, что существуют функции в классе интересующих нас объектов, которые допустимы при слабых предположениях, необходимых для вывода асимптотической теории, но они становится недопустимыми, если мы всерьез воспринимаем утверждения о неопределенности (о доверительных интервалах, о тестировании гипотез и т.п.) в имеющейся выборке. При любой ширине окна ядра будут существовать «гладкие» функции, которые осциллируют слишком сильно, чтобы их точно оценить ядром с этой шириной окна. Мы интуитивно применяем эти ограничения при оценке результатов, даже если они не обсуждаются в исследовательских отчетах. Эти проблемы применительно к тестированию гипотез обсуждаются в работе Faust (1999).

В сфере временных рядов быстро стало понятно, что использующие ядра методы спектрального анализа не обладают большей общностью. ARIMA-модели, в которых количество параметров систематически растет с увеличением размера выборки или как функция от данных, могут привести к асимптотической теории настолько же общей, как и методы спектрального анализа. Класс моделей временных рядов, которые можно хорошо аппроксимировать последовательностью параметрических моделей, не больше, в топологическом смысле, чем класс моделей, удовлетворяющих условиям гладкости, подразумевающимся в методах спектрального анализа. Гладкость функций в функциональном пространстве – предположение того же рода, что и норма затухания последовательностей в пространстве последовательностей.

Вышеописанные проблемы выявляются в той же форме в теории непараметрического оценивания регрессионных функций или функций плотности. По моему впечатлению, все-таки практики вне области временных рядов менее осведомлены о том, что явно непараметрическая модель – это всего лишь еще одна версия того, какая именно модель будет хорошо работать в конкретном приложении, а не способ получения результата с меньшим числом предположений.[1]

[1]Возможно, это не из-за того, или не только из-за того, что эконометристы, работающие со временными рядами, умнее. Спектральные модели во временных рядах очень неуклюжи, если требуется спрогнозировать будущее по данным на текущий момент. Так как это стандартная задача в прикладном анализе временных рядов, для практиков было облегчением получить теоретическое доказательство того, что непараметриче-

Проблема также возникает и в стандартных утверждениях, что анализ, основанный на МНК, ИП и ОММ, свободен от предположений о виде распределения возмущений. Эти заявления обычно используются для утверждений о неопределенности результатов, которые (для данных модели и выборки) оправданы только при более жестком ограничении класса распределений возмущений, чем это требуется в асимптотической теории. Во многих случаях выводы о распределениях, основанные на асимптотической теории, справедливы в точности при предположениях о нормальности ошибок и справедливы в приемлемом приближении для данной выборки, если возмущения близки к нормальным.

По сути, в конечном счете мы обосновываем предположение о нормальности, прибегая к асимптотической теории. Это наводит на мысль, что есть шансы, что это предположение робастно. Рассуждения с помощью энтропии говорят о том, что это предположение, тем не менее, может быть в некотором смысле и консервативным.

4 Модель

Рассмотрим модель

$$\underset{T\times 1}{y} = x\beta + \varepsilon = \underset{T\times k}{Z}\gamma\beta + \nu\beta + \varepsilon, \tag{1}$$

$$\underset{T\times 1}{x} = Z\gamma + \nu, \tag{2}$$

$$\mathbb{V}([\nu\beta + \varepsilon\ \nu]) = \Sigma. \tag{3}$$

Заметим, что ту же модель можно параметризовать «наоборот» без изменения класса возможных вероятностных моделей данных просто переменой мест y и x в этих формулах. Эту симметрию можно расширить, если записать $Y = [y\ x]$, $\Pi = [\gamma\beta\ \gamma]$ и $\eta = [\nu\beta + \varepsilon\ \nu]$, а модель – как

$$Y = Z\Pi + \eta. \tag{4}$$

В модели требуется, чтобы $k \times 2$ матрица Π имела ранг 1, и как только мы накладываем такое ограничение, модель в форме (4) пробегает то же множество вероятностных моделей для данных.

При использовании параметризации (1–2) правдоподобие не стремится к нулю при $\beta \to \infty$ и постоянном Σ, какой бы ни был размер выборки. Это происходит из-за того, что при очень больших β наилучшая подгонка достигается при очень малых $\|\gamma\|$, при этом γ выбирается так, чтобы $\beta\gamma$ обеспечивало бы наилучшую возможную подгонку в уравнении (1) для y. Это, конечно же, приводит в большинстве случаев к плохой подгонке в уравнении (2) для x, но в пределе при больших β и очень малых $\|\gamma\|$ правая часть уравнения для x будет приблизительно нулем. Таким образом, при $\beta \to \infty$ степень подгонки ухудшается, но только до определенного значения, которое не загоняет правдоподобие в нуль.

Это приводит к наличию ловушек для наивных байесовских подходов. Динамические Методы Монте-Карло (ДММК), применяемые непосредственно к правдоподобию, как если бы оно было ненормированной функцией плотности распределения, не сходятся, и аналитическое интегрирование параметров из правдоподобия может не дать результата или выдать несобственные маржинальные «апостериорные распределения».

Плоские априорные распределения для Π в (4) без ограничений на ее ранг (неограниченная приведенная форма, НПФ), напротив, дают интегрируемые апостериорные распределения,

ские модели врожденно не являются более общими. В непараметрических регрессиях, с другой стороны, локальное свойство ограничений гладкости, неявно присутствующее в ядерных методах, более обоснованно, чем ограничения, налагаемые, скажем, ортогональными полиномиальными, параметрическими моделями.

если размер выборки не слишком мал. Таким образом, разумно использовать Евклидово расстояние для введения метрики на подмногообразие матриц Π с уменьшенным рангом, после чего преобразовать плоские априорные распределения (по мере Лебега) на этом подмногообразии к координатам β и γ. Этот вывод сам по себе не гарантирует, что апостериорные распределения при этих несобственных априорных распределениях будут собственными, но это многообещающий подход. Несобственное априорное распределение для β и γ, возникающее в этом подходе, есть

$$\left|\frac{\partial\Pi}{\partial(\beta,\gamma)}\left(\frac{\partial\Pi}{\partial(\beta,\gamma)}\right)'\right|^{\frac{1}{2}} = \|\gamma\|\left(1+\beta^2\right)^{\frac{1}{2}}. \tag{5}$$

Это действительно приводит к собственным апостериорным распределениям.

Даже при таких априорных распределениях хвосты апостериорных распределений стремятся к нулю лишь с полиномиальной скоростью, причем степень этого полинома не увеличивается с ростом выборки. В линейных регрессионных моделях же с плоскими априорными распределениями, наоборот, хвосты уменьшаются с полиномиальной скоростью, которая увеличивается линейно с размером выборки.

Обсудим теперь, к чему это приводит при комбинировании априорных распределений с Гауссовскими хвостами или с хвостами, убывающими как полином высокого порядка, с правдоподобием, взвешенным с помощью (5). Пусть вначале априорное среднее и пик правдоподобия совпадают, а затем пик правдоподобия удаляется от априорного среднего, при этом формы правдоподобия и априорной плотности распределения остаются теми же. Тогда апостериорные среднее, медиана и мода сначала будут удаляться от априорного среднего (как ожидалось), но затем поменяют направление и в конце концов совпадут с априорным средним, когда пик правдоподобия окажется далеко от априорного среднего. Это ситуация изображена на Рис. 1.

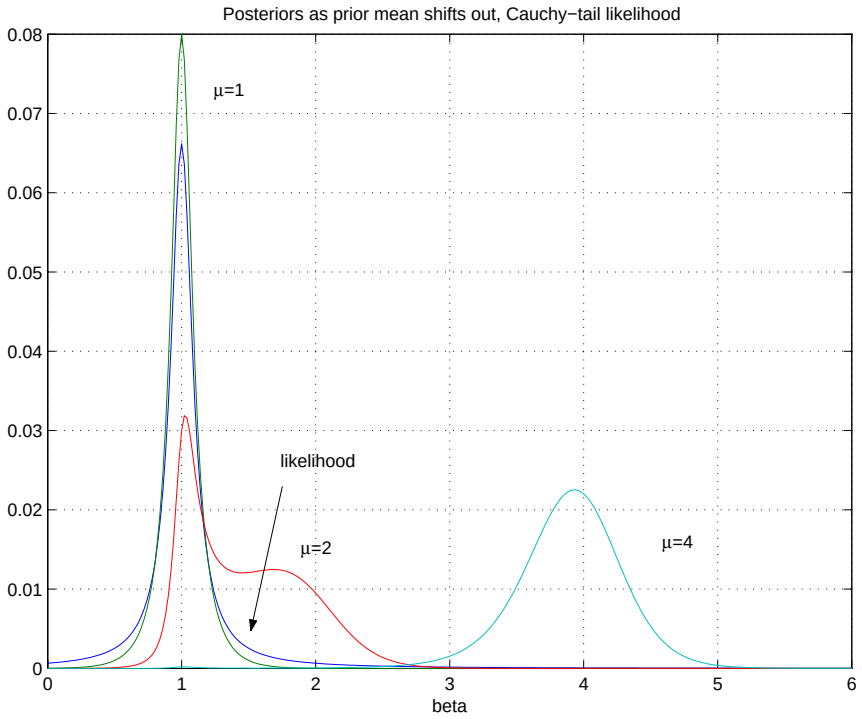

Рис. 1: Апостериорное распределение при увеличивающемся расстоянии между ММП-оценкой и априорным средним. Априорное t-распределение с 9-ю степенями свободы и $\sigma = 1/3$. Функция плотности – Коши с $\sigma = 0{,}1$, центрирована в 1.

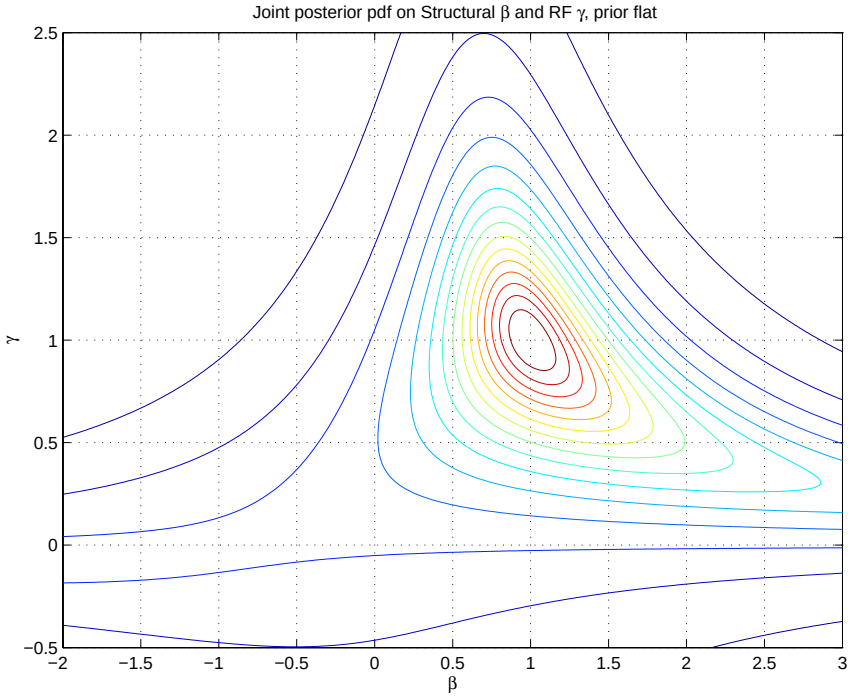

Рис. 2: Совместная плотность распределения структурного параметра β и ПФ-параметра γ. Плоское априорное распределение. Кроме того, $Z'Z = 4$, $\hat{\Sigma} = \begin{pmatrix} 0{,}67 & 0{,}33 \\ 0{,}33 & 0{,}67 \end{pmatrix}$, $\hat{\beta} = \hat{\gamma} = 1$.

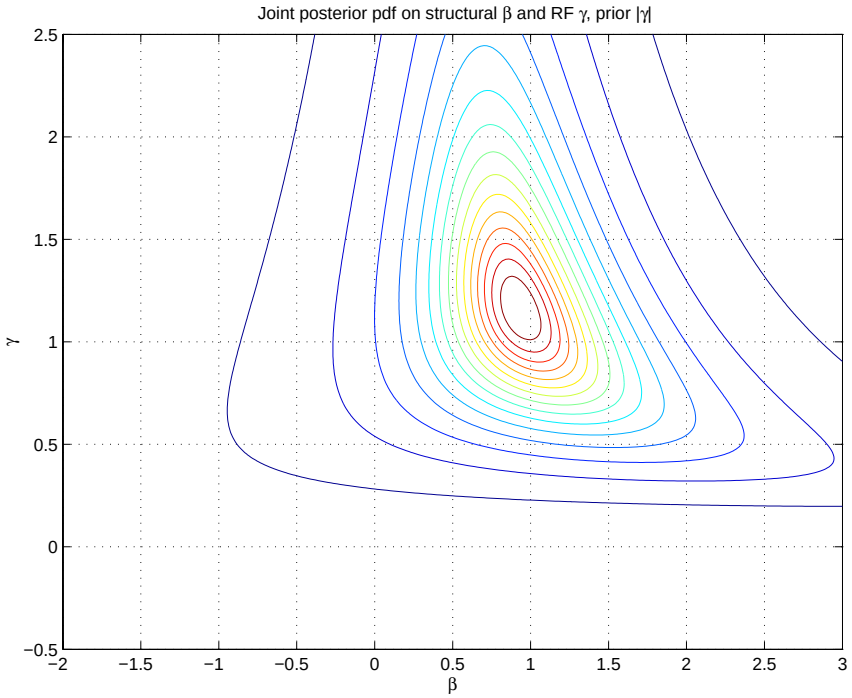

Рис. 3: Совместная плотность распределения структурного параметра β и ПФ-параметра γ. Априорное распределение для $|\gamma|$ (см. Рис. 2).

Другими словами, в модели инструментальных переменных тяжелые хвосты правдоподобия означают, что когда правдоподобие значительно отличается от априорного распределения, априорное распределение доминирует, даже если правдоподобие имеет пик достаточно резкий, чтобы заглушить априорную информацию вблизи него. Это имеет важные практические последствия для интерпретации ИП-оценок, которые отражены в общем признании того, что существует проблема «слабых инструментов». Но с байесовской точки зрения слабые инструменты – это не свойство популяции или даже выборки, а скорее свойство взаимосвязи выборки и априорных вер. ИП иногда дают «изменчивые» оценки, и практики это знают. Не существует, однако, способа «протестировать» оценки на изменчивость; надо учитывать априорные веры и факт, является ли выборочное правдоподобие в окрестности разумных величин для β действительно почти логарифмически плоским.

5 Совместные и маржинальные апостериорные распределения в случае точной идентификации

На Рис. 2 и 3 изображены линии уровня апостериорного распределения в случае точной идентификации, когда совместное апостериорное распределение легко выводится аналитически. Заметим его явно не эллиптическую форму, что означает, что априорная информация о γ будет сильно влиять на точность и форму апостериорного распределения β и наоборот. Также заметим, что в случае плоских априорных распределений для β и γ на графике видна довольно сложная картина вблизи $\|\gamma\| = 0$.

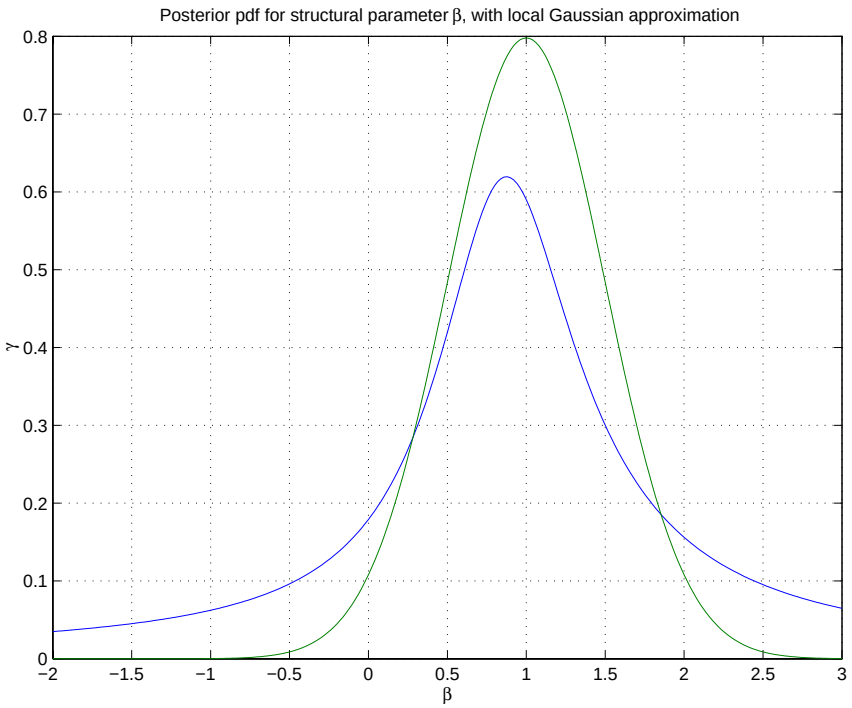

Рис. 4: Апостериорная функция плотности распределения для структурного параметра β. Локальная гауссовская аппроксимация (см. Рис. 2).

На Рис. 4 изображено маржинальное апостериорное распределение β, соответствующее Рис. 3. Заметим, насколько обманчива в этом случае локальная гауссовская аппроксимация.

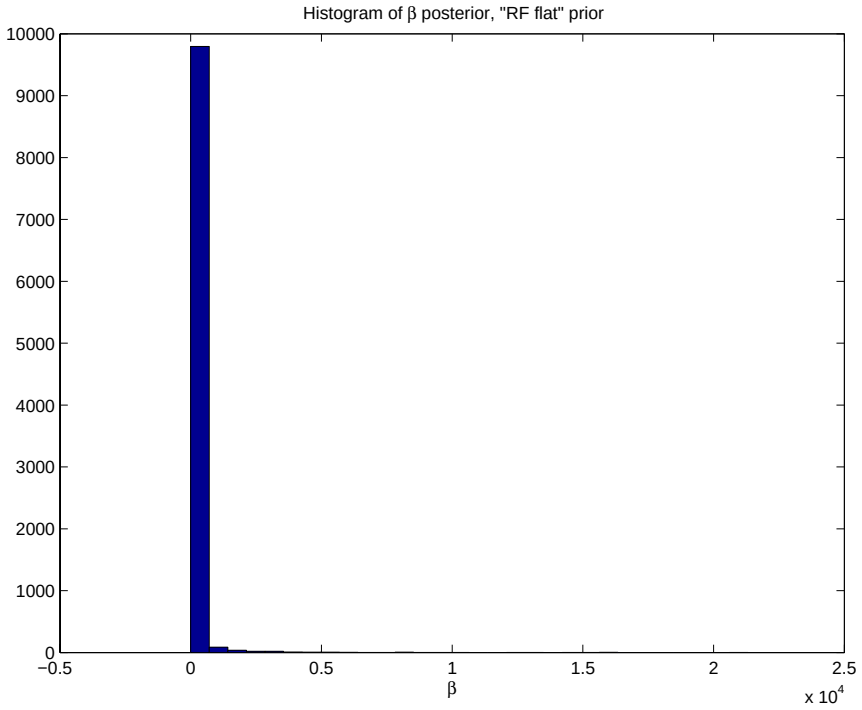

Рис. 5: Гистограмма апостериорного распределения β. Плоское априорное распределение для ПФ-параметра. Кроме того, $x = Z \begin{pmatrix} 1 \\ 0{,}1 \end{pmatrix} + \varepsilon$, $y = Z \begin{pmatrix} 1 \\ 0{,}1 \end{pmatrix} + v$, $\underset{20\times 2}{Z}, \varepsilon, v$ – все $N(0,1)$. Значение ИП-оценки $\hat{\beta} = 1{,}5132$, асимптотическая стандартная ошибка 0,8412.

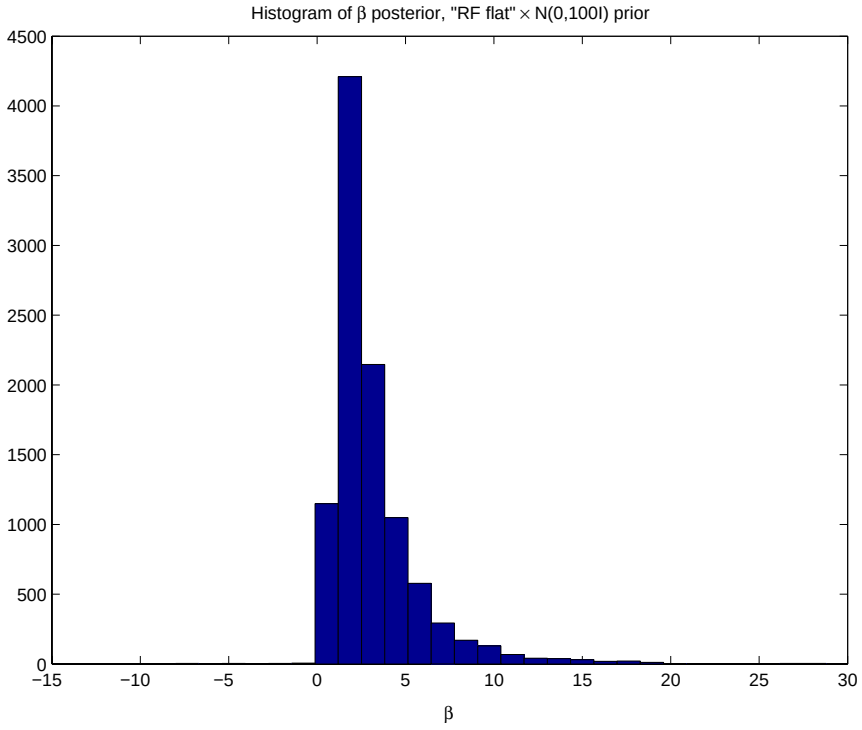

Рис. 6: Гистограмма апостериорного распределения β. Плоское априорное распределение $\times\, N(1, 100I)$ для ПФ-параметра. Модель и выборка как для Рис. 5.

6 Случай слабых инструментов

В выборке, где апостериорные распределения сильно не гауссовские со значительными и медленно спадающими хвостами, даже явная очень слабая априорная информация может существенно влиять на результаты анализа. На представленных далее графиках (Рис. 5–11) изображены эффекты наложения априорного распределения $N(0, 100I)$ на β и γ и в модели, и в выборке, что подразумевает оцененное с существенной неопределенностью значение β около 1. Даже такое слабое априорное распределение сильно влияет на апостериорное, в основном за счет устранения чрезвычайно вытянутых хвостов. Конечно же, апостериорные средние сильно зависят от добавления такой слабой априорной информации.

Для построения этих графиков я сгенерировал выборки размера 10000 при помощи ДММК, выбирая последовательно из условных правдоподобий для $\{\beta \,|\, \gamma, \Sigma\}$, $\{\gamma \,|\, \beta, \Sigma\}$ и $\{\Sigma \,|\, \gamma, \beta\}$, имеющих стандартную форму, и осуществляя один шаг алгоритма Метрополиса-Гастингса для отражения влияния априорного распределения. Когда априорное распределение не используется, в выборке возникают большие значения β, и когда β становиться большим, зависимость β от $\|\gamma\|$ очень сильная. Как известно, сильная зависимость приводит к медленной сходимости алгоритма Гиббса. Графики 10 и 11 показывают, как использование априорного распределения улучшает свойства сходимости алгоритма Гиббса путем устранения чрезвычайно больших экземпляров β.

7 Случай большого количества инструментов

Когда $T \leq k$, то есть когда нет степеней свободы в «регрессии на первом шаге», правдоподобие имеет два бесконечных пика: первый – там, где γ выбрана для идеальной подгонки x в (2), второй – где $\beta\gamma$ выбирается для идеальной подгонки y в (1). Обычно мы не считаем убедительными соответствующие оценки, которые основываются на МНК при прогоне регрессии y на x или при прогоне регрессии x на y, в последнем случае с оцениванием β обращением оценки коэффициента обратной регрессии. Причина в том, что мы на самом деле не верим, что любое из двух уравнений в НПФ имеет идеальную подгонку. Следовательно, имеет смысл попытаться отразить эти веры в априорном распределении, которое могло бы отвлечь от МНК-оценок.

Один из способов сделать это – использовать сопряженные априорные распределения, что можно осуществить добавлением «фиктивных наблюдений» в данные. Для определенности пусть $T = k = 20$, и зададим расширенные данные как

$$x^* = \begin{pmatrix} x \\ 0 \\ {\scriptstyle T\times 1} \end{pmatrix}, \quad y^* = \begin{pmatrix} y \\ 0 \\ {\scriptstyle T\times 1} \end{pmatrix}, \quad Z^* = \begin{pmatrix} Z \\ \lambda I \\ {\scriptstyle T\times T} \end{pmatrix}. \tag{6}$$

Вычисления, лежащие в основе Рис. 12, соответствуют $\lambda = 0{,}5$. Естественно считать, что эти фиктивные наблюдения стремятся «прижать» результаты к нулю. Однако из-за того, что они связывают априорные дисперсии γ и $\beta\gamma$ возле нуля с дисперсией возмущения в уравнениях, они также ослабляют «идеальные подгонки». На рисунке 12 видно, что последний эффект доминирует, так как применение априорного распределения сдвигает апостериорное вверх от априорного среднего (и от МНК-оценки). Этот подход перспективен как способ избежать утилизации информации из-за неиспользования несомненно годных и доступных инструментальных переменных.

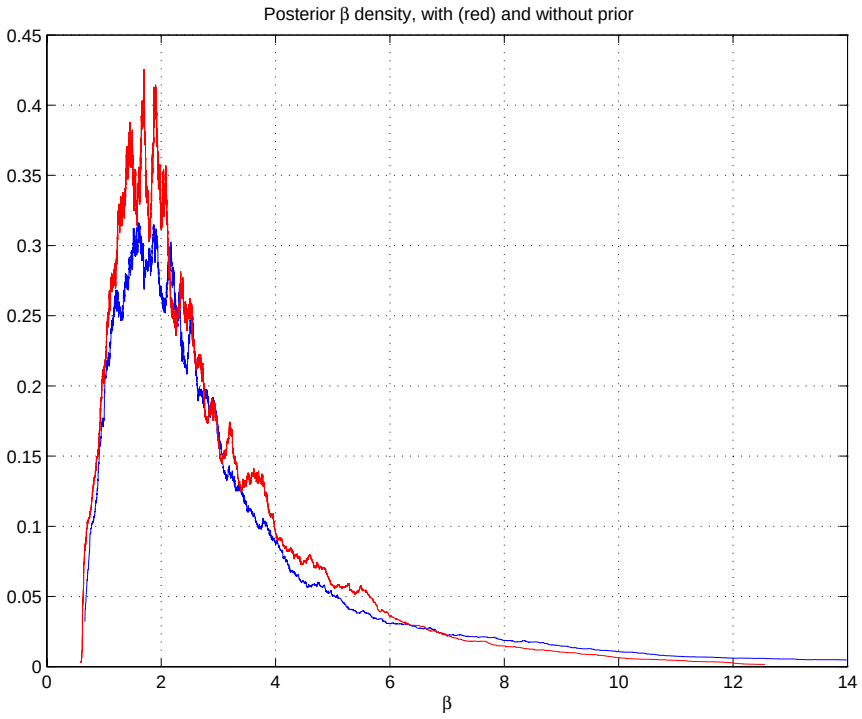

Рис. 7: Апостериорное распределение β при наличии априорного распределения (красным цветом) и без такового. Модель и выборка как для Рис. 5. Плотности оцениваются по методу «ближайших соседей», коих здесь 300.

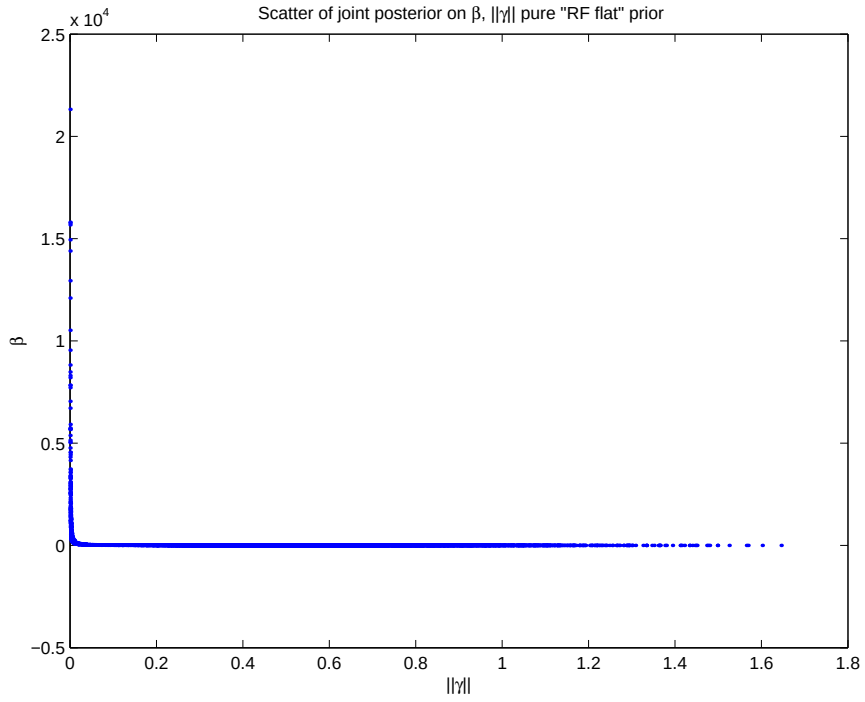

Рис. 8: Точки совместного апостериорного распределения β и $\|\gamma\|$ при плоском априорном распределении для ПФ-параметра. Модель и выборка как для Рис. 5.

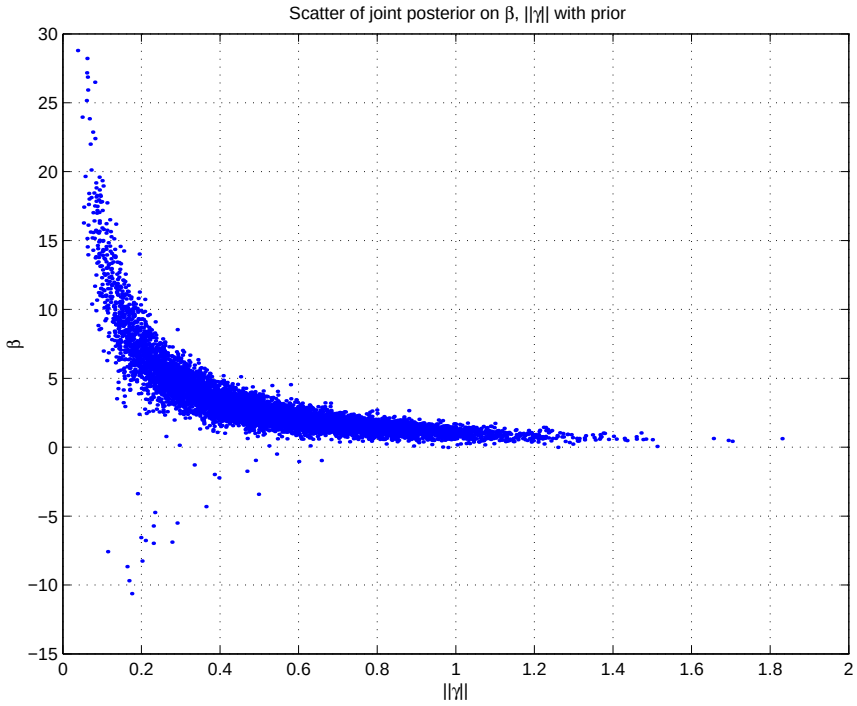

Рис. 9: Точки совместного апостериорного распределения β, $\|\gamma\|$ при априорном распределении для ПФ-параметра. Модель и выборка как для Рис. 5.

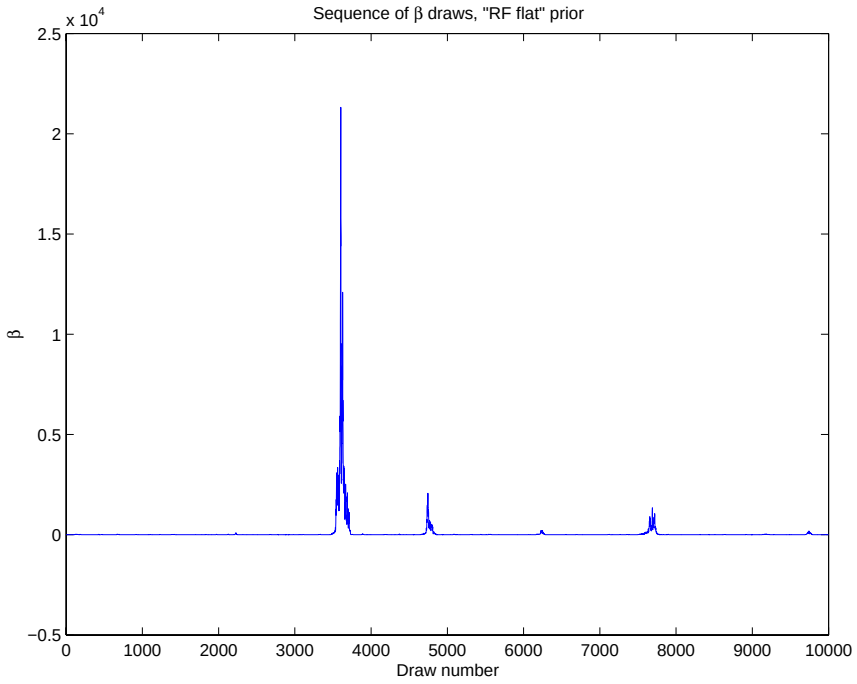

Рис. 10: Последовательность вытягиваний β. Плоское априорное распределение для ПФ-параметра. Модель и выборка как для Рис. 5.

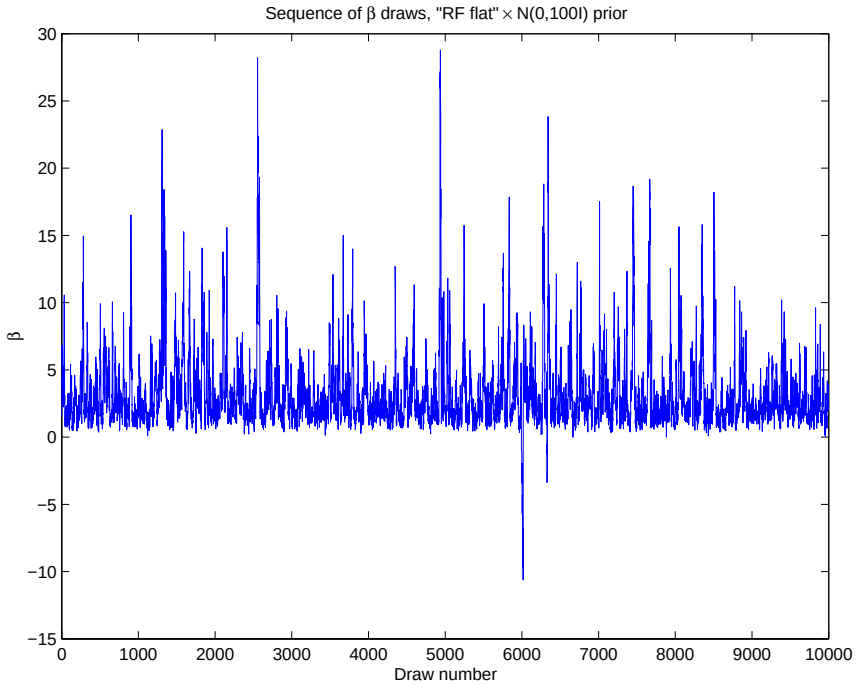

Рис. 11: Последовательность вытягиваний β. Плоское априорное распределение $\times N(1, 100I)$ для ПФ-параметра. Модель и выборка как для Рис. 5.

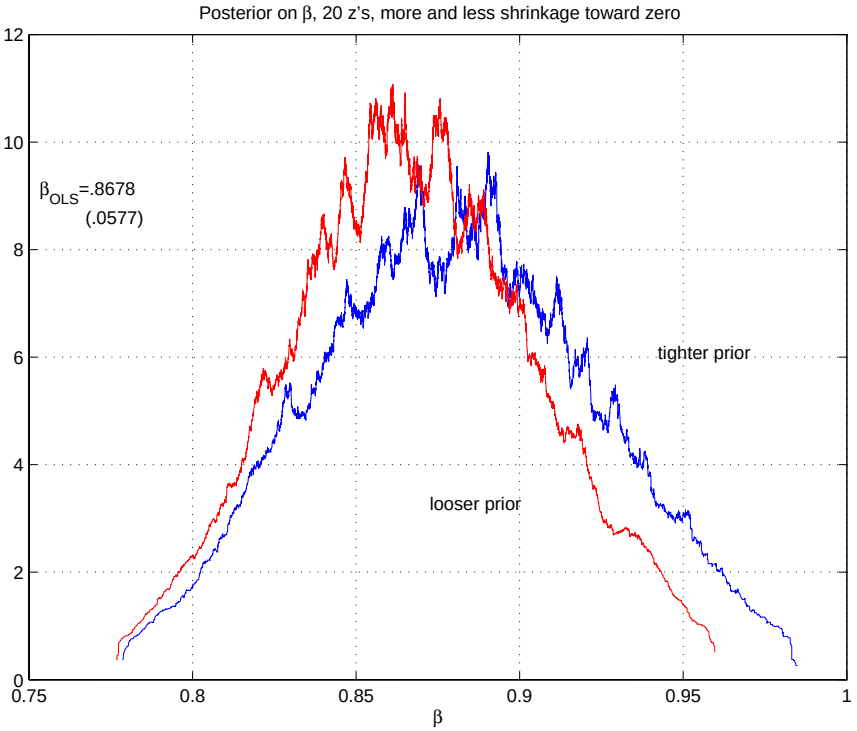

Рис. 12: Матрица Z размером 20×20 тянется из N(0,1). Коэффициенты приведенной формы при Z все равны 1. Выборка такова, что $\hat{\beta}_{OLS} = 0{,}8678$ с обычной стандартной ошибкой 0,0577. Полученная из обратной регрессии $\hat{\beta} = 0{,}9446$.

Список литературы

Faust, J. (1999). Conventional confidence intervals for points on spectrum have confidence level zero. *Econometrica* 67, 629–37.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* 75, 121–146.

Kim, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics* 107, 175–193.

Kitamura, Y. & M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65, 861–874.

Kleibergen, F. & E. Zivot (2000). Bayesian and classical approaches to instrumental variable regression. Technical report, Econometric Institute, Rotterdam.

Sims, C.A. (1974). Distributed lags. In M. Intrilligator & D. Kendrick (eds.). *Frontiers of Quantitative Economics II*. Amsterdam: North-Holland.

Zellner, A., J. Tobias & H.K. Ryu (1997). Bayesian method of moments (BMOM) analysis of parametric and semiparametric regression models. Technical report, University of Chicago.

Thinking about instrumental variables

Christopher A. Sims

Princeton University, Princeton, USA

We take a decision-theoretic view on the question of how to use instrumental variables and method of moments. Since prior beliefs play an inevitably strong role when instruments are possibly "weak", or when the number of instruments is large relative to the number of observations, it is important in these cases to report characteristics of the likelihood beyond the usual IV or ML estimates and their asymptotic (i.e. second-order local) approximate standard errors. IV and GMM appeal because of their legitimate claim to be convenient to compute in many cases, and a (spurious) claim that they can be justified with few "assumptions". We discuss some approaches to making such a claim more legitimately.

Keywords: Bayesian approach, GMM, instrumental variables, weak instruments, instrument selection, entropy

JEL Classification: C11, C13, C44