Теорія страхового та фінансового ризику

Ямненко Р.Є

КНУ імені Тараса Шевченка

29 березня 2017 р.

Зміст

Коваріати

Непараметричний підхід обмежений у своїх можливостях при дослідженні деяких важливих питань аналізу виживання, таких як вплив різних факторів (*коваріатів*) на виживання.

Коваріат — це будь-яка величина, пов'язана із кожною особою, така як вік, стать, вид лікування, рівень лікування, серйозність симптомів і т.ін.

Якщо коваріати розбивають популяцію на невелику кількість однорідних груп, то можна порівняти оцінки Каплана-Мейера або інші непараметричні оцінки для кожної групи, але більш прямий і ясний метод — це побудова моделі, у якій вплив коваріатів на виживання моделюється безпосередньо. Це регресійна модель.

У цьому розділі ми будемо припускати, що значення коваріатів для i-ї особи представлені $1 \times p$ вектором z_i.

Найбільш широко вживаною регресійною моделлю в останні роки стала *модель пропорційних ризиків* також відома, як модель Кокса.

Повністю параметричні моделі

У повністю параметричній моделі робиться дуже сильне припущення про те, що розподіл тривалості залишку життя належить заданому сімейству параметричних розподілів, і регресійна задача зводиться до оцінювання параметрів за даними.

Широко вживаними розподілами є показниковий розподіл (сталий ризик), розподіл Вейбула (монотонний ризик), розподіл Гомперца-Мейкхема (показниковий ризик) і log-логістичний розподіл ("горбатий"ризик). Ці розподіли часто використовують як розподіли втрат з даними про страхові позови, але цензурування спостережень значно ускладнює метод максимальної правдоподібності і часто доводиться застосовувати чисельні наближені методи.

Параметричні моделі можна використовувати для однорідних популяцій (випадок однієї вибірки), на відміну від підходу з лекції 6.

Крім того ці моделі можна підігнати до помірної кількості однорідних груп. У останньому випадку довірчі інтервали для підігнаних параметрів дають критерій відмінності груп, що є кращим за непараметричні процедури. Однак повністю параметричні моделі важко застосовувати без попередньої інформації про вигляд функції ризиків.

Тому напів-параметричний підхід є більш популярним.

Модель Кокса

Модель Кокса пропонує таку форму функції ризиків для i-ї особи

$$\lambda(t; z_i) = \lambda_0(t) \exp(\beta z_i^T),$$

тут T означає транспонування вектора, і притримуючись усталених у статистиці позначень, ми позначаємо ризик через λ замість μ.

β — це $1 \times p$ вектор *регресійних параметрів*. Оскільки βz_i^T — це скалярний добуток, то кожний фактор із z_i в функцію ризику входить мультиплікативно. $\lambda_0(t)$ називається *базовим ризиком*.

У представленій простій моделі лише $\lambda_0(t)$ залежить від часу, але можна також використовувати коваріати залежні від часу.

У моделі Кокса функції ризику різних осіб із векторами коваріатів z_1 і z_2 пропорційні протягом будь-якого часу

$$\frac{\lambda(t; z_1)}{\lambda(t; z_2)} = \frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}$$

тому цю модель називають *моделлю пропорційних ризиків*. Для моделі із пропорційними ризиками можна взяти

$$\lambda(t; z_i) = \lambda_0(t) g(z_i),$$

де $g(z)$ — це довільна функція z. Однак модель Кокса забезпечує додатність інтенсивності ризику і дає лінійну модель для логарифму ризику, що дуже зручно і в теорії і в практиці.

Зручність і корисність цієї моделі випливає із того, що загальний вигляд функції ризику всіх осіб визначається базовим ризиком $\lambda_0(t)$, в той час як експоненційна частина враховує відмінність між особами.

Таким чином, якщо ми не цікавимося точною формою функції ризику, а більше цікавимося впливом коваріатів (факторів), то ми можемо не зважати на $\lambda_0(t)$ і оцінювати β на основі даних, не зважаючи на на форму базового ризику. Тому цей підхід називається *напів-параметричним*.

Корисність і гнучкість моделі Кокса зробили її домінуючою у літературі по аналізу виживання. І ця модель є першою, до якої звертаються статистики при аналізі даних виживання.

Часткова функція правдоподібності

Для того, щоб оцінити β потрібно максимізувати наступну часткову функцію правдоподібності. Нехай $R(t_j)$ позначає множину осіб, які знаходяться у групі ризику протягом спостереження за тривалістю життя j-ї особи. Припустимо, що відбувається тільки одна смерть у момент t_j, тобто $d_j = 1,\ (1 \leq j \leq k)$. Тому ймовірність того, що помре j-та особа при умові, що в момент t_j помре рівно одна особа із $R(t_j)$ дорівнює

$$\frac{\lambda(t_j; z_j)}{\sum\limits_{i \in R(T_j)} \lambda(t_j; z_i)} = \frac{\lambda_0(t_j) \exp(\beta z_j^T)}{\lambda_0(t_j) \sum\limits_{i \in R(t_j)} \exp(\beta z_i^T)} = \frac{\exp(\beta z_j^T)}{\sum\limits_{i \in R(t_j)} \exp(\beta z_i^T)}.$$

Ми тут використали співвідношення ${}_h q_t \approx \lambda(t)h$.

Таким чином часткова функція правдоподібності буде мати вигляд

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta z_j^T)}{\sum_{i \in R(t_j)} \exp(\beta z_i^T)}.$$

Ми бачимо, що часткова функція правдоподібності залежить лише від порядку, в якому спостерігається смерть. Назва "часткова"функція правдоподібності виникає тому, що це частина повної функції правдоподібності, яка включає моменти часу в які спостерігалася смерть, а те, що спостерігалось між моментами смерті, відкидається.

Максимізація $L(\beta)$ проводиться чисельно, і більшість статистичних пакетів містять процедуру підгонки моделі Кокса.

На практиці можуть існувати зв'язки між даними, а саме

(a) $d_j > 1$;

(b) деякі спостереження цензуруються протягом періоду спостереження за тривалістю життя.

У випадку (b) вважають, що цензурування відбулось не в момент t_j, а зразу після нього, тому цензуровані в момент t_j особи включаються до групи ризику $R(t_j)$.

У випадку (a) обчислення L досить складні бо потрібно передрати всі можливі комбінації d_j смертей із множини $R(t_j)$ — групи ризику в момент t_j. Тому використовують наближення Бреслова (Breslow)

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta s_j^T)}{\left(\sum_{i \in R(t_j)} \exp(\beta z_i^T) \right)^{d_j}},$$

де s_j — це сума векторів коваріат z всіх d_j осіб, які спостерігались до моменту їх смерті в момент t_j.

Оцінка одержана із максимізації часткової функції правдоподібності має такі ж асимптотичні властивості, як звичайна оцінка максимальної правдоподібності: вона асимптотично нормальна і незміщена, її асимптотична кореляційна матриця оцінюється матрицею оберненою до інформаційної матриці.

Функція

$$u(\beta) = \left(\frac{\partial \ln L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ln L(\beta)}{\partial \beta_p} \right)$$

називається *функцією міри ефективності*.

Розв'язок рівняння $u(\hat{\beta}) = 0$ дає оцінку максимальної правдоподібності $\hat{\beta}$.

Інформаційна матриця, що спостерігається, $I(\hat{\beta})$ має вигляд

$$\left\{ I(\hat{\beta})_{ij} \right\} = \left\{ -\frac{\partial^2 \ln L(\beta)}{\partial \beta_i \partial \beta_j} \right\} \Bigg|_{\beta=\hat{\beta}}, \quad 1 \leq i, j \leq p.$$

Отже кореляційна матриця

$$C = \{\mathrm{cov}(\tilde{\beta}_i, \tilde{\beta}_j)\}$$

асимптотично (при $n \to \infty$, n — об'єм вибірки) співпадає з $I^{-1}(\hat{\beta})$.

Корисною рисою більшості комп'ютерних пакетів для підгонки моделі Кокса є те, що інформаційна матриця обчислена для $\hat{\beta}$ зазвичай має вигляд добутку процесів, які підганяються (це використовується у алгоритмі Ньютона-Рафсона).

Тому є доступними стандартні похибки компонент $\hat{\beta}$. Це буває корисним при підгонці конкретної моделі.

Підгонка моделі

У практичних задачах наявні декілька факторів, і частиною процесу моделювання є вибір тих факторів, які мають значний вплив на процес. Тому необхідні критерії, які оцінюють як вплив самих факторів (коваріатів) так і їх комбінацій.

Загальним критерієм є критерій *відношення правдоподібності*.

Припустимо, що нам необхідно оцінити ефект введення додаткових коваріатів у модель. Нехай ми маємо модель із p коваріатами і іншу модель із $p + q$ коваріатами, які включають p коваріатів першої моделі. Кожна модель підігнана методом максимальної правдоподібності і нехай $\ln L_p$ і $\ln L_{p+q}$ — це максимум логарифмів функцій правдоподібності відповідно першої і другої моделей.

Статисткою відношення правдоподібності буде статистка

$$-2(\ln L_p - \ln L_{p+q}),$$

яка має асимптотично χ^2-розподіл із q степенями свободи при гіпотезі, що додаткові q коваріатів не мають впливу на процес при наявності початкових p коваріатів.

Строго кажучи, ця статистка грунтується на повній функції правдоподібності, але коли підганяється модель Кокса то вона використовується і для часткової функції правдоподібності.

Наприклад, ми розглядаємо модель для вивчення впливу гіпертонії на виживання, де z_i має дві компоненти: $z_i^{(1)}$ — стать, $z_i^{(2)}$ — тиск.

Припустимо, що ми хочемо перевірити гіпотезу, що паління сигарет не впливає на виживання при наявності перших двох факторів. Визначимо розширений вектор коваріатів

$$z_i' = (z_i^{(1)}, z_i^{(2)}, z_i^{(3)}),$$

де $z_i^{(3)}$ — це фактор, який дорівнює 0 для осіб, які не палять, і дорівнює 1 для курців. Статистика відношення правдоподібності

$$-2(\ln L_2 - \ln L_3)$$

має χ^2-розподіл з 1 степеню свободи при нульовій гіпотезі

$$H_0 : \ \beta_3 = 0.$$

Статистика відношення правдоподібності є основою для різноманітних стратегій побудови моделей. При цьому:

(a) можна почати з нульової моделі (без факторів), а потім вводити фактори по одному, або

(b) почати з повної моделі, яка включає всі можливі фактори, а потім виключати ті фактори, які не мають значного впливу.

Додатково потрібно перевірити взаємодію між факторами, у випадку, коли їх вплив залежить від наявності або відсутності кожного з них.

Статистика відношення правдоподібності є стандартним інструментом при відборі моделей, наприклад вона використовується у Великобританії при виборі представників сімейства функцій Гомперца-Мейкхема для градуювання (див. Розділ 13).