Обзор математических моделей и методов анализа данных о выживаемости

Прокудина Е.И.
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
preliv@gmail.com

Абсаттарова Э.Э.
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
elina.absattarova@yandex.ru

Аннотация[1]

В работе рассмотрены математические модели и методы анализа данных типа продолжительности жизни.

1. Введение

Задача моделирования и анализа продолжительности жизни человека рассматривается в разных областях: в медицине, демографии, страховании, социологии. В частности, задача моделирования продолжительности жизни с учетом различных факторов является одной из подзадач оценки рисков в гериатрии.

Объектом конкретного исследования может быть продолжительность жизни, как новорожденного, так и человека, дожившего до определенного возраста, как здорового, так и имеющего определенное заболевание, проживающего в определенном регионе и т.д.

Постановка задачи анализа выживаемости может быть разной в зависимости от целей исследования, но, как правило, исходными данными являются статистические данные, содержащие информацию о временах жизни группы людей на заданном промежутке времени, отобранных в группу по определенному признаку или совокупности признаков. Необходимо определить показатели, характеризующие продолжительность жизни, и факторы, наиболее влияющие на эти показатели.

Литература, посвященная данной тематике, является достаточно обширной. Рассмотрим основные математические модели и методы анализа данных о выживаемости.

2. Математические модели и методы

Предполагается, что продолжительность жизни человека в возрасте x лет – $T(x)$ (время до определенного события в общем случае) является непрерывной случайной величиной.

В качестве показателей выживаемости используются функциональные и числовые характеристики данной случайной величины: функция распределения $F_x(t)$, функция выживания $s_x(t)$ – вероятность того, что человек в возрасте x проживет еще t лет, кривая смертей $f_x(t)$ – плотность распределения случайной величины, интенсивность смертности или функция риска $\mu_x(t) = f_x(t)/s_x(t)$, математическое ожидание, дисперсия, медиана, мода и др., а также статистические оценки вероятностных характеристик.

Методы решения задачи анализа выживаемости можно разделить на параметрические, непараметрические и полупараметрические [1].

К параметрическим методам относится, в частности, применение аналитических моделей для описания продолжительности жизни как случайной величины.

Первой по времени появления из аналитических моделей смертности является модель де Муавра. В данной модели продолжительность жизни новорожденного равномерно распределена на промежутке $[0,\omega]$, где ω – предельная продолжительность жизни. При этом остаточное время жизни равномерно распределено на $[0,\omega - x]$. Модель применяется на небольших временных отрезках, поскольку плохо описывает эмпирические данные [2, 3].

В модели Гомпертца интенсивность смертности экспоненциально зависит от возраста:

$$\mu_x = Be^{cx},$$

где B, c – положительные константы.

Используется для описания смертности в защищённых средах, где внешние причины смерти отсутствуют (в лабораторных условиях, в зоопарках для большинства многоплодных животных или для людей в развитых странах) [4]

Труды Седьмой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-30 мая, Уфа-Ставрополь, Ханты-Мансийск, Россия, 2019

В первой модели Мэйкхама уточняется модель Гомпертца: к компоненту, описывающему смертность от естественных причин, прибавляется компонент, независящий от возраста и описывающий смертность в результате несчастного случая:

$$\mu_x = A + Be^{cx},$$

где A, B, c – положительные константы [3, 4].

Во второй модели Мэйкхама предполагается, что вероятность насильственной смерти меняется с возрастом по линейному закону:

$$\mu_x = A + Dx + Be^{cx}.$$

Модель Вейбулла представляет собой степенную зависимость

$$\mu_x = kx^s,$$

где k, s – положительные константы [3,4].

Также для описания продолжительности жизни применяются экспоненциальное, гамма, логнормальное и другие распределения [5].

Параметры в распределениях подбираются в основном с помощью методов максимального правдоподобия и наименьших квадратов. В [5] представлены формулы для вычисления параметров различных распределений, полученные с помощью метода максимального правдоподобия.

Для оценки качества моделей, а также сравнения между собой различных моделей, применяются информационные критерии, например, Акаике, Шварца [5].

К параметрическим подходам моделирования продолжительности жизни и прогнозирования смертности можно отнести использование статистических моделей временных рядов, таких как однофакторая модель Ли и Картера, многофакторные модели с эффектом когорты Реншоу-Хабермана, Кернса-Блейка-Дауда [6, 7]. В [7] показано преимущество модели Реншоу-Хабермана для российских данных.

Методы построения таблиц смертности относятся к непараметрическим методам анализа данных о выживаемости. Таблица смертности представляет собой систему взаимосвязанных, упорядоченных по возрасту рядов чисел, описывающих процесс вымирания некоторого поколения с фиксированной начальной численностью населения [8, 9, 10].

При построении таблиц смертности без учета причин смертности применяются метод смертных списков и демографический метод.

Метод смертных списков используется при отсутствии данных о возрастной структуре населения и основан на следующих допущениях: медленное изменение чисел рождений, медленное изменение возрастной смертности, закрытое население. В качестве исходных статистических данных используются данные об умерших, сгруппированные по возрастам. Метод базируется на предположении стационарного населения (общее число умерших равно родившимся), что является его основным недостатком [11, 12].

Демографический метод предполагает наличие не только данных о распределении умерших по возрастам, но и данных о возрастной структуре населения. Исходным показателем при расчете таблиц с помощью этого метода служит возрастной коэффициент смертности [8].

С учетом причин смертности обычно выполняют построение кратких таблиц смертности (в отличие от полных строятся не для однолетних, а для пяти- или десятилетних возрастных групп) в связи с тем, что число смертных случаев по ряду причин в отдельных возрастных группах может быть незначительным. Расчет показателей таблиц смертности в данном случае основан на предположении, что интенсивность смертности от некоторой причины в данном возрасте не зависит от смертности по другим причинам в более молодых возрастах. В таблице смертности по одной из причин вместо ожидаемой продолжительности жизни можно рассматривать средний возраст смерти от изучаемой причины, при условии наличия единственной причины смерти [9, 13].

Прямой метод построения таблицы смертности (метод Лапласа) основан на простом соотношении числа больных, переживших контрольный срок, и числа больных, взятых под наблюдение в начале исследования [12]. Возможности этого метода ограничены получением показателей выживаемости лишь при условии, что все больные были взяты под наблюдение не позже момента, от которого к окончанию исследования должно пройти необходимое для определения полного времени выживания число лет.

В настоящее время широко применяется в медицинской практике динамический метод расчета показателей выживаемости для таблиц смертности. Главным его достоинством является возможность использования всей информации, которая имеется в распоряжении исследователя, например, в группу, для которой определяется показатель n-летней выживаемости, могут включаться и те больные, которые были взяты под наблюдение менее n лет назад, также он позволяет рассчитать показатели внутригодовой выживаемости [14].

Так как таблицы смертности строятся на основе статистических данных, которые могут содержать случайные ошибки, то к таблицам применяется процедура сглаживания. Например, на основе аналитических моделей смертности, подбирая параметры выбранной функции так, чтобы минимизировать разницу между сглаженной и первичной таблицами. Обычно для этого используется модель Мейкхема. Для улучшения

качества сглаживания данную модель можно модифицировать, предполагая полиномиальную зависимость смертности для старших возрастов и склеивая ее с экспонентой на остальном временном интервале. Чтобы получилась гладкая кривая, значения производных двух функций в точках склеивания должны быть равны [15].

Самым простым методом локального сглаживания является «фильтр 1/*n*» – простое локальное усреднение исходных вероятностей, где *n* – количество соседних точек, участвующих в усреднении. Как правило, данный параметр выбирают нечетным числом, и усреднение проводится симметрично относительно выбранной исходной точки. В случае если исходные данные содержат значимую статистическую погрешность, то метод приводит к эффекту, известному как “over smoothing” – сглаженный график имеет много перегибов, чтобы «успеть» за исходным [15].

В [15] также рассматривается метод ядерного сглаживания Надарая-Ватсона, отмечается, что использование нормального (гауссовского) ядра, ядра Епанечникова и трикубического ядра для сглаживания таблиц смертности дают примерно одинаковые результаты.

К непараметрическим методам анализа данных о выживаемости относят также широко применяемый на практике метод Каплана-Мейера, с помощью которого на основе неполных данных строится эмпирическая функция выживания [16].

Корректность применения метода рассматривается в работах [17, 18]. Подчеркивается, что для получения фактов, которые могут быть использованы в доказательных целях, необходимы проспективные контролируемые исследования.

В [19] для построения кривых выживаемости наряду с методом Каплана–Мейера используется метод Катлера–Эдерера, применяемый при наличии выборки большого объема и учитывающий цензурированные данные.

В [20] проведено сравнение метода Каплана–Мейера с методом конкурирующих рисков, который учитывает влияние на терминальное событие существования многих вариантов неблагоприятных исходов, подчеркивается преимущество метода конкурирующих рисков.

К полупараметрическим методам анализа данных о выживаемости относят широко применяемые на практике регрессионный анализ Кокса и регрессионный анализ Кокса с зависящими от времени предикторами, представляющие собой прогнозирование риска наступления события для рассматриваемого объекта и оценку влияния определенных независимых переменных на этот риск. В [21] на практическом примере описаны принципы проведения анализа пропорциональных рисков Кокса с помощью пакета прикладных статистических программ SPSS.

3. Заключение

Выполнен обзор математических методов анализа данных о выживаемости. Рассмотрены параметрические, непараметрические и полупараметрические методы. Отмечены особенности их применения.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00780.

Список используемых источников

1. Gardiner J.C. Survival Analysis: Overview of parametric, nonparametric and semiparametric approaches and new developments // SAS Global forum. – 2010. – 252–2010. –23 P. URL: https://support.sas.com/resources/papers/proceedings10/252-2010.pdf (дата обращения: 01.03.2019).
2. Life Contingencies (Part A2) / W.F. Scott – Edinburg: Department of Actuarial Mathematics and Statistics, Herriot-Watt University, 1996.
3. Основы актуарной математики / Е.М. Бронштейн, Е.И. Прокудина – Уфа: УГАТУ, 2012. – 315 с. URL:http://e-library.ufa-rb.ru/dl/lib_net_r/Bronshtein Osn akt mat-ki 2012.pdf (дата обращения: 03.04.2019).
4. Халявкин А.В., Крутько В.Н. // Подход к моделированию старения с позиций биофизики сложных систем. Труды ИСА РАН 2006. Т. 19 С.117-155.
5. Гайдышев И.П. Подгонка распределений в параметрическом анализе выживаемости. // Вестн. Ом. ун-та. 2016. №4. С. 10-15.
6. Binder G. Construction and Comparison of Mortality Tables Based on Different Techniques. Master Thesis. – ETH Zurich, 2014. – 132 p.
7. Миронкина Ю.Н., Гусева В.И. Моделирование смертности в России с помощью актуарных стохастических моделей // Применение многомерного статистического анализа в экономике и оценке качества: труды XI Международной конференции, ЦЭМИ РАН. М.: ЦЭМИ РАН, 2018. С. 108-110.
8. Демографические таблицы / Л.Е. Дарский, М.С. Тольц – М.: МАКС Пресс, 2013. – 104 с.
9. Демография / Медков В.М. – Ростов-на-Дону: Феникс, 2002. – 448 с.
10. Математические основания геронтологии / В.Н. Крутько, М.Б. Славин, Т.М. Смирнова – М.: Едиториал УРСС, 2002. – 384 с.
11. Смертность и продолжительность жизни населения СССР 1926-1927гг. Таблицы

смертности. – Москва-Ленинград: ПЛАНХОЗГИЗ, 1930. – 138 с.

12. Овчарова Л.Н. Построение и анализ таблиц смертности на основе региональных данных // Вестник Ростовского государственного экономического университета «РИНХ». 2009. № 2. с. 264-273.
13. Методика расчета таблиц дожития с учетом влияния отдельных видов причин смертности на ожидаемую продолжительность жизни. Приложение к приказу исполняющего обязанности Председателя Агентства Республики Казахстан по статистике от 29 декабря 2011 года № 386. – Астана, 2011. – URL: https://online.zakon.kz/Document/?doc_id=31191078#pos=0;146
14. Общая онкология: Руководство для врачей / [Н.П. Напалков и др.]; Под ред. Н.П. Напалкова. – Л.: Медицина. Ленингр. отд-ние, 1989. – 646 с.
15. Костенко Л., Хасанов Р. Сглаживание таблиц смертности: подбор функциональной зависимости и гладкое локальное сглаживание. 6 с. URL: http://www.actuary.ru/upload/medialibrary/5be/5be53593203865a7ee017fd423b0658c.pdf
16. Рапаков Г.Г., Горбунов В.А. Исследование методов анализа времени до события при обработке демографических данных // Вестник ВГУ, Серия: Системный анализ и информационные технологии. 2015. № 4 С. 110-120.
17. Куликов С.М., Паровичникова Е Н., Савченко В.Г. Анализ выживаемости или событийный анализ: типовые ошибки ретроспективного метода // Клиническая онкогематология. 2010. Т. 3. № 2. С.176-183.
18. Жуков Н.В. Как организовать собственные исследования и не запутаться в интерпретации чужих Исследования III фазы (часть 3) // Вместе против рака. 2007. № 3-4. С.34-44.
19. Буре В.М., Парилина Е.М., Рубша А. И., Свиркина Л. В. Анализ выживаемости по медицинской базе данных больных раком предстательной железы // Вестник СПбГУ. Сер. 10. 2014. Вып. 2. С. 27-35
20. Слинин А.С., Быданов О.И., Карачунский А.И. Анализ выживаемости и вероятности возникновения отдельных событий у пациентов с острым лейкозом // Вопросы гематологии/онкологии и иммунопатологии в педиатрии. 2016. Т. 15. №3. С. 34–39.
21. Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А. М. Применение регрессии Кокса в здравоохранении с использованием пакета статистических программ SPSS // Наука и Здравоохранение. 2017. № 6. С. 5-27.