Технічні науки

УДК 51-74

Мулява Ольга Ярославівна
студент
«Інститут Прикладного Системного Аналізу»
Національний Технічний Університет України
«КПІ ім. Ігоря Сікорського»
Мулява Ольга Ярославовна
студент
«Институт Прикладного Системного Анализа»
Национальный Технический Университет Украины
«КПИ им. Игоря Сикорского»
Olha Muliava
student
«Institute for Applied Systems Analysis»
National Technical University of Ukraine
«Igor Sikorsky KPI»

МАТЕМАТИЧНІ МОДЕЛІ В АНАЛІЗІ ВИЖИВАННЯ
МАТЕМАТИЧЕСКИЕ МОДЕЛИ В АНАЛИЗЕ ВЫЖИВАНИЯ
MATHEMATICAL MODELS IN ANALYSIS OF SURVIVAL

Анотація: У роботі наведено аналіз і порівняння моделей, які використовуються в аналізі виживання, зокрема модель пропорційних ризиків та модель прискорення часу відмови. Також наведено обґрунтування використання методів повторних вибірок в аналізі виживання.

Ключові слова: аналіз вживання, модель прискорення часу відмови, модель пропорційних ризиків.

Аннотация: В работе приведен анализ и сравнение моделей, используемых в анализе выживания, в частности модель пропорциональных рисков и модель ускорения времени отказа. Также приведено обоснование использования методов повторных выборок в анализе выживания.

Ключевые слова: анализ выживаемости, модель ускорения времени отказа, модель пропорциональных рисков.

Summary: The paper presents an analysis and comparison of the models used in the survival analysis, in proportional hazards model and accelerated failure time model. Also, the justification of the use of repeat sample methods in the survival analysis is presented.

Key words: survival analysis, accelerated failure time model, proportional hazards model.

Існує три підходи до побудови моделей в аналізі виживання: параметричний, напівпараметричний і непараметричний.

Параметричний підхід є найбільш прямим. За цим підходом ми припускаємо, що базовий ризик $\lambda(t)$ має особливу функціональну форму. Наприклад, моделі на основі розподілу Вейбулла, експоненційного та гамма.

В напівпараметричному підході на базовий ризик $\lambda(t)$ накладаються не такі строгі припущення. Точніше, час ділиться на малі інтервали і припускається, що функція базового ризику є константою на кожному з цих інтервалів, призводячи до кусочно-лінійної експоненційної моделі.

Непараметричний підхід концентрується на оцінюванні параметрів регресії, залишаючи функцію базового ризику повністю невизначеною. Цей підхід спирається на часткову правдоподібність.

Два найбільш поширених підходи, якими можна моделювати залежність між часом до події і факторами, що можуть вплинути на нього, відомі як модель пропорційних ризиків і модель прискорення часу відмови.

Модель пропорційних ризиків

1.1 Опис та інтерпретація моделі

Модель пропорційних ризиків - модель, яка використовується в аналізі виживання, яка може бути використана для оцінки значущості різних коваріат на тривалість життя людей або об'єктів за допомогою функції ризику. Крім того, може бути описаний кількісний вплив цих змінних протягом життя на важливі змінні результату.

Модель пропорційних ризиків є напівпараметричною в тому сенсі, що не робиться ніяких припущень про базову фунцію ризику.

Першим і найважливішим обмеженням є неінформативне цензурування. Для задоволення цього обмеження, дизайн експерименту не допускає, щоб випадки цензурування були пов'язані з ймовірністю виникнення події.

Другим припущенням цієї моделі є припущення пропорційних ризиків. Це означає, що виживання двох індивідів повинні мати функції ризику, що зберігають ту ж саму пропорцію з часом.

Модель пропорційних ризиків має наступний вигляд [1]:

$$\lambda_i(t) = \lambda(t)\exp(x_i^T\beta) \qquad (1.1)$$

Тут, коваріати діють мультиплікативно на функцію загрози. Потрібно взяти до уваги, що експонента забезпечує те, що $\lambda_i(t)$ завжди має додатнє значення. В цій моделі, функція базового ризику для і-го

суб'єкта завжди має однакову загальну форму $\lambda(t)$, проте може, наприклад, подвоїтись або зменшитись вдвоє в залежності від факторів ризику, що впливають на суб'єкт.

Розглянемо випадок постійного базового ризику. В такому випадку, отримуємо модель експоненційної регресії:

$$\lambda_i(t) = \lambda \exp(x_i^T \beta) \tag{1.2}$$

Відмітимо, що якщо x_i має вільний член, з'явиться проблема ідентифікованості – неможливо обчислити λ і β_0. З ряду причин переважно краще оцінювати β_0 ніж λ, тому будемо використовувати цю параметризацію. Звичайно, отримавши оцінку β_0, можна легко оцінити значення та довірчі інтервали для λ через перетворення $\lambda = \exp(\beta_0)$.

Оцінка β за методом максимальної правдоподібності є складною у випадку експоненційної регресії через те, що потрібно розв'язувати нелінійну систему рівнянь. Воно не може бути розв'язане у явній формі, для розв'язання потрібна ітеративна процедура.

Основною ідеєю є створити лінійне наближення до нелінійної системи рівнянь, обчислити з неї $\hat{\beta}$, обчислити нове наближення і повторювати до збіжності. Цей метод відомий як алгоритм Ньютона-Рафсона (Рис. 2.1).

Розглянемо гіпотетичне порівняння між двома індивідами, чиї пре диктори є однаковими окрім змінної j, де вони відрізняються на $\delta_j = x_{1j} - x_{2j}$. Тоді:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \exp(\delta_i \beta_i)\,. \tag{1.3}$$

Для будь-якої моделі пропорційних ризиків, $\lambda_1(t)/\lambda_2(t)$ є константою, незмінною з часом. Ця константа відома як відношення

ризиків, або відносний ризик. Таким чином, інтерпретацією коефіцієнта регресії в моделі пропорційних ризиків є те, що $e^{\delta\beta}$ є відношення ризиків для зміни в δодиниць в коваріаті.

1.2 Оцінювання параметрів в моделі пропорційних ризиків

Позначимо лінійний предиктор $\eta_i = x_i^T \beta$.

Враховуючи незалежне цензурування і вважаючи, що реальний час життя $\tilde{T}_i | x_i \sim Exp(\lambda_i)$, внесок логарифма правдоподібності і-го суб'єкта в експоненційну регресію наступний [1]:

$$l_i(\eta_i) = d_i\eta_i - t_i e^{\eta_i}, \tag{1.4}$$

де $d_i = 1\{\tilde{t}_i \leq c_i\}$, $t_i = \tilde{t}_i \wedge c_i = \min(\tilde{t}_i, c_i)$, c_i – час цензурування і-го суб'єкта, $\tilde{t}_i$ – реальний час життя і-го суб'єкта.

Тоді похідна логарифму правдоподібності і гессіан мають наступний вигляд:

$$u_i(\eta_i) = d_i - t_i e^{\eta_i} \tag{1.5}$$

$$H_i(\eta_i) = -t_i e^{\eta_i} \tag{1.6}$$

Нехай μ позначає вектор, у якого і-й елемент дорівнює $t_i e^{\eta_i}$ і W позначає діагональну матрицю з і-м діагональним елементом $t_i e^{\eta_i}$. Тоді можна переписати похідну логарифму правдоподібності і гессіан у вигляді:

$$u(\eta) = d - \mu \tag{1.7}$$

$$H(\eta) = -W \tag{1.8}$$

Розглянемо наближення у вигляді ряду Тейлора

$$u(\eta) \approx u(\tilde{\eta}) + H(\tilde{\eta})(\eta - \tilde{\eta}) = d - \mu + W(\tilde{\eta} - \eta) \quad (1.9)$$

де μ і W фіксовані на $\tilde{\eta}$.

Підставляючи $\eta = X\beta$ у попереднє рівняння і розв'язуючи для β, отримуємо

$$\hat{\beta} \leftarrow (X^T W X)^{-1} X^T (d - \mu) + \tilde{\beta} \quad (1.10)$$

Так як маємо ітеративну процедуру, ми не отримуємо точне значення $\hat{\beta}$. Натомість, потрібно обчислити значення $\hat{\beta}$, перерахувати μ і W і повторювати до збіжності.

Алгоритм Ньютона-Рафсона збігається до оцінки максимальної правдоподібності, так як правдоподібність логарифмічно увігнута і зростаюча, і це також вірно (зазвичай) для експоненційної регресії.

Алгоритм у вигляді блок-схеми зображений на рис.2.1.

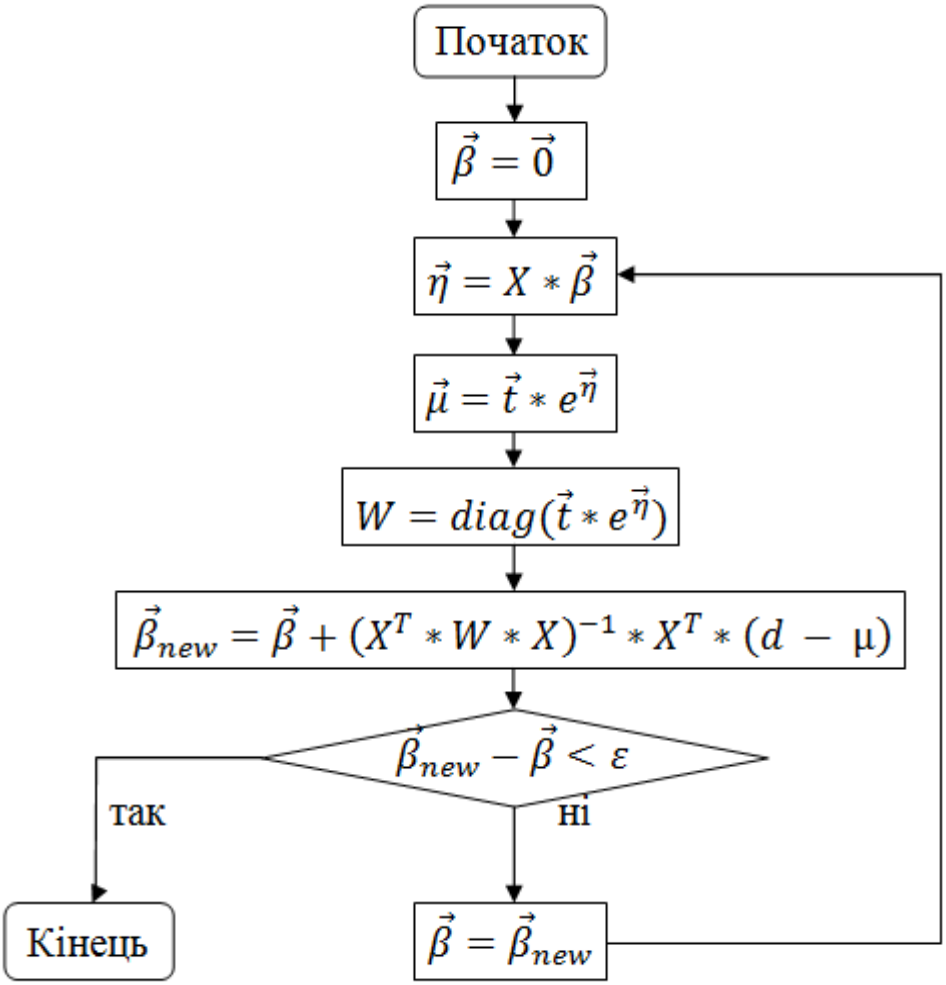

Рисунок 1.1 - Блок-схема алгоритму Ньютона-Рафсона

Існує три підходи до отримання довірчих інтервалів параметрів моделей виживання: заснований на похідній логарифму правдоподібності, метод Вальда і метод відношення правдоподібностей [1].

У методі, заснованому на похідній логарифму правдоподібності, перевіряється гіпотеза $H_0 : \theta = \theta_0$. Для цього обчислюється статистика

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \tag{1.11}$$

і її значення порівнюється з нормальним розподілом. Як і звичайно, перетворюючи цей тест до $\alpha = 0.05$, можна отримати 95% довірчі

інтервали для θ. Відмітимо, що для цього підходу, на відміну від двох наступних, не потрібно оцінювати значення θ.

Метод Вальда заснований на наближенні до похідної логарифму правдоподібності ряду Тейлора оцінки максимальної правдоподібності (H – матриця Гессе):

$$u(\theta) \approx H(\hat{\theta})(\theta - \hat{\theta}) \tag{1.12}$$

Таким чином,

$$\mathcal{I}^{\frac{1}{2}}(\hat{\theta} - \theta_0) \sim N(0,1), \text{ або } \hat{\theta} \sim N(\theta_0, \mathcal{I}^{-1}) \tag{1.13}$$

Тоді оцінка максимальної правдоподібності наближено нормальна, має середнє значення рівне справжньому значенню параметра і варіацію рівну оберненому значенню від інформації.

На основі цього результату, можна легко отримати тести або довірчі інтервали для θ.

Також розглянемо асимптотичний розподіл відношення правдоподібностей. Цей підхід також застосовує розклад в ряд Тейлора, але тут наближується логарифм правдоподібності.

$$l(\theta) \approx l(\tilde{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\hat{\theta})(\theta - \hat{\theta}) \tag{1.14}$$

Таким чином,

$$2\{l(\hat{\theta}) - l(\theta_0)\} \sim \chi_p^2 \tag{1.15}$$

Отримана за алгоритмом Ньютона-Рафсона оцінка $\hat{\beta}$ є оцінкою максимальної правдоподібності, за методом Вальда $\hat{\beta} \sim N(\beta, I^{-1})$. Залишилось лише отримати матрицю інформації відносно β:

$$\hat{\beta} \sim N(\beta, (X^T W X)^{-1}) \tag{1.16}$$

Звідси отримуємо довірчі інтервали для β_j:

$$\beta_j \pm z_{1-\frac{\alpha}{2}} SE_j \tag{1.17}$$

$$SE_j = \sqrt{(X^T W X)^{-1}_{jj}} \tag{1.18}$$

Підхід відношення правдоподібностей на практиці дещо ускладнюється у випадку декількох параметрів, оскільки ми не маємо оцінок параметрів у явних формах. Якби деякий параметр β_j був єдиним параметром, проблема би зводилась до знаходження коренів, де ми б обчислили довірчий інтервал за правилом $2\left(l(\hat{\beta}_j) - l(\beta_j)\right) = \chi^2_{1,.95}$.

Проте, β_j не є єдиним параметром, тому всі оцінки максимальної правдоподібності відповідно зміняться. Іншими словами, оцінювання $l(\beta_j)$ не є простим, тому що потрібно перераховувати $\hat{\beta}_{-j}$ для кожного значення β_j для якого ми намагаємось знайти корені.

Правдоподібність $L(\beta_j, \hat{\beta}_{-j}(\beta_j))$ називається профілюючою правдоподібністю, а процедура перерахування називається профілюванням.

Отримання довірчих інтервалів за допомогою похідної логарифму правдоподібності або відношення правдоподібностей вимагають профілювання, на відміну від підходу Вальда. На практиці, швидше і зручніше використовувати довірчі інтервали Вальда [1].

Модель прискорення часу відмови

2.1 Опис та інтерпретація моделі

Для випадкової величини часу до події T, модель прискорення часу відмови [1] пропонує наступне відношення між коваріатами та $Y = \log T$:

$$Y_i = x_i^T \beta + W_i, \tag{1.19}$$

де $W_i \sim^{iid} f$ є похибкою (або залишком). Такі моделі часто називаються лог-лінійними моделями. Вищенаведені позначення описують широкий клас моделей: в залежності від розподілу, який буде вказано для W, будуть отримані різні моделі, але всі вони будуть мати схожу загальну структуру.

Очевидно, можна припустити, що $W_i \sim^{iid} N(0, \sigma^2)$. Припущення про те, що Y має нормальний розподіл, еквівалентне припущенню, що T має логнормальний розподіл. Тоді, за відсутності цензурування, можна просто використати звичайний метод найменших квадратів, щоб побудувати модель, отримати довірчі інтервали, тощо. Але звичайно, майже завжди у дослідженнях присутнє цензурування, тому потрібно розширити звичайні лінійні методи так, щоб вони справлялись з цензуруванням.

Для будь-якої моделі прискорення часу відмови, маємо:

$$T = e^{\eta_i} T_0, \tag{1.20}$$

де $T_0 = e^W$ і $\eta_i = x_i^T \beta$.

Іншими словами, в той час як в моделі пропорційних ризиків коваріати впливають мультиплікативно на ризик, в моделі прискорення часу відмови коваріати впливають мультиплікативно на час до виникнення події.

Функція виживання: $S_i(t) = S_0(e^{-\eta_i} t)$.

Функція ризику: $\lambda_i(t) = \lambda_0(e^{-\eta_i}t)e^{-\eta_i}$.

Якщо порівнювати модель пропорційних ризиків і модель прискорення часу відмови на графіку логарифм часу проти логарифму ризику, ефект припущення пропорційних ризиків впливає на зміну ризику і спричиняє вертикальне зміщення, в той час як ефект моделі прискорення часу відмови спричиняє горизонтальний зсув. Дві моделі не можуть бути приведені одна до одної, тому процес може бути краще описаний лише однією з моделей, а не обома одночасно [1].

Проте, існує виключення: якщо розподіл є лінійним (на шкалі логарифму часу проти логарифму ризику), тоді будь-якому горизонтальному зміщенню лінії можна поставити у відповідність вертикальне. У цьому масштабі є лінійним розподіл екстремальних значень, а також сімейство розподілів Вейбулла. Тому, розподіл Вейбулла є єдиним розподілом, який задовольняє обидві моделі: пропорційних ризиків і прискорення часу відмови [1].

Згідно з вищенаведеним, вплив зміни на δ_j одиниць в коваріаті j призводить до збільшення часу відмови у $\exp(\delta_i\beta_i)$ разів.

2.2 Оцінювання параметрів моделі прискорення часу відмови

Запишемо модель прискорення часу відмови у наступному вигляді [1]:

$$Y_i = x_i^T\beta + \sigma W_i. \tag{1.21}$$

Відмітимо, що так можуть бути записані моделі Вейбулла і логнормальні. Тоді правдоподібність є наступною:

$$L(\beta, \sigma|y, d) = \prod_i \{\sigma^{-1}f(w_i)\}^{d_i}\{S(w_i)\}^{1-d_i} = \tag{1.22}$$

$$= \prod_i \{\sigma^{-1}\lambda(w_i)\}^{d_i} S(w_i),$$

де f, λ та S – функції щільності, ризику та виживання відповідно, і $w_i = (y_i - x_i^T \beta)/\sigma$.

Загалом, процес оцінки параметрів відбувається як і у попередньому випадку, за виключенням того, що тепер також необхідно оцінювати параметр σ.

Для прикладу:

$$\frac{\partial l}{\partial \beta} = -\sigma^{-1} X^T a, \tag{1.23}$$

$$\frac{\partial l}{\partial \sigma} = -\sigma^{-1}(d + w_i a), \tag{1.24}$$

де $d = \sum_i d_i$ і $a_i = \partial l_i / \partial w_i$. Якщо W має розподіл екстремальних значень, то $a_i = d_i - e^{w_i}$. Взагалі кажучи, a є нелінійною функцією від β і σ, що значить, що для того, щоб розв'язати $u(\beta, \sigma) = 0$ (u - похідна логарифму правдоподібності), потрібно знову використовувати алгоритм Ньютона-Рафсона.

Ускладненням є те, що є можливість отримати неможливі значення σ, наприклад, негативні значення. Для того, щоб уникнути цієї особливості, можна оцінювати σ більш плавно:

$$\sigma_{m+1} = (1 - \tau)\sigma_m + \tau\tilde{\sigma}, \tag{1.25}$$

де $\tilde{\sigma}$ - крок Ньютона-Рафсона.

Отримання довірчих інтервалів для даної моделі відбувається за тим самим алгоритмом, як і для моделі пропорційних ризиків.

Для побудови довірчих інтервалів потрібна (обернена) матриця інформації:

$$I(\hat{\theta}) = -\begin{bmatrix} \frac{\partial^2 l}{\partial \beta^2} & \frac{\partial^2 l}{\partial \sigma \partial \beta} \\ \frac{\partial^2 l}{\partial \beta \partial \sigma} & \frac{\partial^2 l}{\partial \beta^2} \end{bmatrix} \quad (1.26)$$

Потрібно зазначити, що підхід Вальда є більш точним, якщо правдоподібність репараметризувати в термінах $\tau = \log \sigma$.

Використання методів повторних вибірок в оцінюванні параметрів моделей аналізу виживання

У медицині часто мають справу з малим розміром вибірки. Для цього є декілька причин [2]. Часто дані є цензурованими. Тому, через малі розміри вибірок не можна використовувати класичні статистичні методи, оскільки вони можуть дати занадто загальний або навіть помилковий результат.

За допомогою комп'ютерної симуляції можна отримати багато вибірок на основі початкової вибірки і таким чином більш точно оцінити параметри за допомогою отриманого розподілу оцінок.

Класична побудова довірчих інтервалів на основі асимптотичної нормальності (підхід Вальда) часто призводить до помилкових висновків при роботі з цензурованими даними, особливо в невеликих вибірках. Альтернативні методи дозволяють побудувати оцінку довірчого інтервалу, не покладаючись на ці припущення [3].

Також в деяких дослідженнях [4] методи повторних вибірок використовують для дослідження стійкості факторів і вибору, які з них потрібно включати в кінцеву модель. Комбіноване застосування моделі аналізу виживання і методів повторної вибірки допомагає будувати більш

точні моделі у випадку, коли досліджувані суб'єкти дуже відрізняються між собою.

Використання методів bootstrap і jackknife є доцільним для оцінювання параметрів експоненційного закону зі зсувом.

Висновки

У статті наведено огляд моделей, які використовуються в аналізі виживання, зокрема модель пропорційних ризиків та модель прискорення часу відмови. Також наведено обґрунтування використання методів повторних вибірок в аналізі виживання.

Модель пропорційних ризиків є напівпараметричною моделлю, яка оцінює значущість різних коваріат у впливі на тривалість виживання за допомогою функції ризику. У цій моделі коваріати впливають мультиплікативно на ризик. На основі побудобаної правдоподібності можна отримати оцінки параметрів (за допомогою алгоритму Ньютона-Рафсона), та отримати довірчі інтервали. Серед можливих варіантів підхід Вальда до отримання довірчих інтервалів є найбільш зручним, хоча і не завжди точним.

Модель прискорення часу відмови є параметричною моделлю, яка оцінює значущість різних коваріат у впливі на тривалість виживання. В цій моделі коваріати впливають мультиплікативно на час до виникнення події. Оцінка параметрів відбувається за тим же алгоритмом, що і в моделі пропорційних ризиків, довірчі інтервали будуються за допомогою підходу Вальда.

Використання методів повторних вибірок в аналізі виживання є дуже актуальним через присутність малих вибірок в якості вхідних даних та цензурування, що не дає точно оцінити параметри. Класичні статистичні методи не дають результату бажаної якості. За допомогою

методів повторних вибірок можна отримати багато вибірок на основі початкової вибірки і таким чином більш точно оцінити параметри за допомогою отриманого розподілу оцінок, або побудувати оцінку довірчого інтервалу, не покладаючись на припущення стандартних підходів.

Література

1. Survival Data Analysis [Електронний ресурс] / Patrick Breheny // Lecture notes University of Iowa, Iowa, 2015. – Режим доступу: http://myweb.uiowa.edu/pbreheny/7210/f15/notes.html
2. M. Michalak Application bootstrapping Kaplan-Meier estimate for survival curve smoothing / M. Michalak - London: COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY, 2002. - 64.
3. L. Y. Fang Jackknife and bootstrap inferential procedures for censored survival data / L. Y. Fang. - Selangor: AIP Conference Proceedings, 2015. - 215 p.
4. Denne C, Maag S, Heussen N, Häusler M. A new method to analyse the pace of child development: Cox regression validated by a bootstrap resampling procedure // BMC Pediatrics, 2010, 10:12.