Исследование специальных моделей кривых дожития в условиях неполных данных

Коробейников Антон Иванович

Санкт-Петербургский государственный университет

Защита диссертации по специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ»

02 декабря 2010 г.

Введение

Анализ данных типа времени жизни:

- X — время до наступления некоторого события («время наработки на отказ»).
- Приложения в медицине, демографии, социологии, страховой и финансовой математики, теории надежности.

Модели данных

Используемые экспериментаторами модели данных относятся к т.н. «нерегулярному» типу:

- Негладкие плотности.
- Зависящие от параметров носители распределений.

Примеры: модель Гомперца-Макегама, модель ExpCos А.Г. Барта, дополнительный параметр сдвига в классических моделях (Вейбулла, гамма, лог-нормальное распределения).

Проблема: классические методы оценивания либо неприменимы, либо требуют серьезной модификации для работы с подобными моделями.

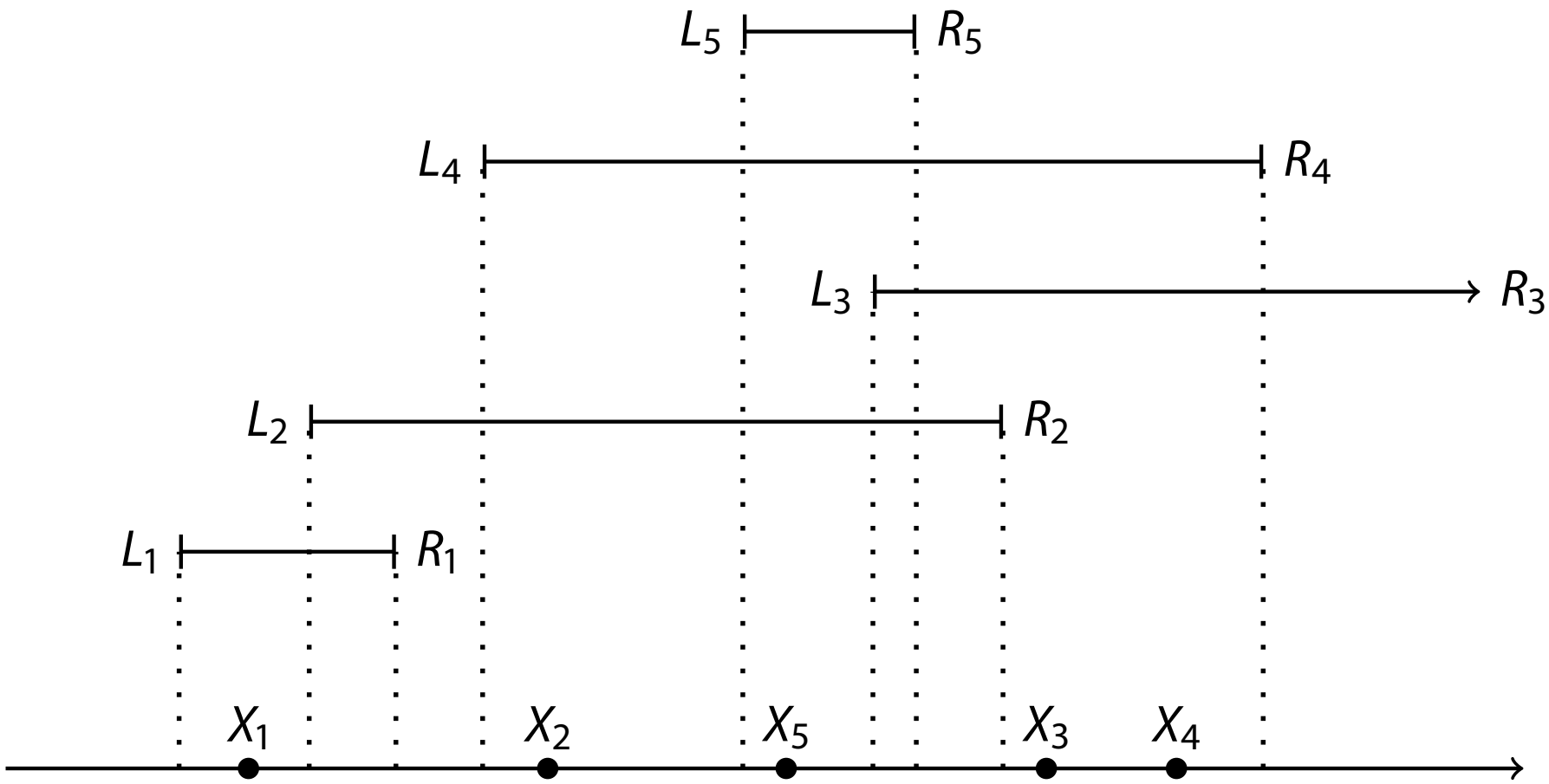
Цензурирование

На практике время наработки на отказ непосредственно почти не наблюдается: данные *цензурированы*.

Проблема: вместо интересуемой случайной величины наблюдается другая, менее информативная.

Для многих реальных задач, связанных с проведением клинических испытаний, характерно т.н. *интервальное цензурирование*: известен некоторый (случайный) интервал, содержащий случайную величину X.

Интервальное цензурирование: пример



Цель

Цели работы:

1. Построение оценок параметров для специальных моделей кривых дожития в условиях интервального цензурирования.
2. Изучение асимптотических свойств полученных оценок теоретически и при помощи моделирования.
3. Разработка средств сравнения параметрических моделей. Модификации информационных критериев типа Акайке на случай интервального цензурирования.
4. Разработка алгоритмов и систем программ, позволяющих производить оценивание параметров при помощи построенных методов.

Публикации

- *Коробейников А. И.* Сравнение оценок параметров специальной модели кривой дожития для выборки с интервальным цензурированием // Вестник СПбГУ, сер. 10. 2009. Т. 2. С. 36–47.
- *Барт А. Г., Коробейников А. И.* Об оценке параметров специальной модели кривой дожития // Математические модели. Теория и приложения. 2007. Т. 8. С. 15–25.
- *Коробейников А. И.* Методы и программное обеспечение задач оценивания параметров в специальном случае модели кривых дожития // Математические модели. Теория и приложения. 2009. Т. 10. С. 28–42.
- *Korobeynikov A.* On the Consistency of ML-estimates for the Special Model of Survival Curves with Incomplete Data // Proc. of 6th St. Petersburg Workshop on Simulation. 2009. Pp. 1039–1045.

Апробация

Конференции:

- 6th Saint Petersburg Workshop on Simulation, Saint Petersburg, June 28 – July 4, 2009.
- 18th Population Approach Group in the Europe (PAGE) Meeting, Saint Petersburg, 23 – 26 June, 2009.
- II Всероссийская научно-практическая конференция с международным участием «Высокотехнологичные методы диагностики и лечения заболеваний сердца, крови и эндокринных органов», Федеральный центр сердца им. В.А. Алмазова, г. Санкт-Петербург, 20 – 22 Мая 2008 г.

Глава 1. Краткое содержание.

В главе 1:

- Построены оценки типа максимума правдоподобия для параметрических моделей в условиях интервального цензурирования смешанного типа.
- Получены достаточные условия состоятельности оценок.
- Исследованы условия идентификации модели.
- Изучены условия асимптотической нормальности оценок.

Известные результаты

Параметрическое оценивание в условиях цензурирования:

- W. Stute, 1992
- P.K. Andersen et al, 1993
- Bickel, Klaassen et al, 1993
- Sun, 2006

Общие проблемы:

- Рассматривается модель случайного правого цензурирования (с ненулевой вероятностью есть полные наблюдения)
- Используются условия регулярности типа Крамера, труднопроверяемые для реальных моделей

Стандартные условия регулярности требуют состоятельности ОМП в качестве одного из условий!

Интервальное цензурирование смешанного типа

- K — количество наблюдений состояния
- T — треугольный массив «возможных моментов наблюдения за состоянием»:
  - $T = \left\{ T_{k,j},\ j = 1, \ldots, k,\ k = 1, \ldots, +\infty \right\}$
  - $0 = T_{k,0} < T_{k,1} < \cdots < T_{k,k} < T_{k,k+1} = +\infty$
- Вектор индикаторов $\Delta_k = (\Delta_{k,1}, \ldots, \Delta_{k,k+1})$ с $\Delta_{k,j} = \mathbb{I}_{(T_{k,j-1}, T_{k,j}]}(X)$
- Наблюдается случайная величина $Y = (K, T_K, \Delta_K)$

Таким образом, Y задает разбиение полуоси $[0; +\infty)$ на $K+1$ (случайный) интервал и указывает интервал, содержащий X

Примеры

- При $K \equiv 1$ приходим к *модели интервального цензурирования первого типа* (Groeneboom, 1992)
  - Обычно представлена как $Y' = (T, \delta)$, где T — момент контроля состояния и $\delta = \mathbb{I}_{[X \leqslant T]}$

- Пример клинических наблюдений (Schick, Yu, 1999):
  - Контроль состояния производится в самом начале
  - Периодическая проверка состояния

$$T_{k,j} = \sum_{i=1}^{j} Z_j, \qquad K = \sup_{j \geqslant 1} \left\{ \sum_{i=1}^{j} Z_i < L \right\},$$

L — максимальная продолжительность наблюдений,

Z_i — промежутки времени между контролем состояния

Распределение наблюдаемой величины

- Параметрическая модель для X с ф.р. $F_\theta \in \{F_\theta,\ \theta \in \Theta\}$
- Условное распределение вектора Δ_K по отношению к (K, T_K) мультиномиально:

$$(\Delta_K \mid K = k, T_K = t_k) \sim \mathrm{Mult}_{k+1}(1, \Delta F_k)$$

где

$$\Delta F_k = (F_\theta(t_{k,1}), F_\theta(t_{k,2}) - F_\theta(t_{k,1}), \ldots, 1 - F_\theta(t_{k,k}))$$

- Распределение Q с.в. Y доминируется некоторой мерой ν, связанной с распределением (K, T_K). Плотность:

$$\frac{dQ}{d\nu} = p_\theta(y) = p_\theta(k, t_k, \delta_k) = \prod_{j=1}^{k+1} \left(F_\theta(t_{k,j}) - F_\theta(t_{k,j-1})\right)^{\delta_{k,j}}$$

Оценки типа максимального правдоподобия

- Пусть $Y_1, \ldots, Y_n$ — n н.о.р. случайных величин с распределением Q.
- $Y_i = \left(K^{(i)}, T_K^{(i)}, \Delta_K^{(i)}\right), i = 1, \ldots, n$
- Логарифм функции правдоподобия для θ:

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta \left(K^{(i)}, T_K^{(i)}, \Delta_K^{(i)}\right),$$

$$m_\theta(k, t_k, \delta_k) = \sum_{j=1}^{k+1} \delta_{k,j} \log \left[F_\theta(t_{k,j}) - F_\theta(t_{k,j-1})\right]$$

- *Приближенная оценка максимального правдоподобия* $\hat{\theta}_n$:

$$l_n(\hat{\theta}_n) - \sup_{\theta \in \Theta} l_n(\theta) \xrightarrow[n \to \infty]{\mathbf{P}} 0$$

Обобщенная состоятельность

Теорема (Обобщенной состоятельности)

Пусть $\mathbf{E}(K) < \infty$ *и функция* $\theta \mapsto F_\theta(x)$ *непрерывна для п.в.* x.

Тогда для любого $\varepsilon > 0$ *и любого компакта* $S \subset \Theta$ *выполняется*

$$\mathbf{P}\left(\operatorname{dist}(\hat{\theta}_n, \Theta_0) > \varepsilon, \hat{\theta}_n \in S\right) \xrightarrow[n\to\infty]{} 0.$$

Кроме того, всегда выполняется

$$\theta \in \Theta_0.$$

Здесь Θ_0 — множество всех точек максимума $\theta \mapsto \int m_\theta \, dQ$.

Идентифицируемость

Введем меру μ:

$$\mu(B) = \sum_{k=1}^{+\infty} \mathbf{P}(K = k) \sum_{j=1}^{k} \mathbf{P}\left(T_{k,j} \in B \,|\, K = k\right)$$

Теорема (Идентифицируемости модели)

Пусть для любых $\theta_1, \theta_2 \in \Theta$ выполняется

$$F_{\theta_1} = F_{\theta_2} \ \mu - п.н. \Rightarrow \theta_1 = \theta_2$$

Тогда в теореме о состоятельности имеем: $\Theta_0 = \{\theta\}$, и в случае компактного Θ имеет место сходимость п.н.:

$$\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta \ п.н.$$

Условия асимптотической нормальности

Теорема

Пусть:

- *функция* $\theta \mapsto \log\left(F_\theta(y) - F_\theta(x)\right)$ *дифференцируема в окрестности* θ *для* $\mu \times \mu$*-п.в.* (x, y), $x < y$ *с* $L^2(Q)$*-суммируемым градиентом;*
- *функция* $\theta \mapsto \int m_\theta \, dQ$ *в точке* θ *допускает разложение по Тейлору до второго члена.*

Тогда:

- *ОМП* $\hat{\theta}_n$ *асимптотически нормальна.*

Метод: теория эмпирических процессов

Глава 2. Краткое содержание.

В главе 2:

- Построены оценки по минимуму расстояния Кульбака-Лейблера для параметрических моделей в условиях интервального цензурирования смешанного типа.
- Получены достаточные условия состоятельности оценок.
- Изучены условия асимптотической нормальности оценок.

Известные результаты

Oakes, 1986: Оценивание на основе ОМКЛ в случае правого цензурирования. Используются условия регулярности типа Cramér'a.

Hjort, 1992: Оценивание на основе ОМКЛ в случае правого цензурирования, в т.ч. случай неточной модели. Используется аппарат считающих процессов и условия регулярности типа Cramér'a.

Suzukawa, 2001: Оценивание на основе ОМП и ОМКЛ в случае правого цензурирования. Сравнение оценок.

Оценки по минимуму расстояния Кульбака-Лейблера

Дивергенция Кульбака-Лейблера:

$$I(f_0, f_\theta) = \int \log \frac{f_0(x)}{f_\theta(x)}\, dF_0(x) = C(f_0) - \int \log f_\theta(x)\, dF_0(x)$$

Свойства:

- $I(f_0, f_\theta) \geqslant 0$.
- $I(f_0, f_\theta) = 0$ тогда и только тогда, когда $f_0(x) = f_\theta(x)$ п.н.

Оценки по минимуму расстояния К-Л:

- Пусть $\hat{F}_n(x)$ — ОМП для $F_0(x)$.
- Тогда определим

$$\tilde{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \hat{I}_n(f_0, f_\theta) = \operatorname*{argmax}_{\theta \in \Theta} \int \log f_\theta(x)\, d\hat{F}_n(x).$$

ОМП, ОМКЛ и цензурирование

ОМП $\hat{\theta}_n$ и ОМКЛ $\tilde{\theta}_n$ совпадают при оценивании по полным данным. $\hat{F}_n$ — эмпирическая функция распределения.

ОМП $\hat{\theta}_n$ и ОМКЛ $\tilde{\theta}_n$ различны в случае цензурирования:

ОМП: Оценивание основывается на наблюдаемой с.в. Y: используется (достаточно сложная) плотность Y относительно (K, T_K) и эмпирическая функция распределения Y.

ОМКЛ: Используется плотность ненаблюдаемой с.в. X и (более сложная) ОМП $\hat{F}_n$.

Оценивание гладких функционалов. Введение.

Необходимое условие состоятельности оценок:

$$\int \log f_\theta(x)\, d\hat{F}_n(x) \xrightarrow[n\to\infty]{\mathbf{P}} \int \log f_\theta(x)\, dF_0(x)$$

Без цензурирования: ЗБЧ

Stute, 1994: Условия состоятельности для гладких функционалов при правом цензурировании.

Groeneboom, 1992: Условия состоятельности при интервальном цензурировании первого типа.

Geskus, 1999: Условия состоятельности при интервальном цензурировании второго типа.

Оценивание гладких функционалов. Асимптотика.

На основе техники из (Geskus, 1999) в диссертации доказана:

Теорема (О состоятельность гладких функционалов)

При наложении некоторых условий регулярности на плотность f_θ и модель интервального цензурирования смешанного типа (K, T_K) гладкие функционалы вида

$$\int \log f_\theta(x)\, d\hat{F}_n(x)$$

асимптотически эффективны:

$$\sqrt{n} \int \log f_\theta(x)\, d\left(\hat{F}_n(x) - F_0(x)\right) \xrightarrow[n\to\infty]{d} N(0, \sigma_0^2).$$

Состоятельность

Теорема (О состоятельности ОМКЛ)

Пусть выполняются условия теоремы о состоятельности гладких функционалов и кроме этого:

1. *Для любого достаточно малого шара* $B \subset \Theta$*:*

$$\int \sup_{\theta \in B} \log f_\theta(x)\, dF_0(x) < \infty$$

2. Θ *— компакт и из* $f_{\theta_1} = f_{\theta_2}$ *п.н. следует* $\theta_1 = \theta_2$*.*

Тогда для ОМКЛ $\tilde{\theta}_n$*:*

$$\tilde{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \hat{I}_n(f_0, f_\theta) \xrightarrow[n \to \infty]{} \theta^* = \operatorname*{argmin}_{\theta \in \Theta} I(f_0, f_\theta) \text{ п.н.}$$

При этом, если найдется θ*:* $f_0 = f_\theta$*, то* $\theta^* = \theta$*.*

Условия асимптотической нормальности ОМКЛ

- В диссертации доказано соотношение:

$$\int \log f_\theta \, d(\hat{F}_n - F_0) = \int c_\theta \, d(Q_n - Q) + o_P(n^{-1/2})$$

для некоторой функции $c_\theta(x)$.

- Например, для интервального цензурирования первого типа с $Y = (T, \mathbb{I}(X \leqslant T))$ имеем:

$$c_\theta(t, \delta) = \frac{\dot{f}_\theta(t)}{f_\theta(t)} \left(-\delta \frac{1 - F_0(t)}{g(t)} + (1 - \delta) \frac{F_0(t)}{g(t)} \right),$$

где $g(t)$ — плотность T.

Соотношение позволяет переписать теорему об асимптотической нормальности из Главы 1 на случай ОМКЛ $\tilde{\theta}_n$

Глава 3: Задача выбора наилучшей модели

- Что «лучше»: ОМП $\hat{\theta}_n$ или ОМКЛ $\tilde{\theta}_n$?
- Как сравнить две параметрические модели?

Информационный критерий (Akaike, 1973): асимптотически несмещенная оценка

$$\int \log f_{\hat{\eta}}(x)\, dF_0(x),$$

где $\hat{\eta}$ — некоторая оценка параметра.

Информационные критерии

Konishi, 1996: GIC: информационный критерий при произвольной процедуре оценивания параметров (AIC, TIC — используют ОМП).

Suzukawa, 2001: Информационные критерии при случайном правом цензурировании.

- Основной прием: замена F_0 на ее оценку $\hat{F}_n$ и рассмотрение величины
$$\int \log f_{\hat{\eta}}(x)\, d\hat{F}_n(x).$$
- Проблема: смещение величины $\int \log f_{\hat{\eta}}(x)\, d\hat{F}_n(x)$ зависит от оценки $\hat{\eta}$ и размерности пространства параметров.

Информационные критерии в случае интервального цензурирования

- В работе предложена оценка смещения

$$d = \int \log f_{\hat{\eta}} \, d\left(\hat{F}_n - F_0\right)$$

при оценивании посредством:

  - $\hat{\eta} = \hat{\theta}_n$
  - $\hat{\eta} = \tilde{\theta}_n$

- В случае интервального цензурирования первого типа получено явное выражение для смещения d
- В остальных случаях отдельный «фрагмент» смещения оценивается при помощи процедур «складного ножа» (jack-knife) и бутстрепа (bootstrap).

Глава 4: Моделирование

Модели кривых дожития:

- Распределения, связанные с экспоненциальным (Вейбулла, гамма, степенное гамма)
- Модель Гомперца-Макегама
- Модель ExpCos А.Г. Барта

Оценки:

ОМП: $\hat{\theta}_n$

ОМКЛ: $\tilde{\theta}_n$

Глава 4: Моделирование

Механизм цензурирования типичен для клинических испытаний:

$$T_{k,j} = \sum_{i=1}^{j} Z_j, \qquad K = \sup_{j \geqslant 1} \left\{ \sum_{i=1}^{j} Z_i < L \right\},$$

L — максимальная продолжительность наблюдений,
Z_i — промежутки времени между контролем состояния

Глава 4: Моделирование

Механизм цензурирования типичен для клинических испытаний:

$$T_{k,j} = \sum_{i=1}^{j} Z_j, \qquad K = \sup_{j \geqslant 1} \left\{ \sum_{i=1}^{j} Z_i < L \right\},$$

L — максимальная продолжительность наблюдений,
Z_i — промежутки времени между контролем состояния

Исследуемые свойства оценок:

- Скорость сходимости к предельному распределению
- Дисперсия и среднеквадратичное отклонение
- Устойчивость к неточной модели

Моделирование: общие результаты

- Параметрическая модель $\{f_\theta, \theta \in \Theta\}$ точна: ОМП $\hat{\theta}_n$ в большинстве случаев лучше ОМКЛ $\tilde{\theta}_n$ как в терминах среднеквадратичного отклонения, так и дисперсии.
- Предполагаемая модель неточна: ОМКЛ $\tilde{\theta}_n$ обладает меньшей дисперсией и среднеквадратическим отклонением, чем ОМП $\hat{\theta}_n$.
  - Чем больше «уровень цензурирования», тем больше отклонение.
- Чем «сложнее» модель, тем меньше отличия между ОМП $\hat{\theta}_n$ и ОМКЛ $\tilde{\theta}_n$.

Глава 5: Анализ реальных данных

Предложенные оценки параметров используются для анализа реальных данных из

Стоматологии (Мадай, Барт, Коробейников и др., 2006):

- Сравнение двух способов лечения

Кардиологии (Бондаренко, Алексеева, Коробейников, 2008):

- Оценка длительности критического периода артериальной гипертензии в разных группах

Фармакологии (Krupitsky, Verbitskaya et all, 2004):

- Выбор наилучшей параметрической модели и сравнение типов лечения в разных группах.

Положения диссертации, выносимые на защиту

- два класса оценок для параметрических моделей в условиях интервального цензурирования смешанного типа
- достаточные условия состоятельности и асимптотической нормальности предложенных оценок параметров в условиях неполных данных
- информационные критерии типа Акайке для сравнения различных параметрических моделей в случае интервального цензурирования смешанного типа
- программный комплекс, реализующий предложенные способы оценивания параметров