Полупараметрический анализ*

Даниэль Макфадден†

Калифорнийский Университет, Беркли, США

Настоящее эссе – обзор двух сфер применения полупараметрической эконометрики: анализа цензурированных данных о продолжительности занятости и анализа данных о заявленной готовности платить за природные ресурсы.

1 Введение

Многие эконометрические задачи можно рассматривать как один из вариантов следующей модели. Имеется случайный вектор $(Y, X) \in \mathbb{R}^k \times \mathbb{R}^m$, такой, что X имеет (неизвестную) плотность распределения $g(x)$, а Y почти наверное характеризуется (неизвестной) функцией условной плотности $f(y|x)$. Также известно преобразование $t(y, x)$ из $\mathbb{R}^k \times \mathbb{R}^m$ в множество действительных чисел $\mathbb{R}$, и условное математическое ожидание этого преобразования, $\theta(x) = \mathbb{E}[t(Y, x)|X = x]$, является объектом эконометрического исследования. Примерами подобных преобразований могут быть: (1) $t(y, x) \equiv y$, когда $\theta(x) = \mathbb{E}[Y|X = x]$ – математическое ожидание Y при условии $X = x$, или *функция регрессии* Y на x; (2) $t(y, x) = yy'$, когда $\theta(x) = \mathbb{E}[YY'|X = x]$ – матрица вторых условных моментов, а в комбинации с первым примером – условная дисперсия $\mathbb{E}[YY'|X = x] - (\mathbb{E}[Y|X = x])(\mathbb{E}[Y|X = x])'$; и (3) $t(y, x) = \mathbb{I}_A(y)$, то есть индикатор-функция множества A, когда $\theta(x)$ – вероятность события A при условии $X = x$. Примерами из экономических приложений могут быть вектор потребительского спроса Y и вектор дохода и цен x, или вектор чистого выпуска фирмы Y и вектор уровней постоянных затрат и цен на переменные факторы x.

Определим возмущение $\varepsilon = \varepsilon(y, x) \equiv t(y, x) - \theta(x)$. Тогда описанную выше постановку можно сформулировать в виде *обобщенной регрессионной модели*

$$t(y, x) = \theta(x) + \varepsilon,$$

где $\mathbb{E}[\varepsilon|x] = 0$. Эконометрические задачи, подходящие под эту модель, можно классифицировать как *полностью параметрические*, *полупараметрические* или *непараметрические*. Модель является полностью параметрической, если *априори* известно, что функция θ и распределение ошибки ε принадлежат семействам с конечным числом параметров. Модель является непараметрической, если о функциональных формах θ и ε ничего неизвестно, за исключением, возможно, некоторых свойств регулярности и формы, таких как непрерывная дифференцируемость или вогнутость. Модель является полупараметрической, если она содержит конечный вектор параметров, обычно представляющий первостепенный интерес, но части θ и/или распределение ε не ограничены семействами с конечным числом параметров. Это определение полупараметрической модели в довольно широком смысле, и оно включает, например, модель линейной регрессии при условиях Гаусса–Маркова, когда распределение ошибок не ограничено параметрическим семейством, и только первые два момента параметризованы. Некоторые эконометристы предпочитают применять термин «полупараметрическая модель» в тех ситуациях, когда задачу можно охарактеризовать с помощью

*Перевод Б. Гершмана и С. Анатольева. Цитировать как: Макфадден, Даниэль (2008) «Полупараметрический анализ», Квантиль, №5, стр. 29–40. Citation: McFadden, Daniel (2008) "Semiparametric analysis," Quantile, No.5, pp. 29–40.

†Адрес: University of California, Berkeley, Department of Economics, 549 Evans Hall #3880, Berkeley, CA 94720-3880, USA. Электронная почта: *mcfadden@econ.berkeley.edu*

конечномерного вектора параметров, являющегося объектом анализа, и бесконечномерного вектора шумовых параметров (который может, например, задавать неизвестную функцию), поскольку именно в таких случаях необходимы неклассические статистические методы.

Наиболее распространенный полупараметрический метод в эконометрике – это обыкновенный МНК, который оценивает параметры модели линейной регрессии, не требуя, чтобы распределение ошибок принадлежало семейству с конечным числом параметров. Современная литература по эконометрической теории расширила полупараметрические методы на различные нелинейные модели. Четыре крупнейшие пересекающиеся области их применения – это модели для цензурированных данных о продолжительности (например, продолжительности занятости), модели с ограниченной зависимой переменной (модели с частичной наблюдаемостью) для дискретных или цензурированных данных (например, о статусе занятости, количестве отработанных часов), модели для данных с (естественным или намеренным) эндогенным самоотбором выборки (например, модель определения заработной платы среди самоотобранных работников или модели для выборок типа «случай-контроль») и модели с аддитивными непараметрическими эффектами. В следующей таблице приведены некоторые приложения соответствующих моделей.

Модель	**Приложения**
Регрессионные и одноиндексные модели для цензурированных данных о продолжительности: $Y\|x \cong Y\|x'\beta$.	Продолжительность занятости, инновационные лаги, мобильность.
Модели с ограниченной зависимой переменной (например, дискретной или цензурированной): $Y^* = x'\beta - \varepsilon,\ \varepsilon\|x \sim F(\cdot)$. Преобразование наблюдаемости $Y = \Psi(Y^*)$: дискретное: $Y = \operatorname{sgn}(Y^*)$, цензурированное: $Y = \min(Y^c, Y^*)$.	Дискретная: статус занятости, выбор брэнда. Цензурированная: количество отработанных часов, уровни расходов.
Эндогенный самоотбор выборки: $Y = x'\beta - \varepsilon,\ \varepsilon\|x \sim f(\cdot),\ x \sim g(\cdot)$. Естественный: (Y, x) наблюдаются $\Leftrightarrow Y > 0$. Намеренный: (Y, x) участвуют в выборке $\Leftrightarrow Y > 0$.	Естественный: самоотобранные работники, домовладельцы. Намеренный: выборка типа «случай-контроль».
Аддитивные непараметрические эффекты: $Y = x'\beta + H(z) + \varepsilon$.	Устойчивый анализ политики.

В большинстве случаев основная задача полупараметрического анализа состоит в оценивании регрессионных коэффициентов, которые определяют положение распределения зависимой переменной; тогда неизвестное распределение является (бесконечномерным) шумовым параметром. Также в некоторых приложениях непосредственный интерес представляет некоторый функционал неизвестного распределения, например, условное математическое ожидание зависимой переменной. Конечной целью анализа могут быть точечные оценки или доверительные интервалы для исследуемых объектов или тестирование гипотез относительно параметров. Обычно важно получить меру точности получаемых оценок, включая скорости сходимости, асимптотические распределения и бутстраповские или другие показатели точности оценок в конечных выборках и качества асимптотических приближений.

Настоящее эссе не является обзором всего спектра полупараметрических моделей в эконометрике и не рассматривает свойства полупараметрических оценок, кроме как в иллюстративных примерах. Хороший обзор основ полупараметрического анализа можно найти в Powell (1994). В данном эссе рассматриваются лишь две сферы применения. Первая – это анализ цензурированных данных о продолжительности занятости – возможно, ведущая сфера

прикладного полупараметрического оценивания. Вторая – это анализ данных о заявленной готовности платить за природные ресурсы.

2 Модели для цензурированных данных о продолжительности занятости

В центре внимания литературы о продолжительности занятости находится воздействие объясняющих переменных, таких как пол, раса, возраст и уровень образования, на риск прекращения работы. Данные о продолжительности занятости обычно являются цензурированными, поскольку периоды занятости начинаются до начала панельного обследования (и дату начала периода не всегда возможно точно определить, используя ретроспективные вопросы) и/или продолжаются после его окончания, или же из-за выбывания объектов наблюдения из панели. В данном разделе рассматривается только цензурирование справа, то есть до окончания периода занятости. При параметрическом анализе моделей продолжительности обычно используются экспоненциальная или вейбулловская кривые выживания или модель пропорциональных рисков Кокса, которая является полупараметрической.

Horowitz & Newmann (1987), возможно, впервые применили на практике методы полупараметрической цензурированной регрессии для анализа данных о продолжительности занятости. Чтобы придать некоторое содержательное наполнение данному экономическому приложению, рассмотрим риски, которые могут привести к окончанию периода занятости. Во-первых, прекращение работы может быть инициировано работником (увольнение по собственному желанию) или работодателем (сокращение, увольнение). На решение работника об увольнении по собственному желанию воздействуют, по-видимому, неденежные характеристики работы (например, безопасность, разнообразие, установленные правила), альтернативные издержки занятости и характеристики работника, такие как уровень образования, раса, преданность работодателю. На решение фирмы об увольнении сотрудника воздействует ожидаемая производительность работника за вычетом заработной платы. Специфический человеческий капитал работника влияет как на альтернативные издержки занятости, так и на ожидаемую производительность. Альтернативные издержки занятости определяются также ожидаемыми страховыми выплатами по безработице и продолжительностью безработицы. Макроэкономические и продуктовые циклы воздействуют на ожидаемую производительность. Следующие аспекты этого словесного описания важны для моделирования продолжительности занятости:

1. Увольнение по собственному желанию и сокращение являются конкурирующими рисками с пересекающимися, но несовпадающими, наборами объясняющих переменных. При структурном оценивании продолжительности необходимо различать эти два вида рисков. Данные о том, заканчивается ли период занятости в результате увольнения по собственному желанию или нет, значительно способствуют идентификации и оцениванию отдельных рисков.

2. Важные объясняющие переменные, такие как уровень макроэкономической активности и запас специфического человеческого капитала работника, меняются во времени, так что структурная модель должна допускать меняющиеся во времени регрессоры. Это довольно легко учесть в случае дискретного времени, используя разнородные марковские модели, но весьма затруднительно в случае непрерывного времени.

3. Ненаблюдаемые переменные, такие как преданность сотрудника работодателю, различаются в популяции и самоотбираются в процессе выживания. Значит, при структурном моделировании продолжительности необходимо определить распределение этих ненаблюдаемых величин. Наличие ненаблюдаемой разнородности также приводит к самоотбору субпопуляции, которая начинает период занятости в интервале наблюдения. Субпопуляция, начинающая период занятости вблизи начала периода наблюдения, будет в

среднем менее преданной работодателю, чем все работники. Те работники, чей первый наблюдаемый период занятости начинается ближе к концу периода наблюдения, будут в среднем более преданными работодателю, если панель достаточно длинная.

4. В структурной модели продолжительности занятости риск должен зависеть исключительно от экономических переменных, но не напрямую от количества прошедшего времени. Следовательно, модели, предполагающие наличие необъясненного «базового» риска, удаляют вариацию, которая должна иметь структурные источники. С точки зрения структурного оценивания экономических факторов продолжительности занятости акцент на эффекте объясняющих переменных смещается при восприятии базового риска как шумового параметра.

5. Экономическая теория не дает конкретных функциональных форм или распределений ненаблюдаемых величин; предположение о том, что наблюдаемые величины входят в модель как параметрическая аддитивная комбинация следует обосновывать как аппроксимацию. Следовательно, анализ, который предполагает, что наблюдаемые величины входят в модель в виде конкретной аддитивной комбинации при неизвестных преобразованиях или распределениях, на самом деле предполагает слишком много о структуре аддитивной комбинации, и, возможно, слишком мало о неизвестных преобразованиях, которые можно достаточно точно аппроксимировать при помощи гибких семейств с конечным числом параметров.

Процесс, порождающий данные о продолжительности занятости, можно охарактеризовать при помощи *кривой выживания* $q(t|x)$, дающей долю популяции с периодами занятости, начинающимися в момент времени 0, которая доживает до момента времени t, при условии наблюдаемой динамики регрессоров $x(\cdot)$. Если присутствуют ненаблюдаемые регрессоры ξ, распределенные в исходной популяции в соответствии с функцией плотности $\nu(\cdot|x,0)$, а $q(t|x,\xi)$ – «структурная» кривая выживания, то процесс, порождающий данные, удовлетворяет следующему соотношению:

$$q(t|x) = \int_{-\infty}^{+\infty} q(t|x,\xi) \cdot \nu(\xi|x,0) d\xi. \tag{1}$$

Функция плотности ненаблюдаемых регрессоров при условии дожития меняется во времени из-за отбора и удовлетворяет уравнению

$$\nu(\xi|x,t) = \nu(\xi|x,0) \cdot \frac{q(t|x,\xi)}{q(t|x)}. \tag{2}$$

Кривую выживания также можно описать с помощью *функции риска*:

$$h(t|x,\xi) = -\nabla_t \ln(q(t|x,\xi)). \tag{3}$$

Средняя норма риска в выжившей популяции равна

$$\begin{aligned} h^*(t|x) &= -\nabla_t \ln(q(t|x)) = \\ &= \frac{\int_{-\infty}^{+\infty} h(t|x,\xi) q(t|x,\xi) \nu(\xi|x,0) d\xi}{q(t|x)} = \int_{-\infty}^{+\infty} h(t|x,\xi) \nu(\xi|x,t) d\xi. \end{aligned} \tag{4}$$

Обращая уравнение (3), получаем

$$q(t|x,\xi) = \exp\left(-\int_0^t h(s|x,\xi) ds\right) \equiv \exp(-\Lambda(t|x,\xi)), \tag{5}$$

где $\Lambda(t|x,\xi)$ – так называемый *интегральный риск*. Средняя продолжительность завершенных периодов занятости равна

$$\mathbb{E}[t|x,\xi] = -\int_0^\infty t \cdot \nabla_t q(t|x,\xi)dt = \int_0^\infty q(t|x,\xi)dt, \tag{6}$$

где второе равенство получено путем интегрирования по частям.

Когда интервал наблюдения конечен, некоторые периоды занятости *прерываются* или *цензурируются справа*; функция выживания, определенная вплоть до момента цензурирования, продолжает характеризовать процесс, порождающий данные. Средняя продолжительность периода занятости, завершенного естественным образом (в момент времени t) или в результате цензурирования (в момент времени t^c) равна

$$\mathbb{E}[\min(t,t^c)] = -\int_0^{t^c} t \cdot \nabla_t q(t|x,\xi)dt + t^c q(t^c|x,\xi) = \int_0^{t^c} q(t|x,\xi)dt. \tag{7}$$

Аналогичные формулы справедливы для средней нормы риска.

При наличии выбывания из выборки момент цензурирования становится случайной величиной с соответствующей функцией выживания $r(t^c|x,\xi)$. В этом случае вероятность того, что наблюдение периода занятости продолжается до момента t, равна $q(t|x,\xi)r(t|x,\xi)$; общий риск завершения наблюдаемого периода занятости естественным путем или в результате цензурирования равен $h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi)$; для периода, заканчивающегося в момент времени t, вероятность цензурирования равна $h(t|x,\xi)/(h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi))$, а средняя продолжительность наблюдаемых периодов занятости равна

$$\int_0^\infty q(t|x,\xi)r(t|x,\xi)dt.$$

Примером параметрической модели продолжительности, когда вектор x неизменен во времени, является модель *Вейбулла*:

$$q(t|x) = \exp(-t^\alpha e^{-x'\beta}), \tag{8}$$

где α – положительный параметр, β – вектор параметров, а x – вектор регрессоров. Соответствующая функция риска имеет вид

$$h(t|x) = \alpha t^{\alpha-1} e^{-x'\beta}, \tag{9}$$

а средняя продолжительность завершенных периодов равна

$$\mathbb{E}[t|x] = e^{x'\beta/\alpha}\Gamma(1+1/\alpha), \tag{10}$$

где $\Gamma(\cdot)$ – гамма-функция. При $\alpha = 1$ получаем *экспоненциальную* модель продолжительности.

Имеются три стратегии статистического оценивания цензурированных данных о продолжительности:

1. Полностью параметрический подход, когда предполагается, что $q(t|x)$ или, в случае ненаблюдаемой разнородности, $q(t|x,\xi)$ и $\nu(\xi|x,0)$ принадлежат семействам с конечным числом параметров.[1]

[1]Типичными примерами являются предположение о вейбулловском или логнормальном распределении для $q(t|x)$ или экспоненциальном распределении для $q(t|x,\xi)$ в комбинации с гамма-распределением для ξ. Параметры распределения можно оценить методом максимального правдоподобия.

2. Полностью непараметрический подход, когда $q(t|x)$ оценивается без каких-либо параметрических ограничений, например, при помощи оценки Каплана–Мейера.[2]

3. Одноиндексный полупараметрический подход, когда $q(t|x)$ зависит от x через скалярную функцию $V(x,\beta)$, которая известна, за исключением конечного вектора параметров β, но $q(t|v)$ не ограничивается параметрическим семейством. В случае ненаблюдаемой разнородности либо $q(t|v,\xi)$, либо $\nu(\xi|v,t)$ могут быть непараметрическими (но не оба одновременно, если нет дополнительных ограничений, ввиду требований идентификации).[3]

Рассмотрим некоторые альтернативные варианты полупараметрических моделей, которые предлагаются в литературе. Пусть x – вектор регрессоров, предполагаемый *неизменным во времени*. Пусть далее β – вектор неизвестных параметров, $V(x,\beta) \equiv x'\beta$ – одноиндексная функция с неизвестными параметрами β, а $q(t|x'\beta)$ – функция выживания. Пусть T^* – случайная величина, обозначающая количество прошедшего времени, а T^c – момент цензурирования, так что наблюдаемая продолжительность соответствует $T = \min(T^*, T^c)$. Имеются четыре альтернативные модели для T:

1. *Модель регрессии*: $\ln T^* = x'\beta + \varepsilon$, где $\varepsilon|x$ имеет неизвестную плотность распределения $f(\varepsilon)$ с нулевым средним. Относительно функции плотности $f(\cdot)$ часто предполагают симметричность и гомоскедастичность. Модели соответствует следующая функция выживания:

$$q(t|x'\beta) = 1 - F(\ln t - x'\beta), \tag{11}$$

где $F(\cdot)$ – кумулятивная функция распределения для $f(\cdot)$. Соответствующая функция риска имеет вид

$$h(t|x'\beta) = \frac{f(\ln t - x'\beta)}{t[1 - F(\ln t - x'\beta)]}. \tag{12}$$

Обобщение этой модели допускает гетероскедастичность ε, когда дисперсия зависит от индекса $x'\beta$, или, в более общем случае, от некоторой другой функции от x. *Модель цензурированной регрессии* – это просто модель вида

$$\ln T = \min(\ln T^c, x'\beta + \varepsilon). \tag{13}$$

В случае неслучайного цензурирования она обладает тем свойством, что

$$\mathbb{E}[\ln T|x] = \int [1 - F(y - x'\beta)]dy \tag{14}$$

[2]Классическая оценка Каплана–Мейера формулируется для данных о продолжительности в случае отсутствия регрессоров. Предположим, что в данных периоды занятости, начинающиеся в один и тот же момент времени 0, прерываются (естественным образом или в результате цензурирования) в моменты времени $t_1 < \cdots < t_J$. Пусть n_j обозначает число периодов, которые завершаются естественным образом в момент времени t_j, а m_j – число периодов, цензурируемых в этот момент времени. Общее число периодов, находящихся «в группе риска» в момент времени t_j, равно $N_j = \sum_{i=j}^{J}(n_i + m_i)$. Оценка Каплана–Мейера для функции риска в момент t_j имеет вид $h^*(t_j) = n_j/N_j$. Соответствующая оценка функции выживания имеет вид $q^*(t_j) = (1-h^*(t_j))q^*(t_{j-1})$, или $q^*(t_j) = \prod_{i=1}^{j}(1-n_j/N_j)$. При наличии категориальных регрессоров оценка Каплана–Мейера, очевидно, применяется отдельно для каждой клетки для всех возможных комбинаций регрессоров. Используя идею оценки ближайших соседей из непараметрического регрессионного анализа, оценку Каплана–Мейера можно адаптировать для общего случая некатегориальных регрессоров. В случае ненаблюдаемой разнородности, вообще говоря, невозможно идентифицировать функции выживания и плотность распределения ненаблюдаемых регрессоров, когда оба этих объекта являются непараметрическими. Heckman & Singer (1984) установили этот результат, а также предложили полупараметрические методы для оценивания параметрической структурной функции выживания $q(t|x,\xi,\beta)$ при наличии непараметрической плотности распределения разнородности $\nu(\xi|x,0)$.

[3]Другие полупараметрические подходы включают многоиндексные модели и методы параметризации квантилей без полной параметризации распределения.

является возрастающей функцией от $x'\beta$.

2. *Модель с преобразованием (обобщенная модель Бокса–Кокса).* Предположим, G является неизвестным монотонно возрастающим преобразованием из $(0, +\infty)$ на множество действительных чисел, и предположим, что

$$G(T^*) = x'\beta + \varepsilon, \tag{15}$$

где $\varepsilon|x$ имеет известную или неизвестную плотность распределения $f(\varepsilon)$. Соответствующая функция выживания имеет вид

$$q(t|x'\beta) = 1 - F(G(t) - x'\beta), \tag{16}$$

а соответствующая функция риска –

$$h(t|x'\beta) = \frac{G'(t)f(G(t) - x'\beta)}{1 - F(G(t) - x'\beta)}. \tag{17}$$

Опять же, модель можно обобщить на случай гетероскедастичности относительно $x'\beta$.

3. *Целенаправленное проецирование (одноиндексная регрессия).* Предположим, H – неизвестное преобразование из множества действительных чисел в себя. Предположим, что

$$\ln T^* = H(x'\beta) + \varepsilon, \tag{18}$$

где $\varepsilon|x$ имеет известную или неизвестную плотность распределения $f(\varepsilon)$. Соответствующая функция выживания имеет вид

$$q(t|x'\beta) = 1 - F(\ln t - H(x'\beta)), \tag{19}$$

а функция риска –

$$h(t|x'\beta) = \frac{f(\ln t - H(x'\beta))}{t[1 - F(\ln t - H(x'\beta))]}. \tag{20}$$

Распределение ошибок обычно предполагается гомоскедастичным, но некоторые оценки этой модели допускают гетероскедастичность относительно $x'\beta$.

4. *Модель пропорциональных рисков.* Предположим, что $h_0(t)$ – неизвестная неотрицательная функция «базового риска», а регрессоры оказывают пропорциональный эффект на риск, то есть

$$h(t|x) = h_0(t)\exp(-x'\beta). \tag{21}$$

Определим базовый интегральный риск:

$$\Lambda_0(t) = \int_0^t h_0(s)ds. \tag{22}$$

Тогда функция выживания принимает вид

$$q(t|x'\beta) = \exp(-\Lambda_0(t)e^{-x'\beta}), \tag{23}$$

и

$$\ln \Lambda_0(T^*) = x'\beta + \varepsilon, \tag{24}$$

где ε имеет распределение экстремальных значений:

$$F(\varepsilon) = 1 - \exp(-e^{-\varepsilon}). \tag{25}$$

Другие распределения ошибки можно получить из модели пропорциональных рисков с ненаблюдаемой разнородностью. Например, следуя работе Lancaster (1979), предположим, что

$$h(t|x,\xi) = h_0(t)\exp(-x'\beta)\xi, \tag{26}$$

где ξ имеет гамма-распределение, $\nu(\xi|x,0) = \xi^{\theta-1}e^{-\xi}/\Gamma(\theta)$. Тогда, применяя соотношение (1), получаем

$$q(t|x) = \left(1 + e^{\Lambda_0(t)-x'\beta}\right)^{-\theta}, \tag{27}$$

откуда следует, что выполняется уравнение (15), когда ε имеет обобщенное логистическое распределение (или e^{ε} имеет распределение Парето):

$$F(\varepsilon) = 1 - (1 + e^{\varepsilon})^{-\theta}. \tag{28}$$

Средний риск для (26) равен

$$h^*(t|x) = \frac{\theta h_0(t)e^{\Lambda_0(t)}}{e^{\Lambda_0(t)} + e^{x'\beta}} \tag{29}$$

и больше не принимает форму пропорциональных рисков. Условное распределение ненаблюдаемых регрессоров при данной функции выживания $\nu(\xi|x,t)$ остается гамма-распределением с параметром θ, но относительно преобразованной величины $(1 + e^{\Lambda_0(t)-x'\beta})\xi$.

Модель пропорциональных рисков (21) является частным случаем модели с преобразованием, когда ошибка имеет распределение (25). Модель пропорциональных рисков с разнородностью (26) – это также частный случай модели с преобразованием. Когда базовый риск является степенной функцией от t, $h_0(t) = \alpha t^{\alpha-1}$, модель (21) упрощается до параметрической вейбулловской модели продолжительности, а также может быть интерпретирована как модель цензурированной регрессии с ошибкам, имеющими распределение экстремальных значений.

Общая «аддитивная одноиндексная модель», включающая как частные случаи четыре описанные модели, имеет вид

$$G(T^*) = H(x'\beta) + \varepsilon, \tag{30}$$

где ε имеет кумулятивную функцию распределения $F(\cdot)$. Соответствующая функция выживания имеет вид

$$q(t|x'\beta) = 1 - F(G(T) - H(x'\beta)). \tag{31}$$

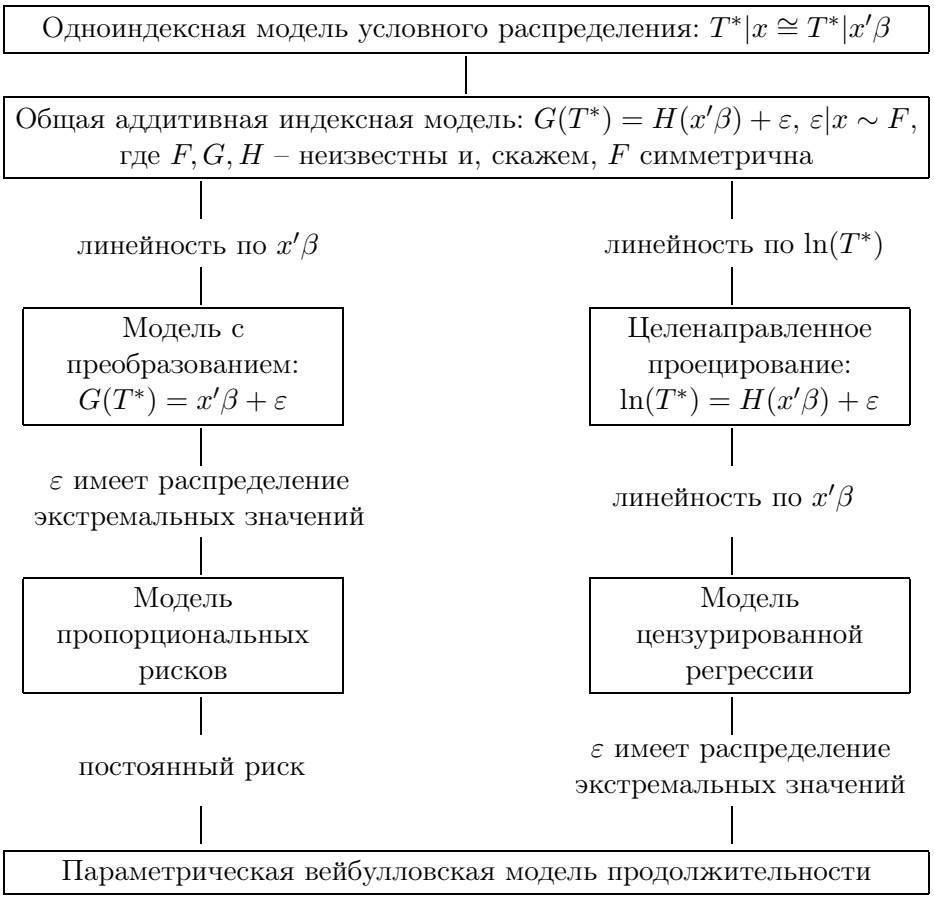
На рисунке 1 показана логическая связь между этими моделями. Все они являются частными случаями *одноиндексной модели*, в которой условное распределение зависимой переменной зависит от регрессоров x исключительно через индекс $x'\beta$. Модель пропорциональных рисков и модель цензурированной регрессии логически различаются, за исключением того факта, что обе они упрощаются до параметрической вейбулловской модели. Обе модели являются частными случаями модели с преобразованием. Модель цензурированной регрессии является частным случаем регрессионной модели целенаправленного проецирования. Модель с преобразованием можно записать как гетероскедастичную модель целенаправленного проецирования: если $G(T^*) = x'\beta + \varepsilon$, где $G(\cdot)$ – монотонно возрастающее преобразование, то $\ln T^* = H(x'\beta) + \zeta$, где $H(x'\beta) = \mathbb{E}_{\varepsilon}[\ln G^{-1}(x'\beta + \varepsilon)]$, а ζ имеет функцию распределения $F(G(\exp(\zeta + H(x'\beta))) - x'\beta)$, которая в общем случае гетероскедастична.

Статистические вопросы, которые возникают при применении этих моделей, включают свойства распределений оценок (асимптотические и, возможно, в конечных выборках), которые получаются при различных предположениях, и эффективность альтернативных оценок.

Рис. 1: Одноиндексные модели

Правила наблюдения: $T = \min(T^c, T^*)$ для данных, цензурированных справа, $T = \mathrm{sgn}(\ln(T^*))$ для биномиальных моделей дискретного выбора.

(Специфика модели растет по мере продвижения вниз по таблице)



До настоящего времени большая часть исследований сконцентрирована на поиске вычислительно доступных оценок, установлении их состоятельности, асимптотической нормальности и границ эффективности.

Хоровиц и Ньюманн используют две оценки для модели цензурированной регрессии – квантильную оценку (Powell, 1986) и одношаговую полупараметрическую ОМНК-оценку (ПОМНК) (Horowitz, 1986). Другие оценки, предложенные для данной модели, включают гибкие параметрические приближения кумулятивной функции распределения (см., например, Duncan (1986), который рассматривает приближения сплайнами – «метод решета»). Chamberlain (1986) и Cosslett (1987) установили для модели цензурированной регрессии существование положительной границы эффективности для параметрической части. Это означает, что можно использовать достаточно грубые оценки непараметрической части, чтобы достичь $\sqrt{N}$ асимптотически нормальной оценки для параметрической части. Доказано, что оценки из Powell (1986) и Horowitz (1986) являются асимптотически нормальными. Ни одна из них не достигает границы эффективности в случае IID-ошибок, и в общем случае одна не является эффективнее другой.

Оценивание модели пропорциональных рисков с неизвестной функцией базового риска подробно изучено, см. Kaplan & Meier (1968), Cox (1972), Kalbfleisch & Prentice (1982) и Meyer (1990). Особенно полезный «полупараметрический» метод оценивания этой модели, приме-

нимый, когда продолжительность измеряется в «неделях», – гибко параметризовать базовый риск; Meyer (1990) показал, что этот метод является $\sqrt{N}$ асимптотически нормальным.

Оценки (одноиндексной) модели целенаправленного проецирования были предложены в Ichimura (1987), Ruud (1986), Stoker (1986) и Powell, Stock & Stoker (1989). Оценка Ичимуры выбирает β, минимизирующую дисперсию $\ln T$ условно на $x'\beta$, используя ядерную оценку условного среднего для получения оценки условной дисперсии. Эта оценка состоятельна, даже если ошибки разнородны относительно индексной функции, так что ее также можно применять для модели с преобразованием. Оценка Ичимуры является $\sqrt{N}$ асимптотически нормальной, и, как недавно было показано, достигает полупараметрической границы эффективности для гомоскедастичной модели целенаправленного проецирования с нормальными ошибками. Она почти наверняка не является эффективной для модели с преобразованием. Оценки Рууда и Стокера основаны на том факте, что при подходящих условиях регрессия $\ln T$ на x пропорциональна β. Эти оценки также $\sqrt{N}$ асимптотически нормальны.

Оценивание модели с преобразованием, применимое также к модели пропорциональных рисков, реализуется с помощью метода максимальной ранговой корреляции, предложенного в Han (1987) и Doksum (1985).

Newey (1990) установил асимптотическую эффективность некоторых ядерных и квантильных оценок модели цензурированной регрессии, когда ошибки имеют симметричное распределение. Эффективность этих оценок при других условиях не установлена. Проблемой, требующей дальнейших исследований, является построение надежных и практичных оценок дисперсии полупараметрических оценок. Интересный эмпирический вопрос заключается в том, можно ли воспринимать модель цензурированной регрессии или модель пропорциональных рисков как ограничения модели с преобразованием (и каковы подходящие и удобные тестовые статистики).

3 Заявленная готовность платить за природные ресурсы

Методом выявления готовности платить (ГП) за природные ресурсы является экспериментальный опрос населения об их условных оценках: участникам обследования задается вопрос, готовы ли они платить величину b, где b – ставка, установленная правилами эксперимента. Пусть d обозначает фиктивную переменную, равную единице при ответе «да» и нулю в противном случае. Выборка из n наблюдений формируется из пар (b, d), а также регрессоров x, характеризующих респондента. Предположим, что ГП распределена в популяции как $w = x'\beta - \varepsilon$, где ε имеет кумулятивную функцию распределения $G(\varepsilon)$, не зависящую от x. Тогда $\mathbb{P}\{d = 1|x'\beta\} = G(x'\beta - b)$, или

$$d = G(x'\beta - b) + \varepsilon. \tag{32}$$

Предположим, что β и функция G неизвестны. Эконометрическая задача состоит в том, чтобы оценить β и, если необходимо, G и при помощи этих оценок измерить положение распределения ГП, условное на x или безусловное. Это пример регрессионной модели целенаправленного проецирования.

Экспериментальные опросы об условных оценках вызывают споры, поскольку они очень чувствительны к психометрическим контекстным эффектам, таких как якорение, при котором респонденты, не уверенные в своих предпочтениях, воспринимают предлагаемую ставку как сигнал о «политкорректном» диапазоне значений оценки. Также некоторые субъекты, по-видимому, действуют стратегически, намеренно принимая ложно высокую ставку, которую в действительности они не заплатили бы, но которая выражает «протестную» позицию. Эти эффекты делают оценки ГП неточными, а их связь с экономикой благосостояния непрочной.

Почему же в экспериментальных опросах об условных оценках для их выявления применяется формат референдума, а не формат, при котором респондентов просили бы дать свободный ответ о ГП? Одной из причин является то, что открытый формат ведет к гораздо более высокой доле отсутствия ответа, так что метод референдума снижает смещение вследствие самоотбора, вызываемого отсутствием ответов. Другая причина состоит в том, что психологически референдум и открытый формат выявляют весьма различное поведение. Некоторые считают, что формат референдума ближе к механизму выборов, обычно применяемому для принятия общественных решений, и имеется преимущество в подражании этому механизму при принятии общественных решений о природных ресурсах.

Один из вопросов, возникающих при разработке экспериментальных опросов об условных оценках, – выбор уровней ставок b. Альтернативами являются случайный выбор b или выбор b на сетке с определенным размером ячеек. На практике используются грубые сетки, что ограничивает точность полупараметрических оценок. Пусть $h(b|x)$ – плотность распределения, из которого вытягиваются уровни ставок b, условно на x. Оно известно исследователю, поскольку выбирается разработчиками эксперимента.

При эконометрическом анализе данных по референдуму о ГП можно использовать тот факт, что (32) является моделью бинарного выбора и одноиндексной моделью (которая гетероскедастична, но только относительно индекса). Тогда доступными методами для оценивания β являются оценка, основанная на максимуме очков из Manski (1978), полупараметрическая ММП-оценка из Cosslett (1987), оценка из Ichimura (1986), минимизирующая ожидаемую условную дисперсию, оценка из Horowitz (1992), являющаяся гладкой версией оценки, основанной на максимуме очков, и оценка из Klein & Spady (1993). Ключевой результат для модели бинарного выбора состоит в том, что при некоторых условиях гладкости, существуют $\sqrt{N}$-состоятельные оценки β_n для β, т.е. величина $\sqrt{N}(\beta_n - \beta)$ асимптотически нормальна. Непараметрическую оценку G можно получить совместно с оцениванием β, как в процедуре Косслетта, или при помощи обычных ядерных методов на втором шаге, после того как оценка β подставляется для формирования индекса; ее непараметрическая оценка обязательно будет иметь скорость сходимости меньшую, чем $\sqrt{N}$.

Особенно простая оценка параметров индекса β была предложена для этой задачи в Lewbel & McFadden (1997): надо просто оценить с помощью МНК модель

$$\frac{d_i - \mathbb{I}\{b_i < 0\}}{h(b_i|x_i)} = x_i\beta + \zeta_i. \tag{33}$$

Авторы показывают, что оценки коэффициентов в данной регрессии являются состоятельными оценками β и асимптотически нормальны со скоростью сходимости $\sqrt{N}$. Эти оценки не является особо эффективными, но их простота делает их отличной отправной точкой для анализа спецификации модели и построения более эффективных оценок. Авторы также устанавливают, что r-й момент ГП, условно на $x = x_0$, можно $\sqrt{N}$-состоятельно оценить следующим образом:

$$M_r = (x_0\beta)^r + r\sum_{i=1}^{n}(b_i + (x_0 - x_i)\beta)^{r-1} \cdot \frac{d_i - \mathbb{I}\{x_i\beta > b_i\}}{\sum_{j=1}^{n} h(b_i + (x_j - x_i)\beta|x_j)}. \tag{34}$$

Оценки (33) и (34) – хорошие примеры статистических процедур полупараметрического оценивания, которые устойчивы в том смысле, что они не зависят от параметрических предположений о распределении ГП и представляют собой вычислительно удобную альтернативу непараметрическим оценкам ядерного типа.

Литература

Chamberlain, G. (1986). Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics* 29, 189–218.

Cosslett, S. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica* 55, 559–585.

Cox, D. (1972). Regression models and life tables. *Journal of Royal Statistical Society B* 34, 187–220.

Doksum, K. (1985). An extension of partial likelihood methods for proportional hazard models to general transformation models. Working paper, University of California, Berkeley.

Duncan, G. (1986). A semiparametric censored regression estimator. *Journal of Econometrics* 29, 5–34.

Han, A. (1987). Nonparametric analysis of generalized regression models: The maximum rank correlation estimator. *Journal of Econometrics* 35, 303–316.

Heckman, J. & B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271–320.

Horowitz, J. (1986). A distribution-free least squares method for censored linear regression models. *Journal of Econometrics* 29, 59–84.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–531.

Horowitz, J. & G. Newmann (1987). Semiparametric estimation of employment duration models. *Econometric Reviews* 6, 5–40.

Horowitz, J. & G. Newmann (1989). Computational and statistical efficiency of semiparametric GLS estimators. *Econometric Reviews* 8, 223–225.

Ichimura, H. (1986). *Estimation of Single Index Models*. Ph.D. Dissertation, MIT.

Kalbfleisch, J. & R. Prentice (1980). *The Stochastic Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E. & P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association* 53, 487–491.

Klein, R. & R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–422.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica* 47, 141–165.

Lewbel, A. and D. McFadden (1997). Estimating features of a distribution from binomial data. Working paper, University of California, Berkeley.

Manski, C. (1978). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–228.

Meyer, B. (1987). Unemployment insurance and unemployment spells. *Econometrica* 58, 757–782.

Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.

Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics* 29, 143–155.

Powell, J., J. Stock & T. Stoker (1989). Semiparametric estimation of weighted average derivatives. *Econometrica* 57, 1403–1430.

Powell, J. (1994). Estimation of Semiparametric Models. Глава в *Handbook of Econometrics IV* под редакцией R. Engle & D. McFadden. Amsterdam: North-Holland.

Ruud, P. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics* 29, 157–187.

Stoker, T. (1986). Consistent estimation of scaled coefficients. *Econometrica* 54, 1461–1481.

Semiparametric analysis

Daniel McFadden

University of California, Berkeley, USA

This essay surveys two areas of application of semiparametric econometrics: the analysis of censored employment duration data, and the analysis of data on stated willingness-to-pay for natural resources.