Эконометрический ликбез: некоторые вопросы микроэконометрики

Модели выживаемости*

Герман Родригес†

Принстонский Университет, Принстон, США

Настоящее эссе представляет собой введение в модели выживаемости в контексте обобщенных линейных моделей. Вводятся понятия функций риска и выживания, а затем рассматриваются наиболее распространенные механизмы цензурирования и получаемые в результате функции правдоподобия. Обсуждаются основные подходы к моделированию времени ожидания, включая модели ускоренной жизни и пропорциональных рисков, и их расширения для случаев меняющихся во времени регрессоров и зависящих от времени коэффициентов. Затем подробно изучается кусочно-экспоненциальная модель выживаемости и отмечается ее эквивалентность модели пуассоновской регрессии. Далее применение этого подхода рассматривается на примере анализа младенческой и детской смертности в Колумбии по данным опроса. В заключении кратко обсуждаются модели в дискретном времени и их эквивалентность модели логистической регрессии.

1 Введение

В настоящем эссе рассматриваются модели анализа данных, обладающих следующими тремя основными характеристиками: (а) зависимая переменная, или отклик, – это *время ожидания* до наступления определенного события, (б) наблюдения являются *цензурированными* в том смысле, что для некоторых объектов наблюдения исследуемое событие не наступило на момент анализа данных, и (в) имеются предикторы, или *объясняющие переменные*, чье воздействие на время ожидания мы желаем оценить или учесть. Начнем с некоторых базовых определений.

2 Функции риска и выживания

Пусть T – неотрицательная случайная величина, представляющая собой время ожидания до наступления некоторого события. Для простоты будем использовать терминологию анализа выживаемости, называя исследуемое событие «смертью», а время ожидания – временем «выживания», хотя изучаемые далее методы находят гораздо более широкое применение. Их можно использовать, например, для анализа возраста при вступлении в брак, продолжительности брака, интервалов между последовательными родами у женщин, времени пребывания в городе (или на определенном месте работы) и продолжительности жизни. Наблюдательный демограф заметит, что эти примеры включают проблемы рождаемости, смертности и миграции.

*Перевод Б. Гершмана и С. Анатольева. Цитировать как: Родригес, Герман (2008) «Модели выживаемости», Квантиль, №5, стр. 1–27. Citation: Rodríguez, Germán (2008) “Survival models,” Quantile, No.5, pp. 1–27.

†Адрес: Office of Population Research, Princeton University, 241 Wallace Hall, Princeton, NJ 08544, USA. Электронная почта: *grodri@Princeton.edu*

2.1 Функция выживания

Предположим, что T – непрерывная случайная величина с функцией плотности распределения (ФПР) $f(t)$ и кумулятивной функцией распределения (КФР) $F(t) = \mathbb{P}\{T \le t\}$, дающей вероятность того, что событие наступило к моменту времени t.

Часто удобно работать с дополнением КФР, называемым функцией *выживания*:

$$S(t) = \mathbb{P}\{T > t\} = 1 - F(t) = \int_t^\infty f(x)dx, \tag{1}$$

и дающим вероятность быть живым в момент времени t, или в более широком смысле, вероятность того, что исследуемое событие не наступило к моменту времени t.

2.2 Функция риска

Альтернативным способом охарактеризовать распределение величины T является *функция риска*, или мгновенная интенсивность осуществления события, определяемая как

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}\{t < T \le t + dt | T > t\}}{dt}. \tag{2}$$

Числитель этого выражения – условная вероятность того, что событие произойдет в интервале $(t, t + dt)$, если оно не произошло ранее, а знаменатель – ширина интервала. Разделив одно на другое, получаем интенсивность осуществления события в единицу времени. Устремляя ширину интервала к нулю и переходя к пределу, получаем мгновенную интенсивность осуществления события.

Условную вероятность в числителе можно записать в виде отношения совместной вероятности того, что T принадлежит интервалу $(t, t + dt)$ *и* $T > t$ (что, конечно, совпадает с вероятностью того, что T принадлежит указанному интервалу), к вероятности условия $T > t$. Первая из них равна $f(t)dt$ для малого dt, а последняя – это $S(t)$, по определению. Деление на dt и предельный переход дают следующий полезный результат:

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3}$$

который некоторые авторы приводят в качестве определения функции риска. Содержательно, интенсивность осуществления события в момент времени t равна плотности событий в момент t, деленной на вероятность дожить до этого момента, не испытав событие ранее.

Заметим из уравнения (1), что $-f(t)$ – это производная $S(t)$. Тогда уравнение (3) можно переписать в виде

$$\lambda(t) = -\frac{d}{dt} \log S(t).$$

Если теперь проинтегрировать обе части от 0 до t и ввести граничное условие $S(0) = 1$ (поскольку событие не может произойти к моменту времени 0), можно преобразовать приведенное выражение и получить формулу для вероятности дожить до момента времени t как функции от рисков во все моменты времени до t:

$$S(t) = \exp\left\{ -\int_0^t \lambda(x)dx \right\}. \tag{4}$$

Это выражение должно быть знакомо демографам. Интеграл в фигурных скобках в этом уравнении называют *кумулятивным риском* и обозначают как

$$\Lambda(t) = \int_0^t \lambda(x)dx. \tag{5}$$

Можно рассматривать $\Lambda(t)$ как сумму всех рисков при переходе от момента времени 0 к t.

Приведенные результаты показывают, что функции выживания и риска дают альтернативные, но эквивалентные описания распределения величины T. Имея функцию выживания, всегда можно ее продифференцировать и получить функцию плотности, а затем найти функцию риска, используя уравнение (3). Имея функцию риска, всегда можно ее проинтегрировать и получить кумулятивный риск, а затем взять от нее экспоненту и найти функцию выживания, используя уравнение (4). Для закрепления введенных понятий рассмотрим пример.

Пример: Простейшее распределение времени жизни получается, если предположить постоянный риск, то есть

$$\lambda(t) = \lambda$$

для всех t. Соответствующая функция выживания имеет вид

$$S(t) = \exp\{-\lambda t\}.$$

Это экспоненциальное распределение с параметром λ. Функцию плотности можно получить, умножив функцию выживания на риск:

$$f(t) = \lambda \exp\{-\lambda t\}.$$

Математическое ожидание равно $1/\lambda$. Это распределение играют центральную роль в анализе выживаемости, хотя, возможно, оно является слишком простым, чтобы быть полезным в приложениях само по себе.

2.3 Ожидаемая продолжительность жизни

Пусть μ обозначает математическое ожидание T. По определению, значение μ можно подсчитать, умножив t на функцию плотности $f(t)$ и взяв интеграл, то есть

$$\mu = \int_0^\infty t f(t) dt.$$

Интегрируя по частям и используя тот факт, что $-f(t)$ – это производная $S(t)$, удовлетворяющая граничным условиям $S(0) = 1$ и $S(\infty) = 0$, можно показать, что

$$\mu = \int_0^\infty S(t) dt. \tag{6}$$

Иными словами, ожидаемая продолжительность жизни – это просто интеграл от функции выживания.

2.4 Замечание о несобственных случайных величинах

До сих пор неявно предполагалось, что исследуемое событие обязательно происходит, то есть $S(\infty) = 0$. Иначе говоря, по прошествии достаточного времени доля выживших снижается к нулю. Из этого условия следует, что кумулятивный риск должен расходиться, то есть $\Lambda(\infty) = \infty$. Интуитивно, событие точно произойдет только в том случае, если кумулятивный риск за долгий период времени достаточно высок.

Есть, тем не менее, множество возможных событий, которые необязательно происходят. Некоторые мужчины и женщины остаются одинокими всю жизнь, вторые и последующие роды могут не произойти, а некоторые люди достаточно счастливы на своем месте работе, чтобы никогда его не покидать. Что делать в таких случаях? Существуют два подхода.

Один подход – заметить, что все равно можно найти функции риска и выживания, которые корректно определены, даже если исследуемое событие может и не произойти. Например,

можно изучать брачность всего населения, включая людей, которые никогда не вступят в брак, и подсчитать доли состоящих в браке и одиноких. В этом примере $S(t)$ будет отражать долю людей, не состоящих в браке в возрасте t, а $S(\infty)$ – долю тех, кто никогда не вступит в брак.

Одно из ограничений данного подхода в том, что, если событие не обязательно должно произойти, время ожидания T может быть неопределенным (или бесконечным) и в таком случае не является собственной случайной величиной. Ее функция плотности, которую можно вычислить по функциям риска и выживания, будет несобственной, то есть не будет интегрироваться к единице. Очевидно, среднее время ожидания не будет определено. В терминах нашего примера, нельзя подсчитать средний возраст вступления в брак для всего населения, просто поскольку не все люди вступают в брак. Но это ограничение не имеет серьезных последствий, если внимание сосредоточено на функциях риска и выживания, а не на времени ожидания. В примере с браком возможно даже вычислить медианный возраст вступления в брак, если определить его как возраст, к которому половина всего населения вступает в брак.

Альтернативный подход – проводить анализ условно на осуществлении события. В терминах нашего примера, можно было бы изучать брачность (возможно, в ретроспективе) для людей, которые в конечном счете вступают в брак, поскольку для этой группы людей время ожидания T всегда корректно определено. В таком случае можно подсчитать не только условные функции риска и выживания, но и среднее время ожидания. В нашем примере можно вычислить средний возраст при вступлении в брак для тех, кто на самом деле вступает в брак. Можно даже подсчитать обычную медиану, определенную как возраст, к которому вступила в брак половина той части населения, которая в конечном счете вступает в брак.

Оказывается, что условные функции плотности, риска и выживания для тех, кто испытывает событие, связаны с соответствующими безусловными функциями для всего населения. Условная функция плотности имеет вид

$$f^*(t) = \frac{f(t)}{1 - S(\infty)}$$

и интегрируется к единице. Условная функция выживания имеет вид

$$S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)},$$

и стремится к нулю при $t \to \infty$. Поделив условную функцию плотности на условную функцию выживания, получаем условную функцию риска:

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{f(t)}{S(t) - S(\infty)}.$$

В качестве упражнения читатель может подсчитать среднее время ожидания для тех, с кем событие случается.

Какой бы подход ни применялся, следует аккуратно указать, какие именно функции риска и выживания используются. Например, условный риск для тех, кто в конечном счете испытывает событие, всегда выше, чем безусловный риск для всего населения. Заметим также, что в большинстве случаев все, что наблюдается, – это произошло событие или нет. Если событие не произошло, может быть, невозможно определить, произойдет ли оно в будущем. В таком случае по данным можно оценить лишь безусловный риск, но этот результат при желании всегда можно выразить в условных величинах, используя приведенные выше выражения.

3 Цензурирование и функция правдоподобия

Вторая отличительная черта анализа выживаемости – цензурирование, то есть тот факт, что для некоторых объектов наблюдения исследуемое событие произошло, а значит, известно точное время ожидания, тогда как для других это событие не произошло, и все, что известно, – это то, что время ожидания превышает время наблюдения.

3.1 Механизмы цензурирования

Существует несколько механизмов, способных генерировать цензурированные данные. При цензурировании *типа I* выборка из n объектов наблюдается в течении фиксированного времени τ. Число объектов, испытывающих событие, или число «смертей», случайно, но общая продолжительность исследования фиксирована. Тот факт, что продолжительность фиксирована, может быть важным практическим преимуществом при разработке последующего дополнительного исследования.

При простом обобщении этой схемы, называемом *фиксированным цензурированием*, каждый объект имеет максимально возможный период наблюдения τ_i, $i = 1, \ldots, n$, который может варьироваться от одного объекта к другому, однако фиксирован заранее. Вероятность того, что объект i будет жив в конце своего периода наблюдения, равна $S(\tau_i)$, а общее число смертей вновь является случайным.

При цензурировании *типа II* выборка из n объектов наблюдается так долго, сколько необходимо, чтобы d объектов испытали событие. В этой схеме число смертей d, которое определяет точность исследования, фиксировано заранее и его можно использовать в качестве параметра. К сожалению, в этом случае общая продолжительность исследования случайна и не может быть точно известна заранее.

При более общей схеме, называемой *случайным цензурированием*, каждый объект имеет потенциальный момент цензурирования C_i и потенциальную продолжительность жизни T_i, которые предполагаются независимыми случайными величинами. Наблюдается $Y_i = \min\{C_i, T_i\}$, то есть минимум из времени цензурирования и времени жизни, и переменная-индикатор, часто обозначаемая d_i или δ_i, которая указывает, закончено наблюдение в результате смерти или цензурирования.

Все эти схемы объединяет тот факт, что механизм цензурирования *неинформативен*, и все они, в сущности, ведут к той же самой функции правдоподобия. Наиболее слабое предположение, требуемое для получения этой функции правдоподобия, состоит в том, что цензурирование наблюдения не должно давать какой-либо информации относительно перспектив выживания этого конкретного объекта за пределами момента цензурирования. На самом деле базовое предположение, которому мы будем следовать, таково: все, что известно о наблюдении, цензурированном в момент времени t – это то, что время жизни для него превышает t.

3.2 Функция правдоподобия для цензурированных данных

Предположим, что имеются n объектов наблюдения со временем жизни, характеризуемым функцией выживания $S(t)$ с соответствующей плотностью $f(t)$ и риском $\lambda(t)$. Предположим также, что объект i наблюдается в течение времени t_i. Если объект умер в момент t_i, его вклад в функцию правдоподобия – значение плотности в этот момент времени, которую можно записать как произведение функций выживания и риска:

$$L_i = f(t_i) = S(t_i)\lambda(t_i).$$

Если объект все еще жив в момент времени t_i, все, что известно при неинформативном цензурировании – это то, что время его жизни превышает t_i. Вероятность этого события

равна

$$L_i = S(t_i),$$

и отражает вклад цензурированного наблюдения в функцию правдоподобия.

Заметим, что оба варианта вкладов содержат функцию выживания $S(t_i)$, поскольку в обоих случаях объект дожил до момента времени t_i. Смерть «домножает» этот вклад на риск $\lambda(t_i)$, а цензурирование – нет. Можно записать оба типа вкладов в виде единого выражения. Для этого пусть d_i является индикатором смерти и равняется единице, если объект i умер, и нулю в противном случае. Тогда функцию правдоподобия можно записать в следующем виде:

$$L = \prod_{i=1}^{n} L_i = \prod_i \lambda(t_i)^{d_i} S(t_i).$$

Логарифмируя и используя выражение, связывающее функцию выживания $S(t)$ и функцию кумулятивного риска $\Lambda(t)$, получаем логарифмическую функцию правдоподобия для цензурированных данных о выживаемости:

$$\log L = \sum_{i=1}^{n} \{d_i \log \lambda(t_i) - \Lambda(t_i)\}. \tag{7}$$

Для закрепления материала рассмотрим пример.

Пример: Предположим, имеется выборка размера n цензурированных наблюдений из экспоненциального распределения. Пусть t_i – период наблюдения, а d_i – индикатор смерти для объекта i.

В случае экспоненциального распределения $\lambda(t) = \lambda$ для всех t. Кумулятивный риск, таким образом, является интегралом от константы, а значит, $\Lambda(t) = \lambda t$. Подставляя два этих результата в уравнение (7), получаем логарифмическую функцию правдоподобия

$$\log L = \sum_{i=1}^{n} \{d_i \log \lambda - \lambda t_i\}.$$

Пусть $D = \sum_i d_i$ обозначает общее число смертей, а $T = \sum_i t_i$ – общее время наблюдения (или подверженность риску). Тогда можно переписать логарифмическую функцию правдоподобия как функцию от этих совокупных величин:

$$\log L = D \log \lambda - \lambda T. \tag{8}$$

Дифференцируя это выражение относительно λ, получаем скор-функцию

$$u(\lambda) = \frac{D}{\lambda} - T,$$

и, приравнивая ее к нулю, находим оценку максимального правдоподобия для риска

$$\hat{\lambda} = \frac{D}{T}, \tag{9}$$

которая равна общему числу смертей, деленному на общую подверженность риску. Демографы узнают в этом выражении общее определение коэффициента смертности. Заметим, что эта оценка является оптимальной (в смысле максимального правдоподобия), только если риск является постоянным и не зависит от возраста.

Можно также подсчитать количество информации в выборке, взяв со знаком минус вторую производную скор-функции:

$$I(\lambda) = \frac{D}{\lambda^2}.$$

Для получения ожидаемой информации необходимо найти ожидаемое число смертей, но оно зависит от схемы цензурирования. Например, при цензурировании типа I с фиксированной продолжительностью τ, ожидаемое число смертей равно $n(1 - S(\tau))$. При цензурировании типа II число смертей фиксируется заранее. При некоторых схемах подсчет данного математического ожидания может быть довольно затруднителен или даже невозможен.

Простой альтернативный вариант – использовать наблюдаемую информацию, оцененную с помощью ММП-оценки λ из уравнения (9). Используя этот подход, асимптотическую дисперсию ММП-оценки риска можно оценить как

$$\hat{\mathbb{V}}(\hat{\lambda}) = \frac{D}{T^2},$$

и затем применять эту оценку для асимптотического тестирования гипотез и построения доверительных интервалов для λ.

В отсутствии цензурированных наблюдений, то есть при $d_i = 1$ для всех i и $D = n$, полученные выше результаты становятся стандартным ММП-оцениванием для экспоненциального распределения, а ММП-оценка параметра λ равна величине, обратной выборочному среднему.

Интересно заметить, между прочим, что логарифмическая функция правдоподобия для цензурированных данных из экспоненциального распределения, приведенная в уравнении (8), в точности совпадает (не считая констант) с логарифмической функцией правдоподобия, возникающей, если считать D пуассоновской случайной величиной со средним значением λT. Чтобы убедиться в этом, следует записать пуассоновскую логарифмическую функцию правдоподобия, когда $D \sim P(\lambda T)$, и заметить, что она отличается от уравнения (8) только наличием члена $D \log(T)$, являющегося константой, зависящей только от данных, но не от параметра λ.

Таким образом, если считать смерть пуассоновской величиной, условно на время подверженности риску, получаются точно такие же оценки (и стандартные ошибки), как в случае, когда периоды подверженности риску рассматриваются как цензурированные наблюдения из экспоненциального распределения. Этот результат разрабатывается ниже, чтобы связать модели выживаемости и обобщенные линейные модели с пуассоновской структурой ошибок.

4 Подходы к моделированию выживаемости

До сих пор речь шла об однородной популяции, в которой продолжительность жизни каждого объекта характеризовалась одной и той же функцией выживания $S(t)$. Рассмотрим теперь третью отличительную черту моделей выживаемости – наличие вектора регрессоров, или объясняющих переменных, которые могут воздействовать на время жизни, – и обратимся к общей задаче моделирования этих эффектов.

4.1 Модели ускоренной жизни

Пусть T_i – случайная величина, представляющая собой (возможно, ненаблюдаемое) время жизни i-го объекта. Поскольку величина T_i должна быть неотрицательной, можно рассмотреть модель для ее логарифма, скажем, обычную линейную модель

$$\log T_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

где ϵ_i – надлежащий остаточный член, распределение которого будет специфицировано далее. Эта модель задает распределение логарифма времени жизни для i-го объекта как простой *сдвиг* стандартного, или базового, распределения, представленного остаточным членом.

Взяв экспоненту от этого уравнения, получаем модель собственно для времени жизни:

$$T_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}T_{0i},$$

где T_{0i} – экспонента от остаточного члена. Удобно также использовать обозначение γ_i для мультипликативного эффекта регрессоров, $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

Интерпретация параметров стандартна. Рассмотрим, например, модель с константой и фиктивной переменной x, отражающей бинарный фактор, скажем, принадлежность к группам 1 или 0. Предположим, соответствующий мультипликативный эффект $\gamma = 2$, так что коэффициент при x – это $\beta = \log(2) = 0{,}6931$. Тогда вывод состоит в том, что люди из первой группы живут вдвое дольше, чем из нулевой.

Существует интересная альтернативная интерпретация, которая объясняет название «модель ускоренной жизни». Пусть $S_0(t)$ обозначает функцию выживания для группы 0, которая будет контрольной группой, а $S_1(t)$ – для группы 1. Для этой модели

$$S_1(t) = S_0(t/\gamma).$$

Иными словами, вероятность того, что индивид из первой группы доживет до возраста t, в точности равна вероятности того, что индивид из нулевой группы доживет до возраста t/γ. Для $\gamma = 2$ получим половину возраста, так что вероятность того, что индивид из первой группы доживет до 40 (или 60) лет будет равна вероятности того, что индивид из нулевой группы доживет до 20 (или 30) лет. Таким образом, можно рассматривать γ как параметр, воздействующий на протекание времени. В нашем примере люди в нулевой группе стареют «в два раза быстрее».

Заметим, что соответствующие функции риска связаны соотношением

$$\lambda_1(t) = \lambda_0(t/\gamma)/\gamma,$$

так что при $\gamma = 2$ в каждом данном возрасте люди из первой группы будут подвержены вдвое меньшему риску, чем вдвое младшие люди из нулевой группы.

Название «модель ускоренной жизни» происходит из промышленных приложений, когда предметы тестируются при гораздо худших условиях, чем встречающиеся в реальной жизни, чтобы тесты можно было выполнить за более короткое время.

Различные предположения о распределении остаточного члена приводят к различным видам параметрических моделей. Если ошибка ϵ_i нормально распределена, получается логнормальная модель для T_i. Оценивание этой модели для цензурированных данных по методу максимального правдоподобия известно в эконометрической литературе как тобит-модель.

Если же ϵ_i имеет распределение экстремального значения с функцией плотности

$$f(\epsilon) = \exp\{\epsilon - \exp(\epsilon)\},$$

то T_{0i} имеет экспоненциальное распределение, и получается модель экспоненциальной регрессии, где T_i экспоненциально распределено с риском λ_i, удовлетворяющим логлинейной модели

$$\log \lambda_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

Примером демографической модели, принадлежащей классу моделей ускоренной жизни, является модель Кола–Макнейла для частоты первого брака, в которой доля индивидов, когда-либо состоявших в браке к возрасту a в данной популяции, записывается в виде

$$F(a) = cF_0\left(\frac{a - a_0}{k}\right),$$

где F_0 – модельное распределение долей женщин, состоявших в браке к определенному возрасту среди когда-либо состоявших в браке, на основе исторических данных по Швеции; c – доля тех, кто в конечном счете вступают в брак, a_0 – возраст вступления в брак, а k – *скорость* протекания брака относительно шведского стандарта.

Модели ускоренной жизни по сути являются обычными моделями регрессии, примененными к логарифму времени жизни, и, не считая факта цензурирования данных, не представляют новых трудностей при оценивании. Как только выбрано распределение остаточного члена, оценивание осуществляется путем максимизации логарифмической функции правдоподобия для цензурированных данных, рассмотренной в предыдущем разделе. Детали можно найти в работе Kalbfleish & Prentice (1980).

4.2 Модели пропорциональных рисков

Большой класс моделей, впервые предложенный в Cox (1972), концентрируется непосредственно на функции риска. Простейший представитель этого класса – модель *пропорциональных рисков*, в которой риск в момент t для индивида с характеристиками $\mathbf{x}_i$ (не включая константу) имеет вид

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{10}$$

В этой модели $\lambda_0(t)$ – это базовая функция риска, которая измеряет риск для индивидов с $\mathbf{x}_i = \mathbf{0}$, служащих точкой отсчета, а $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ – относительный риск, то есть пропорциональное увеличение или уменьшение риска, связанное с набором характеристик $\mathbf{x}_i$. Заметим, что увеличение или снижение риска одинаково для всех моментов времени t.

В качестве иллюстрации рассмотрим пример с двумя выборками, когда имеется фиктивная переменная x, означающая принадлежность к первой или нулевой группе. В этом случае модель принимает вид

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{если } x = 0, \\ \lambda_0(t)e^{\beta} & \text{если } x = 1. \end{cases}$$

Таким образом, $\lambda_0(t)$ представляет собой риск в момент t в нулевой группе, а $\gamma = \exp\{\beta\}$ – это отношение риска в первой группе к риску в нулевой группе в любой момент времени t. Если $\gamma = 1$ (или $\beta = 0$), риски одинаковы в обеих группах. Если $\gamma = 2$ (или $\beta = 0{,}6931$), риск для индивида из первой группы в каждый момент времени вдвое больше риска для индивида того же возраста из нулевой группы.

Заметим, что модель явным образом отделяет эффект времени от эффекта регрессоров. Логарифмируя, легко увидеть, что модель пропорциональных рисков – это простая аддитивная модель для логарифма риска:

$$\log\lambda_i(t|\mathbf{x}_i) = \alpha_0(t) + \mathbf{x}_i'\boldsymbol{\beta},$$

где $\alpha_0(t) = \log\lambda_0(t)$ – логарифм базового риска. Как во всех аддитивных моделях, предполагается что влияние регрессоров $\mathbf{x}$ одинаково для всех моментов времени, или возрастов, t. Нельзя не отметить схожесть между этим выражением и стандартной моделью ковариационного анализа с параллельными прямыми.

Возвращаясь к уравнению (10), можно проинтегрировать обе его части от 0 до t и получить кумулятивные риски

$$\Lambda_i(t|\mathbf{x}_i) = \Lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

которые также пропорциональны. Взяв экспоненту от этого уравнения с противоположным знаком, получаем функции выживания

$$S_i(t|\mathbf{x}_i) = S_0(t)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}, \tag{11}$$

где $S_0(t) = \exp\{-\Lambda_0(t)\}$ – базовая функция выживания. Таким образом, эффект регрессоров $\mathbf{x}_i$ на функцию выживания заключается в возведении ее в степень, равную относительному риску $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$.

В нашем примере с двумя группами и относительным риском $\gamma = 2$ вероятность того, что индивид из первой группы доживет до возраста t, равен квадрату вероятности того, что индивид из нулевой группы доживет до этого же возраста.

4.3 Экспоненциальная и вейбулловская модели

Различные виды моделей пропорциональных рисков можно получить, делая различные предположения о базовой функции выживания, или, что эквивалентно, о базовой функции риска. Например, если базовый риск постоянен во времени, то есть $\lambda_0(t) = \lambda_0$, получаем модель экспоненциальной регрессии, где

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0 \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Интересно, что модель экспоненциальной регрессии принадлежит одновременно к классам моделей пропорциональных рисков и моделей ускоренной жизни. Если базовый риск постоянен, и происходит удвоение или утроение риска, новый риск также будет постоянным (просто более высоким). Возможно, менее очевидным является тот факт, что в случае постоянного базового риска при ускорении течения времени в два или три раза новый риск удваивается или утраивается, но тем не менее остается постоянным во времени, то есть мы остаемся в рамках экспоненциального класса.

Можно задаться вопросом, есть ли другие случаи, при которых две модели совпадают. Есть, но немного. На самом деле, есть всего лишь одно распределение, при котором модели совпадают, и оно включает экспоненциальное как частный случай. Единственный случай, когда два класса совпадают, имеет место при распределении *Вейбулла*, для которого функция выживания принимает вид

$$S(t) = \exp\{-(\lambda t)^p\},$$

а функция риска – вид

$$\lambda(t) = p\lambda(\lambda t)^{p-1},$$

для некоторых параметров $\lambda > 0$ и $p > 0$. При $p = 1$ данная модель сводится к экспоненциальной и имеет постоянный во времени риск. При $p > 1$ риск увеличивается со временем. При $p < 1$ риск уменьшается со временем. Действительно, логарифмируя выражение для функции риска, получаем, что логарифм вейбулловского риска – линейная функция логарифма времени с коэффициентом наклона $p - 1$.

Если взять вейбулловский риск в качестве базового, а затем умножить его на константу γ в рамках модели пропорциональных рисков, снова получим распределение Вейбулла, так что данный класс замкнут относительно пропорциональности рисков. Если взять вейбулловскую функцию выживания в качестве базовой и затем ускорить протекание времени в рамках модели ускоренной жизни, разделив время на константу γ, снова получим распределение Вейбулла, так что этот класс замкнут относительно ускорения времени.

Другие подробности об этом распределении можно найти в Cox & Oakes (1984) или Kalbfleish & Prentice (1980), где доказана эквивалентность двух вейбулловских моделей.

4.4 Меняющиеся во времени регрессоры

До сих пор в явном виде рассматривались только регрессоры, неизменные во времени. Однако локальная природа модели пропорциональных рисков позволяет легко обобщить ее на случай меняющихся во времени регрессоров. Рассмотрим несколько примеров.

Предположим, производится анализ временных интервалов между родами, и изучаются промежутки между последовательными родами. Один из возможных регрессоров – уровень образования матери, который в большинстве случаев можно считать неизменным во времени.

Предположим теперь, что мы хотим добавить в качестве регрессора индикатор кормления грудью того ребенка, с которого начинается отсчет интервала. Предполагая, что ребенка в принципе кормят грудью, эта переменная будет принимать значение единица («да»), начиная

с родов и до окончания грудного кормления, когда переменная сменит значение на ноль («нет»). Это простой пример ситуации, когда регрессор может менять значение лишь один раз.

Более сложный анализ возникает при включении показателя частоты грудного кормления за сутки. Эта переменная может менять значение ежедневно. К примеру, последовательность значений для одной женщины могла бы иметь вид 4, 6, 5, 6, 5, 4, ...

Пусть $\mathbf{x}_i(t)$ обозначает значение вектора регрессоров для индивида i в момент времени t. Тогда модель пропорциональных рисков можно обобщить следующим образом:

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\}. \tag{12}$$

Теперь нет четкого разделения эффектов времени и регрессоров, и иногда может быть сложно идентифицировать эффекты регрессоров, которые сильно коррелируют со временем. Если, например, все дети были отняты от груди в возрасте около 6 месяцев, будет сложно отделить эффект кормления грудью от общих временных эффектов, не имея дополнительной информации. И все же в таких случаях может быть предпочтительным использование меняющегося во времени регрессора как более разумного предиктора риска, чем просто количество прошедшего времени.

Подсчет функций выживания при наличии меняющихся во времени регрессоров немного более затруднителен, поскольку требуется специфицировать траекторию для каждой переменной. В примере с промежутками между родами можно подсчитать функцию выживания для женщин, которые кормят грудью в течение шести месяцев, а затем прекращают. Это реализуется путем использования функции риска, соответствующей $x(t) = 0$ для месяцев от 0 до 6, а затем риска, соответствующего $x(t) = 1$, для месяцев, начиная с шестого. К сожалению, теряется простота уравнения (11): больше нельзя просто возвести в степень базовую функцию выживания.

Меняющиеся во времени регрессоры можно также ввести в контексте моделей ускоренной жизни, но это не так просто и редко осуществляется в практических приложениях. См. подробности в Cox & Oakes (1984, стр. 66).

4.5 Зависящие от времени коэффициенты

Модель также можно обобщить на случай *коэффициентов*, которые меняются во времени и, таким образом, не являются пропорциональными. Вполне возможно, например, что определенные социальные характеристики могут иметь сильное воздействие на риск смертности для детей сразу после рождения, но иметь относительно низкое воздействие позднее. Чтобы учесть модели подобного вида, можно записать

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}(t)\},$$

где параметр $\boldsymbol{\beta}(t)$ теперь является функцией от времени.

Эта модель обладает очень высокой степенью общности. Например, в случае с двумя выборками модель можно записать в виде

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t) & \text{если } x = 0, \\ \lambda_0(t)e^{\beta}(t) & \text{если } x = 1, \end{cases}$$

что, в сущности, позволяет иметь две произвольные функции риска, по одной для каждой группы. Таким образом, это весьма насыщенная модель.

Обычно форму зависимости коэффициентов от времени необходимо специфицировать параметрически, чтобы было возможно идентифицировать модель и оценить параметры. Очевидными кандидатами являются полиномы от времени, когда $\beta(t)$ – линейная или квадратичная функция от времени. Cox & Oakes (1984, стр. 76) показывают, как можно использовать быстро затухающие экспоненты для моделирования переменных коэффициентов.

Заметим, что вновь потеряно простое разделение эффектов регрессоров и времени. Подсчет функции выживания в этой модели снова несколько осложняется тем фактом, что коэффициенты теперь зависят от времени, так что они не выносятся за знак интеграла. Простое уравнение (11) не имеет место.

4.6 Общая модель риска

Изложенные расширения модели на случай меняющихся во времени регрессоров и зависящих от времени коэффициентов можно объединить в наиболее общую версию модели риска:

$$\lambda_i(t, \mathbf{x}_i(t)) = \lambda_0(t) \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}(t)\},$$

где $\mathbf{x}_i(t)$ – вектор меняющихся во времени регрессоров, представляющих собой характеристики индивида i в момент времени t, а $\boldsymbol{\beta}(t)$ – вектор зависящих от времени коэффициентов, представляющих собой эффект, который эти характеристики оказывают в момент времени t.

Пример с грудным кормлением и его влиянием на длину промежутков между последовательными родами является хорошей иллюстрацией, отражающей оба эффекта. Статус грудного кормления сам по себе является меняющимся во времени регрессором $x(t)$, принимающим значение единица, если женщина кормит грудью ребенка через t месяцев после родов. Известно, что воздействие, которое грудное кормление может иметь на задержку овуляции, а следовательно, на снижение риска беременности, быстро падает со временем, так что, возможно, его следует моделировать как зависящий от времени эффект $\beta(t)$. Опять же, дальнейшие действия требуют спецификации функциональной формы зависимости от времени.

4.7 Подгонка модели

Существует, в сущности, три подхода к подгонке моделей выживаемости:

- Первый и, наверное, наиболее простой – это *параметрический* подход, когда предполагается определенная функциональная форма для базового риска $\lambda_0(t)$. Примерами являются модели, основанные на экспоненциальном, вейбулловском, гамма и обобщенном F распределениях.

- Второй подход можно назвать гибким или *полупараметрическим*, когда делаются довольно слабые предположения о базовом риске $\lambda_0(t)$. В частности, можно разбить время на достаточно малые интервалы и предположить, что базовый риск постоянен внутри каждого интервала, что приведет к кусочно-экспоненциальной модели.

- Третий подход – *непараметрический*, при котором регрессионные коэффициенты $\boldsymbol{\beta}$ оцениваются без какой-либо спецификации функции базового риска $\lambda_0(t)$. Этот подход основан на функции частного правдоподобия, предложенной в статье Cox (1972).

Подробное обсуждение этих подходов выходит далеко за рамки настоящего эссе. Остановимся подробнее на промежуточном, полупараметрическом подходе, поскольку (а) он достаточно гибкий, чтобы быть полезным инструментом с широким спектром применения, и (б) он тесно связан с моделью пуассоновской регрессии.

5 Кусочно-экспоненциальная модель

Рассмотрим подгонку модели пропорциональных рисков в обычном виде

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{13}$$

при достаточно слабых предположениях о базовом риске $\lambda_0(t)$.

5.1 Кусочно-постоянный риск

Рассмотрим разбиение времени на J интервалов с точками разбиения $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$. Определим j-й интервал $[\tau_{j-1}, \tau_j)$, продолжающийся от $(j-1)$-й границы до j-й, включая начало и не включая конец.

Предположим далее, что базовый риск *постоянен* внутри каждого интервала, так что

$$\lambda_0(t) = \lambda_j \quad \text{для } t \text{ из } \quad [\tau_{j-1}, \tau_j). \tag{14}$$

Таким образом, базовый риск $\lambda_0(t)$ моделируется, используя J параметров $\lambda_1, \ldots, \lambda_J$, каждый из которых представляет риск для эталонной группы (или индивида) в одном конкретном интервале. Поскольку риск предполагается кусочно-постоянным, соответствующую функцию выживания часто называют кусочно-экспоненциальной.

Конечно, разумный выбор точек разбиения должен позволить достаточно хорошо аппроксимировать почти любой базовый риск, если использовать близко расположенные границы интервалов, когда риск меняется быстро, и более широкие интервалы, когда риск меняется медленнее.

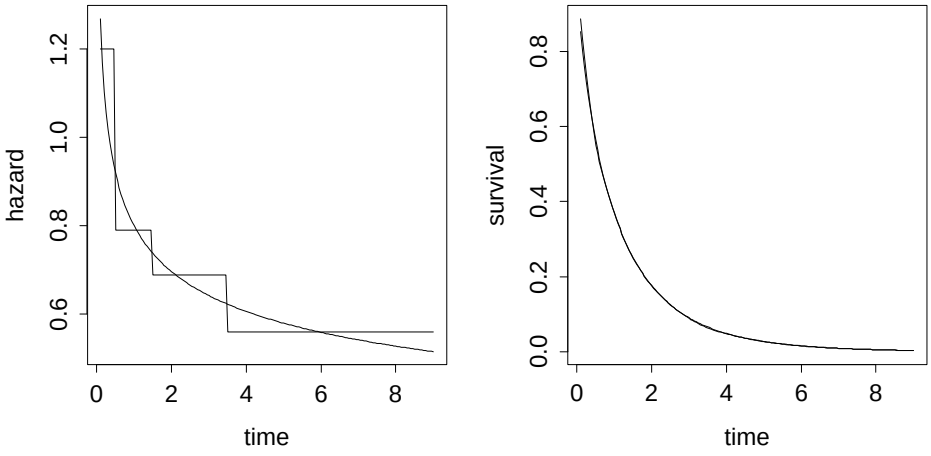


Рис. 1: Аппроксимация кривой выживания с помощью кусочно-постоянной функции риска

На Рис. 1 показано, как распределение Вейбулла с параметрами $\lambda = 1$ и $p = 0{,}8$ можно аппроксимировать с помощью кусочно-экспоненциального распределения с точками разбиения 0,5, 1,5 и 3,5. Диаграмма слева показывает, что кусочно-постоянный риск способен лишь в самых общих чертах повторить гладко убывающую вейбулловскую функцию риска, но при этом, как показано на диаграмме справа, соответствующие кривые выживания неразличимы.

5.2 Модель пропорциональных рисков

Теперь добавим регрессоры в контексте модели пропорциональных рисков из уравнения (13), предполагая, что базовый риск является кусочно-постоянным, как в уравнении (14). Запишем модель в виде

$$\lambda_{ij} = \lambda_j \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}, \tag{15}$$

где λ_{ij} – риск, соответствующий индивиду i в интервале j, λ_j – базовый риск в интервале j, а $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ – относительный риск для индивида со значениями регрессоров $\mathbf{x}_i$ по сравнению с базовым в каждый заданный момент времени.

Логарифмируя, получаем аддитивную логлинейную модель

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{16}$$

где $\alpha_j = \log \lambda_j$ – логарифм базового риска. Это стандартная логлинейная модель, в которой временные интервалы влияют на риск. Поскольку константа в явном виде не включена в модель, не требуется накладывать ограничения на α_j. Если бы мы захотели включить константу, отражающую риск в первом интервале, мы положили бы $\alpha_1 = 0$, как обычно.

Модель можно расширить, включив меняющиеся во времени регрессоры и зависящие от времени коэффициенты, но обсуждение деталей мы отложим до изучения оценивания более простой модели пропорциональных рисков.

5.3 Эквивалентная пуассоновская модель

Holford (1980) и Laird & Oliver (1981) в независимо написанных статьях, опубликованных почти одновременно, заметили, что кусочная модель пропорциональных рисков из предыдущего раздела эквивалентна определенной модели пуассоновской регрессии. Сначала сформулируем результат, а затем набросаем его доказательство.

Вспомним, что наблюдаются t_i – общее время, прожитое i-м индивидом, и d_i – индикатор смерти, который принимает значение единица, если индивид умер, и ноль в противном случае. Введем аналогичные показатели для каждого интервала времени, проживаемого индивидом i. Можно рассматривать этот процесс как создание ряда псевдонаблюдений, по одному для каждой комбинации индивида и интервала.

Сначала создадим показатели подверженности риску. Пусть t_{ij} обозначает время, прожитое i-м индивидом в j-м интервале, то есть между τ_{j-1} и τ_j. Если индивид прожил до конца интервала, то есть $t_i > \tau_j$, то время, прожитое в интервале, равно ширине интервала и $t_{ij} = \tau_j - \tau_{j-1}$. Если индивид умер или был цензурирован в этом интервале, то есть $\tau_{j-1} < t_i < \tau_j$, то время, прожитое в интервале, равно $t_{ij} = t_i - \tau_{j-1}$, разнице между общим прожитым временем и нижней границей интервала. Рассматриваются только те интервалы, в течение которых индивид жил, но очевидно, что время жизни в интервале равно нулю, если индивид умер до его начала и $t_i < \tau_{j-1}$.

Теперь создадим индикаторы смерти. Пусть d_{ij} принимает значение единица, если индивид i умер в интервале j, и ноль в противном случае. Пусть $j(i)$ обозначает интервал, куда попадает t_i, то есть интервал, когда индивид i умер или был цензурирован. Мы используем функциональное обозначение, чтобы подчеркнуть, что этот интервал меняется от одного индивида к другому. Если t_i попадает, скажем, в интервал $j(i)$, то d_{ij} должно равняться нулю для всех $j < j(i)$ (то есть для всех более ранних интервалов) и будет равно d_i для $j = j(i)$ (то есть для интервала, когда индивид i наблюдался последний раз).

Теперь кусочно-экспоненциальную модель можно оценить, воспринимая индикаторы смерти d_{ij} как если бы это были независимые пуассоновские величины со средними значениями

$$\mu_{ij} = t_{ij}\lambda_{ij},$$

где t_{ij} – время подверженности риску, определенное выше, а λ_{ij} – риск для индивида i в интервале j. Логарифмируя это выражение и учитывая, что риски удовлетворяют модели пропорциональных рисков в уравнении (15), получаем

$$\log \mu_{ij} = \log t_{ij} + \alpha_j + \mathbf{x}_i'\boldsymbol{\beta},$$

где $\alpha_j = \log \lambda_j$, как и прежде.

Таким образом, кусочно-экспоненциальная модель пропорциональных рисков эквивалентна пуассоновской логлинейной модели для псевдонаблюдений, по одному для каждой комбинации индивида и интервала, где индикатор смерти – это зависимая переменная, а логарифм времени подверженности риску входит в уравнение как сдвиг.

Важно отметить, что не предполагается, что d_{ij} имеют независимые пуассоновские распределения, поскольку это, очевидно, не так. Если индивид i умер в интервале $j(i)$, он был

жив во все предыдущие интервалы $j < j(i)$, так что индикаторы, очевидно, не могут быть независимыми. Более того, каждый индикатор может принимать только значения единица или ноль, так что он не может иметь пуассоновское распределение, при котором значениям, большим единицы, соответствует ненулевая вероятность. Результат более тонкий. Совпадают не распределения, а функции правдоподобия. При данной реализации кусочно-экспоненциального процесса выживаемости можно найти такую реализацию набора независимых пуассоновских наблюдений, которая имеет такую же функцию правдоподобия, а значит, ведет к таким же оценкам и результатам тестирования гипотез.

Доказательство несложно. Вспомним из раздела 3.2, что вклад i-го индивида в логарифмическую функцию правдоподобия имеет общий вид

$$\log L_i = d_i \log \lambda_i(t_i) - \Lambda_i(t_i),$$

где $\lambda_i(t)$ – риск, а $\Lambda_i(t)$ – кумулятивный риск для i-го индивида в момент t. Пусть $j(i)$ обозначает интервал, куда попадает t_i, как и раньше.

В кусочно-экспоненциальной модели первый член логарифмической функции правдоподобия можно записать как

$$d_i \log \lambda_i(t_i) = d_{ij(i)} \log \lambda_{ij(i)},$$

используя тот факт, что риск равен $\lambda_{ij(i)}$, когда t_i находится в интервале $j(i)$, а индикатор смерти d_i относится непосредственно к последнему интервалу жизни индивида i и поэтому равен $d_{j(i)}$.

Кумулятивный риск во втором члене выражения – это интеграл, который можно записать в виде суммы следующим образом:

$$\Lambda_i(t_i) = \int_0^{t_i} \lambda_i(t)dt = \sum_{j=1}^{j(i)} t_{ij}\lambda_{ij},$$

где t_{ij} – количество времени, проведенное индивидом i в интервале j. Чтобы увидеть это, заметим, что необходимо проинтегрировать риск от 0 до t_i. Разобьем этот интеграл на сумму интегралов, по одному для каждого интервала, где риск постоянный. Если индивид полностью проживает интервал, вкладом в интеграл будет риск λ_{ij}, умноженный на ширину интервала. Если индивид умирает или цензурируется внутри интервала, вкладом в интеграл будет риск λ_{ij}, умноженный на время, потраченное с начала интервала вплоть до смерти или момента цензурирования, которое равно $t_i - \tau_{j-1}$. Но это в точности совпадает с определением времени подверженности риску t_{ij}.

Одно небольшое отклонение от симметрии в результатах состоит в том, что риск ведет к *одному* члену при $d_{ij(i)} \log \lambda_{ij(i)}$, а кумулятивный риск – к $j(i)$ членам, по одному на каждый интервал от $j = 1$ до $j(i)$. Мы знаем, однако, что $d_{ij} = 0$ для всех $j < j(i)$, так что можно добавить члены при $d_{ij} \log \lambda_{ij}$ для всех предыдущих j; пока $d_{ij} = 0$, они не дают никакого вклада в логарифмическую функцию правдоподобия. Этот прием позволяет записать вклад i-го индивида в логарифмическую функцию правдоподобия как сумму $j(i)$ вкладов, по одному для каждого интервала, прожитого индивидом:

$$\log L_i = \sum_{j=1}^{j(i)} \{d_{ij} \log \lambda_{ij} - t_{ij}\lambda_{ij}\}.$$

Тот факт, что вклад индивида в логарифмическую функцию правдоподобия – это *сумма* нескольких членов (то есть вклад в функцию правдоподобия – это произведение нескольких членов), означает, что можно воспринимать каждый член как соответствующий независимому наблюдению.

Последний шаг – определить вклад каждого псевдонаблюдения, и здесь заметим, что он совпадает, если не считать константы, с функцией правдоподобия, когда d_{ij} имеет пуассоновское распределение со средним $\mu_{ij} = t_{ij}\lambda_{ij}$. Чтобы убедиться в этом, запишем пуассоновскую логарифмическую функцию правдоподобия в виде

$$\log L_{ij} = d_{ij} \log \mu_{ij} - \mu_{ij} = d_{ij} \log(t_{ij}\lambda_{ij}) - t_{ij}\lambda_{ij}.$$

Это выражение совпадает с логарифмической функцией правдоподобия, полученной выше, за исключением члена $d_{ij} \log(t_{ij})$, но это константа, зависящая от данных, но не от параметров, так что ее можно игнорировать с точки зрения оценивания. Доказательство окончено.

Данный результат обобщает наблюдение, сделанное в конце раздела 3.2, о взаимосвязи между функцией правдоподобия для цензурированных экспоненциальных данных и пуассоновской функцией правдоподобия. Расширение состоит в том, что вместо одного «пуассоновского» индикатора смерти для каждого индивида мы имеем по одному подобному индикатору для каждого интервала, прожитого каждым индивидом.

Создание псевдонаблюдений может значительно увеличить размер набора данных, вероятно настолько, что анализ станет невозможным. Заметим, однако, что количество различных комбинаций регрессоров может быть мало, даже если общее число псевдонаблюдений велико. В этом случае можно группировать наблюдения, складывая показатели времени подверженности риску и индикаторы смерти. В этой более общей модели можно определить d_{ij} как число смертей, а t_{ij} – как общее время подверженность риску для индивидов с характеристиками $\mathbf{x}_i$ в интервале j. Как всегда в агрегированных пуассоновских моделях, оценки, стандартные ошибки и тесты отношения правдоподобия будут в точности такими же, как для индивидуальных данных. Конечно, остатки моделей будут разными, представляя собой точность подгонки агрегированных, а не индивидуальных данных, но это можно рассматривать как низкую цену по сравнению с удобством работы с небольшим количеством объектов.

5.4 Меняющиеся во времени регрессоры

Из предыдущего анализа должно быть очевидно, что можно легко включить в модель меняющиеся во времени регрессоры, если они меняют значение только на границах интервалов. При создании псевдонаблюдений, требуемых для формирования пуассоновской логарифмической функции правдоподобия, обычно повторяют вектор регрессоров $\mathbf{x}_i$, создавая копии $\mathbf{x}_{ij}$, по одной для каждого интервала. Тем не менее, ничто в нашем анализе не требует, чтобы эти векторы были одинаковыми. Следовательно, можно переопределить вектор $\mathbf{x}_{ij}$ как представляющий значения регрессоров для индивида i в интервале j, и продолжать как обычно, переписав модель в виде

$$\log \lambda_{ij} = \alpha_j + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

Требование того, чтобы регрессоры сменяли значение только на границах интервала, может показаться ограничительным, но в действительности модель более гибкая, чем кажется на первый взгляд, поскольку всегда можно снова разбить псевдонаблюдения. Например, если бы мы хотели ввести изменение регрессоров для индивида i в середине интервала j, можно было бы разбить одно псевдонаблюдение на два, первое со старыми и второе с новыми значениями регрессоров. Каждая половина получает свою собственную меру времени подверженности риску и индикатор смерти, но обе будут помечены как принадлежащие одному интервалу, так что им будет соответствовать один и тот же базовый риск. Все шаги приведенного выше доказательства по-прежнему верны.

Конечно, дальнейшее разбиение наблюдений увеличивает размер набора данных, и всегда будут существовать практические ограничения на то, как долго можно следовать этому

подходу, даже при использовании сгруппированных данных. Альтернативой является использование более простых индикаторов, таких как среднее значение регрессора в интервале, возможно лагированное, чтобы избежать прогнозирования текущего риска с помощью будущих значений регрессоров.

5.5 Зависящие от времени коэффициенты

Оказывается, что кусочно-экспоненциальная модель позволяет легко включить в анализ непропорциональные риски или зависящие от времени коэффициенты, снова предполагая, что эти коэффициенты меняются лишь на границах интервалов.

В качестве иллюстрации предположим, что есть один регрессор, принимающий значение x_{ij} для индивида i в интервале j. Предположим далее, что этот регрессор является фиктивной переменной, так что его возможные значения – единица и ноль. На данный момент неважно, является ли значение постоянным для индивида или меняется от интервала к интервалу.

В модели пропорциональных рисков мы бы записали

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij},$$

где β отражает эффект регрессора на логарифм риска в каждый данный момент времени. Беря экспоненту, получаем, что риск при $x = 1$ равен $\exp\{\beta\}$, помноженному на риск при $x = 0$, и этот эффект один и тот же во все моменты времени. Это простая аддитивная модель относительно времени и имеющегося регрессора.

Чтобы позволить коэффициенту зависеть от времени, запишем

$$\log \lambda_{ij} = \alpha_j + \beta_j x_{ij},$$

где β_j отражает эффект регрессора на риск в интервале j. Беря экспоненту, получаем, что риск в интервале j при $x = 1$ равен $\exp\{\beta_j\}$, помноженному на риск в интервале j при $x = 0$, так что эффект может меняться от интервала к интервалу. Поскольку воздействие регрессора зависит от интервала, получаем своего рода взаимодействие регрессора и времени, что становится более очевидным, если записать модель в виде

$$\log \lambda_{ij} = \alpha_j + \beta x_{ij} + (\alpha\beta)_j x_{ij}.$$

Эти модели напоминают модели ковариационного анализа. Здесь α играет роль константы, а β – роль коэффициента наклона. Модель пропорциональных рисков имеет различные константы и один и тот же коэффициент наклона, так что она аналогична модели параллельных прямых. Модель с зависящими от времени коэффициентами имеет различные константы и различные коэффициенты наклона и является аналогичной модели с взаимодействием.

Итак, можно учесть непропорциональность рисков просто путем введения взаимодействий со временем. Конечно, можно также тестировать предположение о пропорциональности рисков, проверяя значимость взаимодействия со временем. Теперь мы готовы рассмотреть пример.

6 Младенческая и детская смертность в Колумбии

Для иллюстрации применения кусочно-экспоненциальных моделей выживаемости возьмем данные о младенческой и детской смертности в Колумбии из работы Сомосы (Somoza, 1980). Данные собраны в 1976 году в ходе опроса, проведенного в рамках Всемирного обследования рождаемости. Выборка состояла из женщин в возрасте от 15 до 49 лет. Анкета содержала историю материнства, включая пол, дату рождения и (если применимо на момент проведения опроса) возраст смерти для каждого ребенка опрашиваемой женщины.

6.1 Подсчет событий и подверженности риску

Как часто случается с данными о выживаемости, большая часть работы уходит на их подготовку для анализа. В рассматриваемом случае мы начали с таблиц, приведенных в статье Сомосы, в которых содержатся данные о выживших детях, сгруппированные по текущему возрасту, и умерших детях, сгруппированные по возрасту смерти. В обеих таблицах возраст указан в соответствии с разбиением возрастов из таблицы 1, использующим короткие интервалы времени в начале жизни, когда риск смерти высок, но быстро падает, и широкие интервалы в более позднем возрасте. Используя эти два бита информации, мы подсчитали количество смертей и время подверженности риску по возрастным группам, предполагая, что дети, которые умерли или были цензурированы внутри интервала, проживали в среднем половину длины интервала.

Таблица 1: Младенческие и детские смерти и подверженность риску по возрасту ребенка и когорте рождения, Колумбия, 1976.

Точный возраст	Когорта рождения					
	1941–59		1960–67		1968-76	
	число смертей	подвер-сть риску	число смертей	подвер-сть риску	число смертей	подвер-сть риску
0–1 мес	168	278,4	197	403,2	195	495,3
1–3 мес	48	538,8	48	786,0	55	956,7
3–6 мес	63	794,4	62	1165,3	58	1381,4
6–12 мес	89	1550,8	81	2294,8	85	2604,5
1–2 года	102	3006,0	97	4500,5	87	4618,5
2–5 лет	81	8743,5	103	13201,5	70	9814,5
5–10 лет	40	14270,0	39	19525,0	10	5802,5

В таблице 1 показаны результаты этих подсчетов в терминах числа смертей и общего числа человеко-лет подверженности риску от рождения и до десятилетнего возраста по категориям возраста ребенка для трех групп детей (или когорт), рожденных в 1941–59, 1960–67 и 1968–76 гг. Цель анализа – оценить величину ожидаемого снижения младенческой и детской смертности в этих когортах и изучить, снижалась ли смертность одинаково быстро во всех возрастах или быстрее в определенных возрастных группах.

6.2 Подгонка пуассоновских моделей

Пусть y_{ij} обозначает число смертей для когорты i ($i = 1, 2, 3$) в возрастной группе j ($j = 1, 2, \ldots, 7$). В свете результатов предыдущего раздела можно рассматривать y_{ij} как реализации пуассоновских случайных величин со средними μ_{ij}, равными

$$\mu_{ij} = \lambda_{ij} t_{ij},$$

где λ_{ij} – риск, а t_{ij} – общее время подверженности риску для группы i в возрасте j. Иными словами, ожидаемое число смертей равно произведению коэффициента смертности на время подверженности риску.

Предостережение по поводу единиц измерения: риск должен измеряться в тех же самых единицах времени, которые использовались для измерения подверженности риску. В нашем примере время измеряется в годах, и следовательно, λ_{ij} представляет собой риск на человеко-год подверженности. Если бы время измерялось в месяцах, λ_{ij} представляло бы собой риск на человеко-месяц подверженности, и было бы в точности равно одной двенадцатой от размера риска на человеко-год.

Для моделирования рисков будем использовать связующее лог-преобразование, так что линейный предиктор принимает вид

$$\eta_{ij} = \log \mu_{ij} = \log \lambda_{ij} + \log t_{ij},$$

то есть является суммой двух частей, $\log t_{ij}$, *сдвига*, или известной части линейного предиктора, и $\log \lambda_{ij}$, логарифма риска.

Наконец, рассмотрим логлинейную модель для риска в обычном виде

$$\log \lambda_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta},$$

где $\mathbf{x}_{ij}$ – вектор регрессоров. В случае, если читатель задумался о том, что произошло с базовым риском, поясним, что он содержится внутри вектора параметров $\boldsymbol{\beta}$. Вектор регрессоров $\mathbf{x}_{ij}$ может включать константу, набор фиктивных переменных, представляющих возрастные группы (например, форму риска в зависимости от возраста), набор фиктивных переменных, представляющих когорты рождения (то есть отвечающих за изменение риска во времени) и даже набор мультипликативных фиктивных переменных, представляющих произведения возрастов и когорт рождения (эффекты взаимодействия).

Таблица 2: Сумма квадратов остатков (СКО) для различных моделей, оцененных по данным о младенческой и детской смертности в Колумбии

Модель	Название	$\log \lambda_{ij}$	СКО	Степени свободы
ϕ	Нулевая	η	4239,8	20
A	Возраст	$\eta + \alpha_i$	72,7	14
C	Когорта	$\eta + \beta_j$	4190,7	18
$A + C$	Аддитивная	$\eta + \alpha_i + \beta_j$	6,2	12
AC	Насыщенная	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

В таблице 2 показаны суммы квадратов отклонений для пяти возможных моделей, включая нулевую модель, две однофакторные модели, двухфакторную аддитивную модель и двухфакторную модель с эффектом взаимодействия, которая является насыщенной для этих данных.

6.3 Эквивалентные модели выживаемости

Нулевая модель предполагает, что риск постоянен с рождения до десятилетнего возраста и одинаков для всех когорт. Таким образом, она соответствует *экспоненциальной модели выживаемости без регрессоров*. Очевидно, такая модель не подходит под данные, сумма квадратов остатков, равная 4239,8 при 20 степенях свободы – это астрономическая величина. ММП-оценка η равна $-3,996$ со стандартной ошибкой 0,0237. Беря экспоненту, получаем оценку риска, равную 0,0184. То есть ожидается около 18 смертей на тысячу человеко-лет подверженности риску. Можно проверить, что 0,0184 – это просто отношение общего числа смертей к общему времени подверженности риску. Умножая 0,0184 на подверженность риску в каждой ячейке таблицы, получаем ожидаемое число смертей. Сумма квадратов остатков, приведенная выше, – это просто удвоенная сумма наблюдаемых смертей, помноженных на логарифм отношения наблюдаемых смертей к ожидаемым.

Модель с возрастом позволяет риску меняться от одной возрастной группы к другой, но предполагает, что риск в каждом возрасте одинаков для всех когорт. Она эквивалентна, таким образом, *кусочно-экспоненциальной модели выживаемости без регрессоров*. Снижение СКО по сравнению с нулевой моделью составляет 4167,1 при 6 степенях свободы, и очень значимо. Риск смерти значительно варьируется с возрастом во время нескольких первых

месяцев жизни. Иными словами, риск, очевидно, не является постоянным. Заметим, что при СКО 72,7 и 14 степенях свободы эта модель не подходит под данные. Значит, предпосылка о том, что все когорты подвержены одинаковому риску, не кажется надежной.

В таблице 3 показаны оценки параметров для однофакторных моделей A и C и для аддитивной модели $A+C$ в формате, напоминающем множественный классификационный анализ. Хотя модель A не подходит под данные, будет поучительным кратко обсудить оценки параметров. Константа, указанная в скобках, соответствует риску $\exp\{-0{,}7427\} = 0{,}4758$, что составляет приблизительно половину смерти на человеко-год подверженности риску в первый месяц жизни. Оценка для возраста от 1 до 3 месяцев соответствует мультипликативному эффекту $\exp\{-1{,}973\} = 0{,}1391$, что составляет 86%-ное снижение риска после выживания в первый месяц жизни. Этот понижающий тренд продолжается до возраста 5–10 лет, когда мультипликативный эффект равен $\exp\{-5{,}355\} = 0{,}0047$, указывая на то, что риск в этих возрастах составляет лишь полпроцента того, каким он был в первый месяц жизни. Можно убедиться в том, что ММП-оценки возрастных эффектов можно подсчитать напрямую по общему числу смертей и общему времени подверженности риску в каждой возрастной группе. Можете ли Вы вычислить СКО вручную?

Рассмотрим теперь модель, включающую только когорту рождения, в которой предполагается, что риск постоянен с момента рождения до десяти лет, но меняется от одной когорты к другой. Эта модель эквивалентна *трем экспоненциальным моделям выживаемости*, по одной для каждой когорты рождения. Как и следовало ожидать, она безнадежно неадекватна, с СКО в несколько тысяч, поскольку не принимает во внимание значительные возрастные эффекты, которые только что обсуждались. Интересно, тем не менее, взглянуть на оценки параметров в таблице 3. В первом приближении, общая норма смертности для старшей когорты равнялась $\exp\{-3{,}899\} = 0{,}0203$, или около 20 смертей на тысячу человеко-лет подверженности риску. Мультипликативный эффект для когорты, рожденной в 1960–1967 гг., равен $\exp\{-0{,}3020\} = 0{,}7393$, что соответствует 26%-ному снижению общей смертности. Однако мультипликативный эффект для самой молодой когорты равен $\exp\{0{,}0742\} = 1{,}077$, что соответствует 8%-ному *увеличению* общей смертности. Можете ли Вы найти объяснение этой явной аномалии? Мы дадим ответ на этот вопрос после рассмотрения следующей модели.

Таблица 3: Оценки параметров для моделей A, C и $A + C$ по данным о младенческой и детской смертности в Колумбии

Фактор	Категория	Общий эффект	Чистый эффект
База			−0,4485
Возраст	0–1 мес	(−0,7427)	–
	1–3 мес	−1,973	−1,973
	3–6 мес	−2,162	−2,163
	6–12 мес	−2,488	−2,492
	1–2 года	−3,004	−3,014
	2–5 лет	−4,086	−4,115
	5–10 лет	−5,355	−5,436
Когорта	1941–59	(−3,899)	–
	1960–67	−0,3020	−0,3243
	1968–76	0,0742	−0,4784

Рассмотрим теперь аддитивную модель с эффектами как возраста, так и когорты, в которой риск может меняться с возрастом и может быть различным для разных когорт, но эффект возраста (когорты) предполагается одинаковым для каждой когорты (возраста).

Эта модель эквивалентна *модели пропорциональных рисков*, в которой предполагается общая форма риска для каждого возраста, а когорта пропорционально влияет на риск во всех возрастах. Сравнивая модель пропорциональных рисков с моделью A, отметим снижение СКО на 66,5 при потере двух степеней свободы, что составляет очень значимый эффект. Это свидетельствует о сильных когортных эффектах, очищенных от влияния возраста. С другой стороны, полученное СКО 6,2 при 12 степенях свободы, очевидно, не является значимым, что означает, что модель пропорциональных рисков адекватно описывает смертность в Колумбии в зависимости от возраста и когорты рождения. Иными словами, предположение о пропорциональности рисков достаточно разумно, откуда следует, что снижение смертности в Колумбии было одним и тем же во всех возрастах.

Рассмотрим оценки параметров в самом правом столбце таблицы 3. Константа – это базовый риск в возрасте 0–1 месяцев для самой первой когорты, то есть рожденных в 1941–59 годы. Параметры возраста, представляющие базовый риск, практически не меняются по сравнению с моделью, учитывающей только возрастной эффект, и соответствуют сильному снижению смертности с момента рождения до десятилетнего возраста, причем половина снижения приходится на первый год жизни. Когортные эффекты, подправленные на возраст, дают более разумную картину снижения смертности во времени. Мультипликативные эффекты для когорт, рожденных в 1960–1967 гг. и 1968–1976 гг., равны $\exp\{-0{,}3243\} = 0{,}7233$ и $\exp\{-0{,}4784\} = 0{,}6120$, что соответствует снижению смертности на 28 и 38 %% в каждом возрасте, по сравнению с когортой, рожденной в 1941–59 годы. Это удивительное снижение младенческой и детской смертности, которое было одним и тем же для всех возрастов. Иными словами, смертность новорожденных, младенцев и тех, кто только начинает ходить, снизилась примерно на 38 процентов среди этих когорт.

Тот факт, что общий эффект для младшей когорты был положителен, а чистый эффект отрицателен и значителен, можно объяснить следующим образом. Поскольку обследование проводилось в 1976 году, дети, родившиеся между 1968 и 1976 годами, были подвержены по большей части смертности в молодых возрастах, когда нормы смертности значительно выше, чем в старших возрастах. Например, ребенок, родившийся в 1975 году, был подвержен только риску смерти в первый год жизни. Общий эффект игнорирует этот факт и, таким образом, переоценивает смертность в этой группе в возрасте от 0 до 10. Чистый эффект делает необходимую поправку на повышенный риск в молодых возрастах, в сущности сравнивая смертность в данной когорте со смертностью в более ранних когортах, когда они были в том же возрасте, а следовательно, способен раскрыть истинное снижение.

Заключительное предостережение по поводу интерпретации: данные получены ретроспективно со слов матерей, которым было от 15 до 49 лет в момент проведения опроса. Эти женщины представляют собой репрезентативную выборку как матерей, так и рождений для недавних периодов, но несколько смещенную для более ранних периодов. Выборка исключает матерей, умерших до проведения опроса, а также тех, кто был старше в момент рождения ребенка. Например, рождения в 1976, 1966 и 1956 гг. относятся к матерям, которым на момент рождения ребенка было меньше 50, 40 и 30 лет, соответственно. При более аккуратном анализе данных следовало бы включить возраст матери при рождении ребенка в качестве дополнительной контрольной переменной.

6.4 Оценка вероятностей выживания

До сих пор наше внимание было сконцентрировано на риске или смертности, но, конечно, как только подсчитан риск, легко найти кумулятивный риск, а значит, вероятности выживания. В таблице 4 представлены результаты такого упражнения, с использованием оценок параметров из модели пропорциональных рисков в таблице 3.

Рассмотрим сначала базовую группу, а именно когорту детей, рожденных до 1960 года. Для получения лог-риска для каждой возрастной группы надо сложить константу и возрастной

Таблица 4: Подсчет вероятностей выживания для трех когорт на основе модели пропорциональных рисков

Возрастная группа	Ширина	База			Выживаемость для когорты		
		лог-риск	риск	кум.риск	<1960	1960–67	1968–76
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0–1 мес	1/12	−0,4485	0,6386	0,0532	0,9482	0,9623	0,9676
1–3 мес	2/12	−2,4215	0,0888	0,0680	0,9342	0,9520	0,9587
3–6 мес	3/12	−2,6115	0,0734	0,0864	0,9173	0,9395	0,9479
6–12 мес	1/2	−2,9405	0,0528	0,1128	0,8933	0,9217	0,9325
1–2 года	1	−3,4625	0,0314	0,1441	0,8658	0,9010	0,9145
2–5 лет	3	−4,5635	0,0104	0,1754	0,8391	0,8809	0,8970
5–10 лет	5	−5,8845	0,0028	0,1893	0,8275	0,8721	0,8893

эффект, например, лог-риск для возраста 1–3 мес. равен $-0{,}4485 - 1{,}973 = -2{,}4215$. Это дает числа в столбце (3) таблицы 3. Далее берем экспоненту для получения риска в столбце (4), например, риск для возраста 1–3 мес. равен $\exp\{-2{,}4215\} = 0{,}0888$. Далее подсчитываем кумулятивный риск, умножаем его на ширину интервала и суммируем по всем интервалам. На этом шаге необходимо выразить ширину интервала в тех же единицах, которые используются для подсчета подверженности риску, в данном случае в годах. Таким образом, кумулятивный риск в конце периода 1–3 мес. равен $0{,}6386 \times 1/12 + 0{,}0888 \times 2/12 = 0{,}0680$. Наконец, заменим знак и возьмем экспоненту для подсчета функции выживания. Например, базовая функция выживания в возрасте 3 месяцев равна $\exp\{-0{,}0680\} = 0{,}9342$.

Для подсчета значений функций выживания, показанных в столбцах (7) и (8), для других двух когорт, можно было бы умножить базовые риски на $\exp\{-0.3242\}$ и $\exp\{-0.4874\}$, чтобы получить риски для когорт 1960–67 и 1968–76, соответственно, а затем повторить шаги, описанные выше, для получения функций выживания. Этот подход был бы необходим при наличии зависящих от времени коэффициентов, но в настоящем случае можно воспользоваться упрощением, которое дает модель пропорциональных рисков. А именно, функции выживания для двух младших когорт можно подсчитать как базовую функцию выживания, *возведенную в степень*, равную относительным рискам $\exp\{-0.3242\}$ и $\exp\{-0.4874\}$, соответственно. Например, вероятность дожить до трех месяцев равна 0.9342 для базовой группы, и, соответственно, оказывается равной $0.9342^{\exp\{-0.3242\}} = 0.9520$ для когорты, рожденной в 1960–1967 гг., и $0{,}9342^{\exp\{-0{,}4874\}} = 0{,}9587$ для когорты 1968–1976 гг.

Заметим, что вероятность умереть во время первого года жизни уменьшилась со 106,7 из тысячи для детей, рожденных до 1960 г., до 78,3 из тысячи для детей, рожденных в 1960–1967 гг., и, наконец, до 67,5 из тысячи для наиболее молодой когорты. Результаты, представленные в терминах вероятностей, часто доступнее для широкой аудитории, чем результаты, представленные в терминах норм риска. (К сожалению, демографы обычно называют вероятность умереть в течение первого года жизни «нормой младенческой смертности». Это неверно, поскольку эта величина представляет собой вероятность, а не норму. В нашем примере норма значительно меняется на протяжении первого года жизни. Если вероятность умереть в течение первого года жизни равна, скажем, q, то средняя норма равна приблизительно $-\log(1-q)$, что несильно отличается от q для малых значений q.)

Концентрируясь на событиях и подверженности риску, мы смогли объединить анализ младенческой и детской смертности в рамках одной модели и использовать всю доступную информацию. Альтернативный подход – сосредоточиться на младенческой смертности (когда смерть наступает на первом году жизни), и решать задачу цензурирования, рассматривая только детей, рожденных по крайней мере за год до обследования, для которых известно, дожили ли они до одного года. Затем можно было бы исследовать вероятность дожития до

возраста 1 год, используя обычные логит-модели. В качестве дополняющего анализа можно было бы взглянуть на дожитие от возраста 1 год до, скажем, 5 лет, работая с данными о детях, рожденных по крайней мере за 5 лет до обследования и доживших до одного года, а затем анализируя, доживают они или нет до пятилетнего возраста, снова используя логит-модель. Будучи простым, данный подход не полностью использует информацию, полагаясь только на полные (нецензурированные) данные. Cox & Oakes (1984) показывают, что этот так называемый подход с уменьшенной выборкой может давать несостоятельные результаты. Другой недостаток этого подхода в том, что он концентрируется на дожитии до ключевых возрастов, но не позволяет исследовать форму риска в промежуточный период.

7 Модели в дискретном времени

Обсудим кратко два расширения модели пропорциональных рисков для дискретного времени, начиная с определения функций риска и выживания в дискретном времени, а затем перейдем к моделям на основе логит- и дополнительного лог-лог-преобразований.

7.1 Дискретные функции риска и выживания

Пусть T – дискретная случайная величина, принимающая значения $t_1 < t_2 < \ldots$ с вероятностями

$$f(t_j) = f_j = \mathbb{P}\{T = t_j\}.$$

Определим функцию выживания в момент t_j как вероятность того, что время жизни T не меньше t_j:

$$S(t_j) = S_j = \mathbb{P}\{T \geq t_j\} = \sum_{k=j}^{\infty} f_j.$$

Далее определим риск в момент t_j как вероятность умереть в этот момент времени при условии, что индивид дожил до этого момента, то есть

$$\lambda(t_j) = \lambda_j = \mathbb{P}\{T = t_j | T \geq t_j\} = \frac{f_j}{S_j}. \tag{17}$$

Заметим, что в дискретном времени риск – это условная вероятность, а не норма. Тем не менее, общий результат, выражающий риск как отношение плотности к функции выживания, остается верным.

Следующий интересный результат для дискретного времени состоит в том, что функцию выживания в момент времени t_j можно записать в терминах риска за все предыдущие моменты времени, $t_1, \ldots, t_{j-1}$:

$$S_j = (1 - \lambda_1)(1 - \lambda_2) \ldots (1 - \lambda_{j-1}). \tag{18}$$

Иными словами, данный результат утверждает, что, чтобы дожить до момента t_j, необходимо сначала дожить до t_1, затем дожить до t_2 при условии дожития до t_1 и так далее, наконец, пережить t_{j-1} условно на дожитии до этого момента. Этот результат аналогичен результату, связывающему функцию выживания в непрерывном времени с интегральным или кумулятивным риском во все предыдущие моменты времени.

Примером процесса выживаемости, происходящего в дискретном времени, может быть время до зачатия, измеряемое в менструальных циклах. В этом случае возможные значения T – это все положительные целые числа, f_j – это вероятность зачатия во время j-го цикла, S_j – вероятность зачатия во время j-го цикла или позже, а λ_j – вероятность зачатия во время

j-го цикла при условии, что оно не произошло ранее. Результат, связывающий функцию выживания и риск, утверждает, что, чтобы добраться до j-го цикла без зачатия, необходимо его отсутствие во время первого цикла, затем во время второго при условии неудавшегося зачатия во время первого цикла, и т.д., и, наконец, во время $(j-1)$-го цикла при условии его отсутствия на более ранней стадии.

7.2 Дискретная функция выживания и логистическая регрессия

Cox (1972) предложил в качестве расширения модели пропорциональных рисков для дискретного времени работать с условными вероятностями смерти в каждый момент времени t_j при условии дожития до этого момента времени. В частности, он предложил такую модель:

$$\frac{\lambda(t_j|\mathbf{x}_i)}{1-\lambda(t_j|\mathbf{x}_i)} = \frac{\lambda_0(t_j)}{1-\lambda_0(t_j)} \exp\{\mathbf{x}_i'\boldsymbol{\beta}\},$$

где $\lambda(t_j|\mathbf{x}_i)$ – риск в момент t_j для индивида со значениями регрессоров $\mathbf{x}_i$, $\lambda_0(t_j)$ – базовый риск в момент t_j, а $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ – относительный риск, связанный со значениями регрессоров $\mathbf{x}_i$.

Логарифмируя, получаем модель относительно *логита* риска, или условной вероятности умереть в момент t_j при условии дожития до этого момента:

$$\operatorname{logit} \lambda(t_j|\mathbf{x}_i) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{19}$$

где $\alpha_j = \operatorname{logit}\lambda_0(t_j)$ – логит базового риска, а $\mathbf{x}_i'\boldsymbol{\beta}$ – эффект регрессоров на логит риска. Заметим, что, в сущности, время в модели является дискретным фактором, поскольку для каждого возможного времени смерти t_j присутствует один параметр α_j. Интерпретация параметров $\boldsymbol{\beta}$, связанных с другими регрессорами, следует той же логике, что и при логистической регрессии.

В действительности есть дальнейшая аналогия с логистической регрессией: можно оценить модель пропорциональных рисков в дискретном времени с помощью оценки логистической регрессии для набора псевдонаблюдений, сгенерированных следующим образом. Предположим, что индивид i умирает или цензурируется в момент времени $t_{j(i)}$. Создадим индикаторы смерти d_{ij}, которые принимают значение единица, если индивид i умер в момент времени j, и ноль в противном случае, по одному для каждого момента времени от t_1 до $t_{j(i)}$. С каждым из этих индикаторов свяжем копию вектора регрессоров $\mathbf{x}_i$ и индекс j, обозначающий момент времени. Тогда модель пропорциональных рисков (19) можно оценить, воспринимая d_{ij} как независимые наблюдения Бернулли с вероятностью успеха, задаваемой риском λ_{ij} для индивида i в момент времени t_j.

В более общем плане можно сгруппировать псевдонаблюдения с одинаковыми значениями регрессоров. Пусть d_{ij} обозначает число смертей, а n_{ij} – общее число индивидов со значениями регрессоров $\mathbf{x}_i$, наблюдаемых в момент времени t_j. Тогда можно воспринимать d_{ij} как биномиальную величину с параметрами n_{ij} и λ_{ij}, где последняя удовлетворяет модели пропорциональных рисков.

Доказательство этого результаты следует тем же шагам, что и доказательство эквивалентности пуассоновской функции правдоподобия и функции правдоподобия для кусочно-экспоненциальных данных о выживаемости при неинформативном цензурировании в разделе 5.3, и основано на уравнении (18), в котором вероятность дожить до времени t_j записана как произведение условных рисков за все предыдущие моменты времени. Важно отметить, что не предполагается, что псевдонаблюдения независимы и имеют бернуллиевское или биномиальное распределение. Мы лишь отмечаем, что функция правдоподобия для модели выживаемости в дискретном времени при неинформативном цензурировании совпадает с биномиальной функцией правдоподобия, которую бы мы получили, воспринимая индикаторы смерти как независимые бернулллиевские или биномиальные случайные величины.

Меняющиеся во времени регрессоры и зависящие от времени коэффициенты можно легко включить в эту модель, следуя тем же шагам, что и ранее. В случае меняющихся во времени регрессоров заметим, что важны только значения регрессоров в дискретные моменты времени $t_1 < t_2 < \ldots$. Зависящие от времени коэффициенты моделируются как взаимодействия между регрессорами и дискретным фактором (или набором фиктивных переменных), представляющим время.

7.3 Дискретная функция выживания и связующее дополнительное лог-лог-преобразование

Альтернативное расширение модели пропорциональных рисков для дискретного времени начинается с функции выживания, которую в рамках модели пропорциональных рисков можно записать в виде

$$S(t_j|\mathbf{x}_i) = S_0(t_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

где $S(t_j|\mathbf{x}_i)$ – вероятность того, что индивид со значениями регрессоров $\mathbf{x}_i$ доживет до момента t_j, а $S_0(t_j)$ – базовая функция выживания. Вспомнив уравнение (18) для дискретной функции выживания, получаем похожее выражение для дополнения функции риска, а именно:

$$1 - \lambda(t_j|\mathbf{x}_i) = [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},$$

и, разрешая относительно риска для индивида i в момент времени t_j, получаем модель

$$\lambda(t_j|\mathbf{x}_i) = 1 - [1 - \lambda_0(t_j)]^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}.$$

Дополнительное лог-лог-преобразование делает правую часть линейной функцией параметров. Применяя это преобразование, получаем модель

$$\log(-\log(1 - \lambda(t_j|\mathbf{x}_i))) = \alpha_j + \mathbf{x}_i'\boldsymbol{\beta}, \tag{20}$$

где $\alpha_j = \log(-\log(1 - \lambda_0(t_j)))$ – дополнительное лог-лог-преобразование базового риска.

Модель можно оценить по дискретным данным о выживаемости, генерируя псевдонаблюдения как ранее и подгоняя обобщенную линейную модель с биномиальной структурой ошибок и связующим дополнительным лог-лог-преобразованием. Иными словами, эквивалентность между биномиальной функцией правдоподобия и функцией правдоподобия для функции выживания в дискретном времени при неинформативном цензурировании сохраняется как для логит, так и для дополнительного лог-лог преобразований.

Интересно отметить, что эту модель можно получить, группируя время в модели пропорциональных рисков в непрерывном времени. Чтобы это увидеть, предположим, что время непрерывно, и нам интересна стандартная модель пропорциональных рисков

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

Предположим, однако, что время сгруппировано в интервалы с границами $0 = \tau_0 < \tau_1 < \ldots < \tau_J = \infty$, и что все, что наблюдается, – это переживает ли индивид интервал или умирает в нем. Заметим, что эта конструкция накладывает некоторые ограничения на цензурирование. Если индивид цензурируется в некоторой точке внутри интервала, неизвестно, пережил бы он этот интервал или нет. Следовательно, необходимо цензурировать его в конце предыдущего интервала, являющегося последней точкой, для которой имеется полная информация. В отличие от кусочно-экспоненциальной модели в данном случае нельзя использовать информацию о подверженности риску в части интервала. С другой стороны, как оказывается, необязательно предполагать, что риск постоянный в каждом интервале.

Пусть λ_{ij} обозначает дискретный риск или условную вероятность того, что индивид i умрет в интервале j при условии того, что он был жив в начале интервала. Эта вероятность равна дополнению условной вероятности выживания в интервале при условии, что индивид был жив в начале интервала, и ее можно записать в виде

$$\begin{aligned}
\lambda_{ij} &= 1 - \mathbb{P}\{T_i > \tau_j | T_i > \tau_{j-1}\} \\
&= 1 - \exp\left\{ - \int_{\tau_{j-1}}^{\tau_j} \lambda(t|\mathbf{x}_i) dt \right\} \\
&= 1 - \exp\left\{ - \int_{\tau_{j-1}}^{\tau_j} \lambda_0(t) dt \right\}^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}} \\
&= 1 - (1 - \lambda_j)^{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}},
\end{aligned}$$

где λ_j – базовая вероятность смерти в интервале j при условии дожития до начала этого интервала. Вторая строчка следует из уравнения (4), связывающего функцию выживания и интегральный риск, третья строчка следует из предпосылки о пропорциональности рисков, а последняя строчка определяет λ_j.

Как отмечено в Kalbfleish & Prentice (1980, стр. 37), «эта дискретная модель, таким образом, является единственной подходящей для сгруппированных данных из модели пропорциональных рисков в непрерывном времени». На практике, тем не менее, модель со связующим логит-преобразованием применяется гораздо чаще, чем модель с дополнительным лог-лог-преобразованием, вероятно потому, что логистическая регрессия лучше известна, чем обобщенные линейные модели со связующим дополнительным лог-лог-преобразованием, и поскольку программное обеспечение для первой из них более широко доступно, чем для второй. В действительности логит-модель часто используется в случаях, когда более адекватной была бы кусочно-экспоненциальная модель, возможно потому, что логистическая регрессия лучше известна, чем пуассоновская.

В заключение полезно было бы сделать некоторые предложения относительно выбора подхода к анализу выживаемости при использовании обобщенных линейных моделей:

- Если время на самом деле дискретно, разумно использовать модель в дискретном времени со связующим логит-преобразованием, которая имеет прямую интерпретацию в терминах условных вероятностей, и легко реализуется с помощью стандартного программного обеспечения для логистической регрессии.

- Если время непрерывно, но наблюдается в сгруппированной форме, то связующее дополнительное лог-лог-преобразование кажется более подходящим. В частности, результаты на основе дополнительного лог-лог-преобразования должны быть более устойчивы к выбору категорий, чем результаты на основе логит-преобразования. Тем не менее, в этом случае нельзя учесть частичную подверженность риску при дискретном времени, независимо от применяемого преобразования.

- Если время непрерывно и хочется предполагать, что риск постоянен внутри каждого интервала, то кусочно-экспоненциальный подход на основе пуассоновской функции правдоподобия предпочтителен. Этот подход достаточно устойчив к выбору категорий и уникален в плане возможности использования информации в случаях частичной подверженности риску.

Наконец, если время на самом деле непрерывно, и хочется оценить эффекты регрессоров, не делая каких-либо предположений о базовом риске, то метод частного правдоподобия из Cox (1972) является очень привлекательным.

Литература

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

Cox, D.R. & D. Oakes (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics* 36, 299–306.

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

Laird, N. & D. Oliver (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of American Statistical Association* 74, 231–240.

Somoza, J.L. (1980). Illustrative analysis: Infant and child mortality in Colombia. *World Fertility Survey Scientific Reports* 10. London: World Fertility Survey.

Survival models

Germán Rodríguez

Princeton University, Princeton, USA

This essay is an introduction to survival models in the context of generalized linear models. We introduce the hazard and survival functions and describe the most common censoring mechanisms and the resulting likelihood function. We discuss the main approaches to modeling waiting times, including accelerated life and proportional hazard models, with extensions to time-varying covariates and time-dependent effects. We then focus on the piece-wise exponential survival model and note its equivalence with Poisson regression models. We illustrate this approach with an application to the analysis of infant and child mortality in Colombia using survey data. We conclude with a brief discussion of discrete time models and their equivalence with logistic regression.

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

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

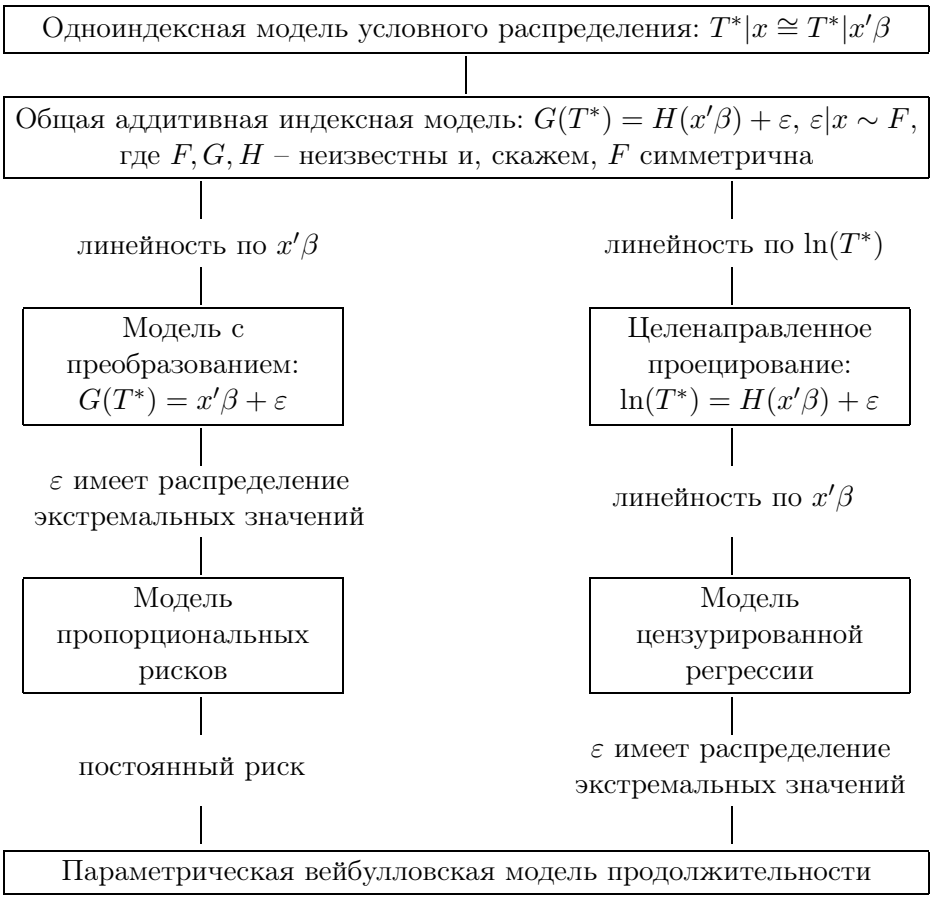
На рисунке 1 показана логическая связь между этими моделями. Все они являются частными случаями *одноиндексной модели*, в которой условное распределение зависимой переменной зависит от регрессоров x исключительно через индекс $x'\beta$. Модель пропорциональных рисков и модель цензурированной регрессии 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.