



POLITECHNIKA WARSZAWSKA
WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH



PRACA DYPLOMOWA MAGISTERSKA
NA KIERUNKU MATEMATYKA
SPECJALNOŚĆ STATYSTYKA MATEMATYCZNA I ANALIZA DANYCH

**ESTYMACJA W MODELU COXA
METODĄ STOCHASTYCZNEGO SPADKU GRADIENTU
Z PRZYKŁADAMI ZASTOSOWAŃ W ANALIZIE DANYCH
Z THE CANCER GENOME ATLAS**

AUTOR:
MARCIN PIOTR KOSIŃSKI

PROMOTOR:
PROF. NDZW. DR HAB. INŻ PRZEMYSŁAW BIECEK

WARSZAWA, LIPIEC 2015

.....
podpis promotora

.....
podpis autora

Spis treści

Wprowadzenie	5
Podstawy modelu statystycznego	7
1. Estymacja metodą największej wiarygodności	9
1.1. Estymacja	9
1.2. Metoda największej wiarygodności	11
1.3. Asymptotyczne własności estymatora największej wiarygodności	11
1.3.1. Zgodność estymatora największej wiarygodności	12
1.3.2. Asymptotyczna normalność estymatora największej wiarygodności	13
2. Model Coxa	17
2.1. Wprowadzenie do modelu Coxa i nomenklatura	17
2.2. Założenia modelu proporcjonalnego ryzyka Coxa	18
2.3. Estymacja w modelu Coxa	20
2.3.1. Analityczna estymacja współczynników	21
3. Numeryczne metody estymacji	23
3.1. Algorytmy spadku wzdłuż gradientu	24
3.2. Algorytm stochastycznego spadku wzdłuż gradientu I	25
3.2.1. Właściwości stochastycznego spadku wzdłuż gradientu	25
3.3. Ewentualne obrazki obrazujące/porównujące algorytmy	26
4. Estymacja w modelu Coxa metodą stochastycznego spadku gradientu	27
5. Analiza danych genomicznych	29
5.1. Opis i pobranie danych	29
5.2. Analiza	29
A. Wykorzystane narzędzia	31
B. Kody w R	33
C. Dokumentacja pakietu RTCGA	35
Literatura	45

Wprowadzenie

+++ Analiza przeżycia. +++

Najbardziej charakterystyczną cechą typowych danych, jakimi posługuje się w analizie przeżycia, jest obecność obiektów, w których końcowe zdarzenie nastąpiło (wówczas ma się do czynienia z obserwacjami *kompletnymi*), oraz obiektów, w których to zdarzenie (jeszcze) nie nastąpiło (obserwacja *ucięta*). Ta specyficzna postać danych statystycznych doprowadziła do powstania specjalnych metod stosowanych tylko w analizie czasu trwania zjawisk. Jednym z takich modeli jest model proporcjonalnych hazardów Coxa. Jak podaje [2], model proporcjonalnych hazardów Coxa jest jednym z najszerszej stosowanych modeli w onkologicznych publikacjach naukowych, ale także jedną z najmniej rozumianych metod statystycznych. Wynika to z łatwego dostępu do pakietów statystycznych zawierających programy do analizy przeżyć, modeli regresji i analiz wielowariantowych, ale prawie nigdy nie zawierających dobrego opisu podstawowych zasad działania modelu Coxa. Dostarczają one wyłącznie instrukcje, jak wprowadzić dane i uruchomić odpowiednie procedury w celu uzyskania wyniku. Poniższy praca zawiera pełny opis metodologii modelu proporcjonalnych hazardów Coxa, w tym wyjaśnienie najważniejszych pojęć.

++++ Stochastyczny Spadek Gradientu ++++ W przeciągu ostatniej dekady, rozmiary danych rosły szybciej niż prędkość procesorów. W tej sytuacji możliwości statystycznych metod uczenia maszynowego stały się ograniczone bardziej przez czas obliczeń niż przez rozmiary zbiorów danych. Jak podaje [4], bardziej szczegółowa analiza wykazuje jakościowo różne kompromisy w przypadkach problemów uczenia maszynowego na małą i na dużą skalę. Rozwiązania kompromisowe w przypadku dużej skali danych związane są ze złożonością obliczeniową zasadniczych algorytmów optymalizacyjnych, których należy dokonywać w nietrywialny sposób. Jednym z takich rozwiązań są algorytmy optymalizacyjne oparte o stochastyczny spadek gradientu, które wykazują niesamowitą wydajność dla problemów wielkiej skali.

Podstawy modelu statystycznego

W pracy zakłada się znajomość podstaw statystyki matematycznej. Aby ujednolicić oznaczenia w niniejszym rozdziale wprowadzona została klasyczna nomenklatura oparta o [25].

Definicja 0.1. *Model statystyczny* określamy przez podanie rodziny $\{\mathbb{P}_\theta : \theta \in \Theta\}$ rozkładów prawdopodobieństwa na przestrzeni próbkowej Ω oraz zmiennej losowej $X : \Omega \rightarrow \mathcal{X}$, którą traktujemy jako obserwację. Zbiór \mathcal{X} nazywamy przestrzenią obserwacji, zaś Θ nazywamy przestrzenią parametrów.

Symbol θ jest nieznanym parametrem opisującym rozkład badanego zjawiska. Może być jednowymiarowy lub wielowymiarowy. Determinując opis zjawiska poprzez podanie parametru θ , jednoznacznie wyznaczany jest rozkład rozważanego zjawiska spośród całej rodziny rozkładów prawdopodobieństwa $\{\mathbb{P}_\theta : \theta \in \Theta\}$, co umożliwia określenie prawdziwości tezy.

Zakłada się, że przestrzeń próbkowa Ω jest wyposażona w σ -ciało \mathcal{F} . Wtedy:

Definicja 0.2. *Przestrzeń statystyczną* nazywa się trójkę $(\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$.

Wprowadzenie σ -ciała \mathcal{F} sprawia, że przestrzeń statystyczna staje się przestrzenią mierzalną, a więc można na niej określić rodzinę miar $\{\mathbb{P}_\theta : \theta \in \Theta\}$, dzięki której da się ustalić prawdopodobieństwa zajścia wszystkich zjawisk w rozważanej teorii.

W celu budowania niezbędnych pojęć potrzebna jest również definicja losowej próby statystycznej, zazwyczaj nazywanej *próbką*.

Definicja 0.3. *Losową próbą statystyczną* nazywamy zbiór obserwacji statystycznych wylosowanych z populacji, które są realizacjami ciągu zmiennych losowych o rozkładzie takim jak rozkład populacji.

Rozdział 1

Estymacja metodą największej wiarogodności

*The making of maximum likelihood was one of the most important developments in 20th century statistics. It was the work of one man but it was no simple process (...).
John Aldrich o R. A. Fisher'ze*

1.1. Estymacja

Estymacja to dział wnioskowania statystycznego będący zbiorem metod pozwalających na uogólnianie wyników badania próby losowej na nieznaną postać i parametry rozkładu zmiennej losowej całej populacji oraz szacowanie błędów wynikających z tego uogólnienia [37].

W statystyce parametrycznej zakłada się, że rozkład prawdopodobieństwa opisujący doświadczenie należy do rodziny $\{\mathbb{P}_\theta : \theta \in \Theta\}$, ale nie zna się parametru θ . Można go jednak szacować dzięki estymatorom opartym na statystykach.

Definicja 1.1. *Statystyka*, dla $X = (X_1, \dots, X_n)$, to odwzorowanie mierzalne $T : \mathcal{X} \rightarrow \mathcal{R}$.

Definicja 1.2. *Estymatorem* parametru θ nazywamy dowolną statystykę $T = T(X)$, gdzie X to próba z badanego rozkładu, o wartościach w zbiorze Θ .

Interpretuje się T jako przybliżenie θ i często estymator θ oznacza symbolem $\hat{\theta}$. Niekiedy w kręgu zainteresowań jest również estymacja $g(\theta)$, gdzie g to ustalona funkcja.

Pewne estymatory mające odpowiednie własności są preferowane nad inne ze względu na większą precyzję bądź ufność oszacowania danego estymatora. Poniżej przedstawione są 2 ważne definicje związane z jakością estymatorów [25], gdy rozmiar próbki X_1, \dots, X_n jest duży. Mówi się wtedy o własnościach asymptotycznych estymatorów, które z matematycznego punktu widzenia są twierdzeniami granicznymi, w których n dąży do nieskończoności. Dzięki tym twierdzeniom możliwe jest opisanie w przybliżeniu zachowania estymatorów dla dostatecznie dużych próbek. Niestety, teoria asymptotyczna nie dostarcza informacji o tym, jak duża powinna być próbka, żeby przybliżenie było dostatecznie dobre.

Definicja 1.3. Estymator $\hat{g}(X_1, \dots, X_n)$ wielkości $g(\theta)$ jest **nieobciążony**, jeśli dla każdego n

$$\mathbb{E}\hat{g}(X_1, \dots, X_n) = g(\theta).$$

Definicja 1.4. Estymator $\hat{g}(X_1, \dots, X_n)$ wielkości $g(\theta)$ jest **zgodny**, jeśli dla każdego $\theta \in \Theta$

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{g}(X_1, \dots, X_n) - g(\theta)| \leq \varepsilon) = 1,$$

dla każdego $\varepsilon > 0$.

Definicja 1.5. Estymator $\hat{g}(X_1, \dots, X_n)$ wielkości $g(\theta)$ jest **mocno zgodny**, jeśli

$$\mathbb{P}_\theta\left(\lim_{n \rightarrow \infty} \hat{g}(X_1, \dots, X_n) = g(\theta)\right) = 1.$$

Zgodność (mocna zgodność) znaczy tyle, że

$$\hat{g}(X_1, \dots, X_n) \rightarrow g(\theta), \quad (n \rightarrow \infty)$$

według prawdopodobieństwa (prawie na pewno). Interpretacja jest taka: estymator jest uznany za zgodny, jeśli zmierza do estymowanej wielkości przy nieograniczonym powiększaniu badanej próbki.

Jednak zgodność (nawet w mocnym sensie) nie jest specjalnie satysfakcjonującą własnością estymatora, a zaledwie minimalnym żądaniem, które powinien spełniać każdy przyzwoity estymator. Dlatego od niektórych estymatorów żąda się silniejszych właściwości, takich jak asymptotyczna normalność.

Definicja 1.6. Estymator $\hat{g}(X_1, \dots, X_n)$ wielkości $g(\theta)$ jest **asymptotycznie normalny**, jeśli dla każdego $\theta \in \Theta$ istnieje funkcja $\sigma^2(\theta)$, zwana asymptotyczną wariancją, taka że

$$\sqrt{n}(\hat{g}(X_1, \dots, X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)), \quad (n \rightarrow \infty).$$

\xrightarrow{d} oznacza
zbieżność wg
rozkładu.

Oznacza to, że rozkład prawdopodobieństwa statystyki $\hat{g}(X_1, \dots, X_n)$ jest dla dużych n zbliżony do rozkładu

$$\mathcal{N}\left(g(\theta), \frac{\sigma^2(\theta)}{n}\right).$$

Inaczej mówiąc, estymator jest asymptotycznie normalny, gdy:

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta\left(\frac{\sqrt{n}}{\sigma(\theta)}(\hat{g}(X_1, \dots, X_n) - g(\theta)) \leq a\right) = \Phi(a),$$

gdzie $\Phi(x)$ to dystrybuenta standardowego rozkładu normalnego $\mathcal{N}(0, 1)$.

Asymptotyczna normalność mówi, że estymator nie tylko zbiega do nieznanego parametru, ale również że zbiega wystarczająco szybko, jak $\frac{1}{\sqrt{n}}$, czyli, że

$$\mathbb{P}_\theta\left(\frac{\sqrt{n}}{\sigma(\theta)}(\hat{g}(X_1, \dots, X_n) - g(\theta)) \leq a\right) - \Phi(a) = f(n) \in \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Jeśli estymator jest asymptotycznie normalny, to jest zgodny, choć nie musi być *mocno zgodny*.

W dalszej części tego rozdziału zostanie wprowadzone pojęcie estymatora największej wiarygodności oraz zostaną udowodnione dla niego jego właściwości, co utwierdzi w przekonaniu, że metoda największej wiarygodności, przy odpowiednich założeniach, jest metodą konstrukcji rozsądnych estymatorów.

1.2. Metoda największej wiarygodności

Metodę największej wiarygodności wprowadził R. A. Fisher w 1922 r. [13], dla której po raz pierwszy procedurę numeryczną zaproponował już w 1912 r. [12]. O burzliwym procesie powstawania metody, o zmianach w jej uzasadnieniu, o koncepcjach, które powstały w obrębie tej metody takich jak parametr, statystyka, wiarygodność, dostateczność czy efektywność oraz o podejściach, które Fisher odrzucił tworząc podstawy pod nową teorię można przeczytać w obszernej pracy dokumentalnej [1].

Metoda ta, jako alternatywa dla metody najmniejszych kwadratów [21], [15], była rozwijana i szeroko stosowana później przez wielu statystyków, i wciąż znajduje obszerne zastosowania w wielu obszarach estymacji statystycznej, np. [18], [19], [23].

Aby zdefiniować estymator oparty o metodę największej wiarygodności, należy najpierw wprowadzić pojęcie funkcji wiarygodności.

Definicja 1.7. *Funkcją wiarygodności nazywamy funkcję $L : \Theta \rightarrow \mathbb{R}$ daną wzorem*

$$L(\theta) = L(\theta; x_1, \dots, x_n) = f(\theta; x_1, \dots, x_n),$$

którą rozważamy jako funkcję parametru θ przy ustalonych wartościach obserwacji x_1, \dots, x_n , gdzie

$$f(\theta; x_1, \dots, x_n) = \begin{cases} \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n), & \text{dla rozkładów dyskretnych,} \\ f_\theta(x_1, \dots, x_n), & \text{dla rozkładów absolutnie ciągłych.} \end{cases}$$

Oznacza to, że wiarygodność jest właściwie tym samym, co gęstość prawdopodobieństwa, ale rozważana jako funkcja parametru θ , przy ustalonych wartościach obserwacji $x = X(\omega)$.

Definicja 1.8. *Estymatorem największej wiarygodności parametru θ , oznaczanym $ENW(\theta)$, nazywamy wartość parametru, w której funkcja wiarygodności przyjmuje supremum*

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

Takie supremum może nie istnieć, dlatego niektóre pozycje w literaturze, w definicji estymatora największej wiarygodności, supremum zastępują wartością największą [31], [38], [27].

1.3. Asymptotyczne własności estymatora największej wiarygodności

W tym podrozdziale zostanie wykazane, że estymator największej wiarygodności jest

- i) zgodny,
- ii) asymptotycznie normalny,
- iii) asymptotycznie nieobciążony.

Asymptotyczna nieobciążoność wynika z asymptotycznej normalności.

Dowody w tym rozdziale są znane w literaturze i opierają się o [28] i [38].

1.3.1. Zgodność estymatora największej wiarygodności

Chcąc wykazać zgodność estymatora największej wiarygodności przy pewnych warunkach regularności przydatna będzie poniższa definicja i następujący Lemat.

Definicja 1.9. *Funkcja log-wiarygodności to funkcja spełniająca równanie*

$$\ell(\theta) = \log(L(\theta)),$$

gdzie przyjmuje się $\ell(\theta) = -\infty$ jeśli $L(\theta) = 0$.

Lemat 1.1. *Gdy θ_0 to prawdziwe maksimum funkcji wiarygodności, to dla każdego $\theta \in \Theta$*

$$\mathbb{E}_{\theta_0} \ell(\theta) \leq \mathbb{E}_{\theta_0} \ell(\theta_0).$$

Dowód. Rozważając różnicę, przy założeniu ciągłości rozkładu:

$$\begin{aligned} \mathbb{E}_{\theta_0} \ell(\theta) - \mathbb{E}_{\theta_0} \ell(\theta_0) &= \mathbb{E}_{\theta_0} (\ell(\theta) - \ell(\theta_0)) = \mathbb{E}_{\theta_0} (\log f(\theta; X) - \log f(\theta_0; X)) \\ &= \mathbb{E}_{\theta_0} \log \frac{f(\theta; X)}{f(\theta_0; X)}, \end{aligned}$$

i pamiętając o tym, że $\log t \leq t - 1$, można dojść do

$$\begin{aligned} \mathbb{E}_{\theta_0} \log \frac{f(\theta; X)}{f(\theta_0; X)} &\leq \mathbb{E}_{\theta_0} \left(\frac{f(\theta; X)}{f(\theta_0; X)} - 1 \right) = \int \left(\frac{f(\theta; x)}{f(\theta_0; x)} - 1 \right) f(\theta_0; x) dx \\ &= \int f(\theta; x) dx - \int f(\theta_0; x) dx = 1 - 1 = 0. \end{aligned}$$

Obie całki równają się 1 jako, że są całkami z funkcji gęstości, zaś równość w nierówności zachodzi tylko wtedy, gdy $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$. ■

Dzięki temu wynikowi możliwe jest udowodnienie poniższego Twierdzenia.

Twierdzenie 1.2. *Pod pewnymi warunkami regularności nałożonymi na rodzinę rozkładów prawdopodobieństwa, estymator największej wiarygodności ENW(θ) jest zgodny, tzn.*

$$ENW(\theta) \rightarrow \theta \quad \text{dla} \quad n \rightarrow \infty.$$

Dowód.

1) Z definicji w $ENW(\theta)$ przyjmowana jest wartość największa funkcji $L(\theta)$, a więc tym bardziej funkcji $\ell(\theta) = \log L(\theta)$ oraz funkcji

$$\ell_n(\theta) = \frac{1}{n} \ell(\theta) = \frac{1}{n} \log L(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(\theta; X_i)$$

(zakładając ciągłość rozkładu i niezależność X_1, \dots, X_n), gdyż ekstremum jest niezmiennicze ze względu na monotoniczną transformację i liniowe przekształcenie jakim jest dzielenie.

2) Z Lematu 1.1 wynika, że θ_0 maksymalizuje $\mathbb{E}_{\theta_0} \ell(\theta)$.

3) Z Prawa Wielkich Liczb, które jest spełnione gdy założy się, że X_i to realizacje ciągu zmiennych losowych o skończonych wartościach oczekiwanych, wynika, że

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(\theta; X_i) \rightarrow \mathbb{E}_{\theta_0} \ell(\theta),$$

co ostatecznie oznacza, że $ENW(\theta)$ jest zgodny. ■

\mathbb{E}_{θ_0} oznacza
wartość oczekiwaną
względem rozkładu
parametryzowane-
go przez
 θ_0 .

1.3.2. Asymptotyczna normalność estymatora największej wiarygodności

Fisher w swojej karierze wprowadził wiele pożytecznych pojęć stosowanych do dziś. Jednym z nich jest Informacja Fishera, która zostanie wykorzystana w dowodzie asymptotycznej normalności estymatora największej wiarygodności.

Definicja 1.10. Niech X będzie zmienną losową o gęstości f_θ , zależnej od jednowymiarowego parametru $\theta \in \Theta \subset \mathbb{R}$. **Informację Fishera** zawartą w obserwacji X nazywa się funkcję

$$\mathcal{I}(\theta) = \mathbb{E}_\theta(\ell'(\theta; X))^2 = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right)^2, \quad (1.1)$$

gdzie odpowiednio

$$\begin{aligned} \mathcal{I}(\theta) &= \int \left(\frac{\partial}{\partial\theta} \log f_\theta(x)\right)^2 f_\theta(x) dx && \text{dla zmiennej ciągłej;} \\ \mathcal{I}(\theta) &= \sum_x \left(\frac{\partial}{\partial\theta} \log f_\theta(x)\right)^2 \mathbb{P}_\theta(X = x) && \text{dla zmiennej dyskretnej.} \end{aligned}$$

W dowodzie asymptotycznej normalności estymatora największej wiarygodności kluczowymi założeniami są poniższe warunki regularności. Rodzina gęstości musi być dostatecznie regularna aby pewne kroki rachunkowe w dalszych rozumowaniach były poprawne.

Definicja 1.11. Warunki regularności.

- i) Informacja Fishera jest dobrze określona. Zakłada się, że Θ jest przedziałem otwartym, istnieje pochodna $\frac{\partial}{\partial\theta} \log f_\theta$, całka/suma we wzorze (1.1) jest bezwzględnie zbieżna (po obłożeniu funkcji podcałkowej modulem całka istnieje i jest skończona) i $0 < \mathcal{I}(\theta) < \infty$.
- ii) Wszystkie gęstości f_θ mają jeden nośnik, tzn. zbiór $\{x \in X : f_\theta(x) > 0\}$ nie zależy od θ .
- iii) Można przenosić pochodną przed znak całki, czyli zamienić kolejność operacji różniczkowania $\frac{\partial}{\partial\theta}$ i całkowania $\int \dots dx$.

Wprowadzając takie założenia, otrzymano przydatne właściwości Informacji Fishera.

Stwierdzenie 1.3. Jeśli spełnione są warunki regularności (1.11) to:

$$\begin{aligned} (i) \quad & \mathbb{E}_\theta \frac{\partial}{\partial\theta} \log f_\theta(X) = 0, \\ (ii) \quad & \mathcal{I}(\theta) = \mathbb{V}ar_\theta\left(\frac{\partial}{\partial\theta} \log f_\theta(X)\right), \\ (iii) \quad & \mathcal{I}(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2} \log f_\theta(X)\right). \end{aligned}$$

Dowód tego stwierdzenia można znaleźć w [25].

Patrząc na postać pochodnej funkcji log-wiarygodności

$$\ell'(\theta_0; X) = (\log f(\theta_0; X))' = \frac{f'(\theta_0; X)}{f(\theta_0; X)},$$

można wywnioskować, że nieformalnie interpretacja Informacji Fishera jest miarą tego jak szybko zmieni się funkcja gęstości jeśli delikatnie zmieni się parametr θ w okolicach θ_0 . Biorąc kwadrat i wartość oczekiwaną, innymi słowy uśredniając po X , otrzymuje się uśrednioną wersję tej miary. Jeżeli Informacja Fishera jest duża, oznacza to, że gęstość zmieni się szybko gdyby poruszyć parametr θ_0 , innymi słowy - gęstość z parametrem θ_0 jest znacząco inna i może zostać łatwo odróżniona od gęstości z parametrami nie tak bliskimi θ_0 . Stąd wiemy, że możliwa estymacja θ_0 oparta o takie dane jest dobra. Z drugiej strony, jeżeli Informacja Fishera jest mała, oznacza to, że gęstość dla θ_0 jest bardzo podobna do gęstości z parametrami nie tak bliskimi do θ_0 , a co za tym idzie, dużo ciężiej będzie odróżnić tę gęstość, czyli estymacja będzie słabsza.

Dzięki pojęciu Informacji Fishera i warunkom regularności, możliwe jest udowodnienie poniższego Twierdzenia.

Twierdzenie 1.4. *Pod pewnymi warunkami regularności nałożonymi na rodzinę rozkładów prawdopodobieństwa, estymator największej wiarygodności jest asymptotycznie normalny,*

$$\sqrt{n}(ENW(\theta) - \theta_0) \rightarrow \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta_0)}\right).$$

Z Twierdzenia widać, że im większa Informacja Fishera tym mniejsza asymptotyczna wariancja estymatora prawdziwego parametru θ_0 .

Dowód. Ponieważ $ENW(\theta)$ maksymalizuje $\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(\theta; X)$, to $\ell'_n(\theta) = 0$.

Dalej, korzystając z Twierdzenia o Wartości Średniej:

$$\frac{g(a) - g(b)}{a - b} = g'(c) \text{ albo } g(a) = g(b) + g'(c)(a - b), \text{ dla } c \in [a, b],$$

gdzie $g(\theta) = \ell'_n(\theta)$, $a = ENW(\theta)$, $b = \theta_0$, można zapisać równość

$$0 = \ell'_n(ENW(\theta)) = \ell'_n(\theta_0) + \ell''_n(\theta_1)(ENW(\theta) - \theta_0), \text{ dla } \theta_1 \in [ENW(\theta), \theta_0],$$

a z niej przejść do postaci

$$\sqrt{n}(ENW(\theta) - \theta_0) = -\frac{\sqrt{n}\ell'_n(\theta_0)}{\ell''_n(\theta_1)}. \quad (1.2)$$

Z Lematu (1.1) wynika, że θ_0 maksymalizuje $\mathbb{E}_{\theta_0}\ell(\theta_0)$ czyli

$$\mathbb{E}_{\theta_0}\ell'(\theta_0) = 0, \quad (1.3)$$

a to można wstawić do licznika w równaniu (1.2)

$$\begin{aligned} \sqrt{n}\ell'_n(\theta_0) &= \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n \ell'(\theta_0) - 0\right) \\ &= \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n \ell'(\theta_0) - \mathbb{E}_{\theta_0}\ell'(\theta_0)\right) \rightarrow \mathcal{N}\left(0, \text{Var}_{\theta_0}(\ell'(\theta_0))\right), \end{aligned} \quad (1.4)$$

gdzie zbieżność wynika z Centralnego Twierdzenia Granicznego.

Następnie można rozważyć mianownik w równaniu (1.2). Dla wszystkich θ wynika

$$\ell''(\theta) = \frac{1}{n} \sum_{i=1}^n \ell''(\theta) \rightarrow \mathbb{E}_{\theta_0} \ell''(\theta)$$

z Prawa Wielkich Liczb.

Dodatkowo, ponieważ $\theta_1 \in [ENW(\theta), \theta_0]$ a $ENW(\theta)$ jest zgodny (poprzedni podrozdział), to ponieważ $ENW(\theta) \rightarrow \theta_0$, to też $\theta_1 \rightarrow \theta_0$, a wtedy

$$\ell''_n(\theta_1) \rightarrow \mathbb{E}_{\theta_0} \ell''(\theta_0) = -\mathcal{I}(\theta_0)$$

z punktu (iii) ze Stwierdzenia (1.3).

Wtedy prawa strona równania (1.2), dzięki (1.4)

$$-\frac{\sqrt{n}\ell'_n(\theta_0)}{\ell''_n(\theta_1)} \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{V}ar_{\theta_0}(\ell'(\theta_0))}{(\mathcal{I}(\theta_0))^2}\right).$$

Ostatecznie wariancja

$$\mathbb{V}ar_{\theta_0}(\ell'(\theta_0)) \stackrel{z \text{ def}}{=} \mathbb{E}_{\theta_0}(\ell'(\theta_0))^2 - (\mathbb{E}_{\theta_0} \ell'(\theta_0))^2 = \mathcal{I}(\theta_0) - 0,$$

co wynika z definicji Informacji Fishera i (1.3).

■

Rozdział 2

Model Coxa

W tym rozdziale zostanie przedstawiony model proporcjonalnych hazardów Coxa. Głównym celem tej pracy jest wykorzystanie numerycznej metody estymacji współczynników metodą stochastycznego spadku gradientu w omawianym modelu. Więcej o estymacji metodą stochastycznego spadku gradientu napisane jest w rozdziale 3.2. Definicje i twierdzenia w tym rozdziale oparte są o [9], [34], [2] i [7].

2.1. Wprowadzenie do modelu Coxa i nomenklatura

Analiza przeżycia, w której model Coxa znalazł największe zastosowania, polega na modelowaniu wpływu czynników na czas do wystąpienia pewnego zdarzenia. Zdarzeniem może być np. śmierć pacjenta, awaria urządzenia, zerwanie umowy przez klienta, odejście z pracy pracownika lub deaktywacja pewnej usługi. Analizując czasy do wystąpienia zdarzenia wykorzystuje się funkcję przeżycia bądź niosącą równoważną informację funkcję hazardu.

Definicja 2.1. *Funkcja przeżycia w grupie j , to funkcja, która spełnia*

$$S_j(t) = \mathbb{P}(T_{j,i}^* \geq t) = 1 - F_j(t), t \in \mathbb{R} \quad (2.1)$$

gdzie $F_j(t)$ to dystrybuanta rozkładu zadanego gęstością $f_j(t)$.

Definicja 2.2. *Funkcja hazardu to funkcja, która wyraża się wzorem*

$$\begin{aligned} \lambda_j(t) &= \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq T^* \leq t+h | T^* \geq t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq T^* \leq t+h)}{h} \cdot \frac{1}{\mathbb{P}(T^* \geq t)} \\ &= \lim_{h \rightarrow 0} \frac{F_j(t+h) - F_j(t)}{h} \cdot \frac{1}{S_j(t)} = \frac{f_j(t)}{S_j(t)}. \end{aligned} \quad (2.2)$$

W tych definicjach T^* oznacza czas do wystąpienia zdarzenia. Zakłada się, że wewnątrz każdej grupy $j = 1, 2, \dots, m$ wyznaczonej przez poziomy zmiennych objaśniających, czasy $T_{j,i}^*$ dla $i = 1, \dots, n_j$ to niezależne zmienne losowe z tego samego rozkładu o gęstości $f_j(t)$.

Wartość funkcji hazardu w momencie t traktuje się jako chwilowy potencjał pojawiającego się zdarzenia (np. śmierci lub choroby), pod warunkiem że osoba dożyła czasu t . Funkcja hazardu nazywana jest również funkcją ryzyka, intensywnością umieralności (*force of mortality*), umieralnością chwilową (*instantaneous death rate*) lub chwilową częstością niepowodzeń (awarii) (*failure rate*). Ostatniego określenia używa się w teorii odnowy [8], w której analizuje się awaryjność elementów przemysłowych.

Model proporcjonalnych hazardów Coxa [9] jest obecnie najczęściej stosowaną procedurą do modelowania relacji pomiędzy zmiennymi objaśniającymi a przeżyciem, lub innym cenzurowanym zdarzeniem. Model ten umożliwia analizę wpływu czynników prognostycznych na przeżycie. Sir David Cox opracował tego typu model dla tabeli przeżyć i zilustrował zastosowanie modelu dla przypadku białaczki, ale model może być stosowany do obliczania przeżyć w odniesieniu do innych chorób, jak w przypadku przeżyć w chorobach nowotworowych lub kardiologicznych po transplantacji serca lub zawałach serca [26].

Definicja 2.3. *Model Coxa zakłada postać funkcji hazardu dla i -tej obserwacji X_i jako*

$$\lambda_i(t) = \lambda_0(t)e^{X_i(t)'\beta}, \quad (2.3)$$

gdzie λ_0 to niesprecyzowana nieujemna funkcja nazywana bazowym hazardem, a β to wektor współczynników rozmiaru p , co odpowiada liczbie zmiennych objaśniających w modelu Coxa.

Takie sformułowanie modelu gwarantuje, że funkcja hazardu jest nieujemna.

Model Coxa, dla wersji kiedy współczynniki są stałe w czasie, nazywany jest **modelem proporcjonalnych hazardów**, gdyż stosunek (proporcja) hazardów dla dwóch obserwacji X_i oraz X_j jest stały w czasie:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} = e^{(X_i - X_j)\beta}.$$

Oznacza to, że hazard dla jednej obserwacji można uzyskać poprzez przemnożenie hazardu dla innej obserwacji przez pewną stałą c_{ij} :

$$\lambda_i(t) = \frac{e^{X_i'\beta}}{e^{X_j'\beta}} \cdot \lambda_j(t) = c_{ij} \cdot \lambda_j(t).$$

W modelu proporcjonalnych hazardów istotnym elementem jest estymacja stałych c_{ij} .

2.2. Założenia modelu proporcjonalnego ryzyka Coxa

Model Coxa znalazł szerokie zastosowanie w sytuacjach, gdy analiza wymaga wykorzystania cenzurowanych danych. Model Coxa jest w stanie wykorzystać je do estymacji współczynników w modelu, przekładających się na proporcje hazardów. Z uwagi na aspekt praktyczny podyktowany warunkami technicznymi prób klinicznych i badań biologicznych, zbiory danych klinicznych zawierają cenzurowane czasy zdarzeń. Oznacza to, że w wielu przypadkach niemożliwe jest obserwowanie czasu zdarzeń dla wszystkich obserwacji w zbiorze. Niekiedy jest to uwarunkowane zbyt długim czasem do wystąpienia zdarzenia. Czasem jest to związane z zaplanowanym okresem próby klinicznej, który jest krótszy niż czas do zdarzenia dla

pacjentów, którzy mogli zostać włączeni do próby klinicznej pod koniec jej trwania i nie udało się dla nich zaobserwować czasów zdarzeń. W wielu przypadkach pacjenci, traktowani jako obserwacje w zbiorze, znikają z pola widzenia w momencie, gdy np. przestają pojawiać się na wizytach kontrolnych. Może być to spowodowane negatywnymi relacjami z lekarzem prowadzącym lub przeprowadzką. W takich sytuacjach wykorzystuje się daną obserwację do momentu jej ostatniej kontroli. Nie rezygnuje się z tej obserwacji w analizie i wykorzystuje się o niej informacje w pełni dla czasu, w którym przebywała pod obserwacją. Jest to ogromna zaleta modelu Coxa.

Z przyczyny cenzurowanych danych potrzebne są założenia modelu dotyczące cenzurowania czasów, które opierają się o następujące definicje.

Definicja 2.4. *Cenzurowanie prawostronne polega na zaobserwowaniu czasu*

$$T = \min(T^*, C),$$

gdzie T^* to prawdziwy czas zdarzenia, zaś C jest nieujemną zmienną losową.

Definicja 2.5. *Cenzurowanie jest niezależne jeśli zachodzi*

$$\lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq T^* \leq t + h | T^* \geq t)}{h} = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq T^* \leq t + h | T^* \geq t, Y(t) = 1)}{h},$$

gdzie $Y(t) = 1$ jeśli do chwili t nie wystąpiło zdarzenie ani cenzurowanie, czyli jednostka pozostaje narażona na ryzyko zdarzenia oraz $Y(t) = 0$ w przeciwnym wypadku.

Interpretacja tej definicji jest następująca: jednostka cenzurowana w chwili t jest reprezentatywna dla wszystkich innych narażonych na ryzyko zdarzenia w chwili t . Innymi słowy cenzurowanie nie wybiera z populacji osobników bardziej albo mniej narażonych na zdarzenie. Cenzurowanie działa niezależnie od mechanizmu występowania zdarzenia.

Definicja 2.6. *Cenzurowanie jest nie-informatywne jeśli zachodzi*

$$g(t; \theta, \phi) \equiv g(t; \phi), \quad (2.4)$$

gdzie $g(t; \theta, \phi)$ jest funkcją gęstości dla cenzurowań C_i wyrażonych jako niezależne zmienne losowe o jednakowym rozkładzie, zaś prawdziwe czasy T_i^* są interpretowane jako niezależne zmienne losowe o jednakowym rozkładzie i funkcji gęstości $f(t; \theta)$, czyli θ parametryzuje jedynie rozkład czasów zdarzeń.

Oznacza to, że cenzurowanie nie daje informacji o parametrach rozkładu czasów zdarzeń, ponieważ nie zależy od parametrów od których zależy hazard.

Model proporcjonalnych hazardów Coxa oparty jest na założeniach:

- i) Współczynniki modelu $\beta_k, k = 1, \dots, p$ są stałe w czasie, co przekłada się na to, że stosunek hazardów dla dwóch obserwacji jest stały w czasie.
- ii) Postać funkcjonalna efektu zmiennej niezależnej, czyli postać modelu $\lambda_i(t) = \lambda_0(t)e^{X_i(t)'\beta}$.
- iii) Obserwacje są niezależne.
- iv) Cenzurowanie czasów jest nie-informatywne.
- v) Cenzurowanie czasów jest niezależne (od mechanizmu występowania zdarzenia).

2.3. Estymacja w modelu Coxa

Funkcja hazardu jest wykładniczą funkcją zmiennych objaśniających, nieznana jest natomiast postać bazowej funkcji hazardu, co bez dalszych założeń uniemożliwia estymację standardową metodą największej wiarygodności. Rozwiązaniem Cox'a jest maksymalizacja tylko tego fragmentu funkcji wiarygodności, który zależy jedynie od estymowanych parametrów. W modelu Coxa proporcjonalnych hazardów estymacja współczynników β oparta jest o częściową funkcję wiarygodności, którą wprowadził Cox w 1972 r. [9].

Dla konkretnego czasu zdarzenia t_i , gdzie w zbiorze obserwowanych jest K czasów zdarzeń, prawdopodobieństwo warunkowe ze względu na licznosc zbioru ryzyka w czasie t_i , że czas zdarzenia dotyczy i -tej jednostki spośród wciąż obserwowanych jest równe

$$\frac{e^{X'_i \beta}}{\sum_{l \in \mathcal{R}(t_i)} e^{X'_l \beta}}, \quad (2.5)$$

gdzie *zbiór ryzyka* $\mathcal{R}(t_i)$, w chwili t_i , rozumiany jest jako zbiór indeksów obserwacji, które są w danym czasie t_i pod obserwacją.

Chcąc estymować współczynniki metodą największej wiarygodności należy rozważyć funkcję wiarygodności, która dla niezależnego cenzurowania prawostronnego ma postać:

$$L(\beta, \varphi) = L_p(\beta) \cdot L^*(\beta, \varphi), \quad (2.6)$$

gdzie

$$L_p(\beta) = \prod_{i=1}^n f(t_j; \beta)^{\delta_j} S(t_j; \beta)^{1-\delta_j} = \prod_{i=1}^n \lambda(t_j; \beta)^{\delta_j} S(t_j; \beta) \quad (2.7)$$

to częściowa funkcja wiarygodności, a $L^*(\beta, \varphi)$ zależy od cenzurowania (parametr φ).

Wtedy dla niezależnego cenzurowania i dla czasów zdarzeń, które nie zaszły jednocześnie **częściowa funkcja wiarygodności w modelu Coxa** ma postać:

$$L_p(\beta) = \prod_{i=1}^K \frac{e^{X'_i \beta}}{\sum_{l=1}^n Y_l(t_i) e^{X'_l \beta}}, \quad (2.8)$$

gdzie $Y_l(t_i) = 1$, gdy obserwacja X_l jest w zbiorze ryzyka w czasie t_i , i $Y_l(t_i) = 0$ w przeciwnym przypadku, n to liczba obserwacji w zbiorze, a K to wspomniana wyżej liczba zaobserwowanych czasów zdarzeń. Zaletą takiej postaci funkcji częściowej wiarygodności jest to, że w jej wzorze nie występuje funkcja bazowego hazardu, zatem estymacja współczynników może odbywać się bez znajomości jej postaci.

Jeśli dodatkowo cenzurowanie jest nie-informatywne, to $L_p(\beta)$ jest **pełną** funkcją wiarygodności, bowiem wówczas

$$L^*(\beta, \varphi) \propto L^*(\varphi)$$

co bierze się z definicji cenzurowania nie-informatywnego (2.4)

$$g(t; \theta, \phi) \equiv g(t; \phi).$$

Ponieważ model proporcjonalnych hazardów Coxa zakłada niezależność i nie-informatywność cenzurowania zatem można uważać, że częściowa funkcja wiarygodności daje pełną informację o współczynnikach i wnioskowanie w oparciu o nią jest uzasadnione i poprawne.

W sytuacjach, gdy nie jest spełnione założenie nie-informatywności cenzurowania i częściowa funkcja wiarygodności nie jest funkcją wiarygodności w sensie bycia proporcjonalną do prawdopodobieństwa obserwowanego zbioru, można ją traktować jako funkcję wiarygodności dla celów asymptotycznego wnioskowania o współczynnikach modelu, zobacz [?].

2.3.1. Analityczna estymacja współczynników

Standardowo w celu znalezienia maximum, aby ułatwić obliczenia, można rozważaną funkcję obłożyć monotoniczną transformacją jaką jest logarytm, tak aby w konsekwencji otrzymać **częściową funkcję log-wiarygodności**

$$\ell_p(\beta) = \sum_{i=1}^K X_i' \beta - \sum_{i=1}^K \log \left(\sum_{l \in \mathcal{R}(t_i)} e^{X_l' \beta} \right). \quad (2.9)$$

Bezpośrednie kalkulacje dają p -wymiarowy wektor pochodnych, dla $k = 1, \dots, p$

$$U_k(\beta) = \frac{\partial \ell_k(\beta)}{\partial \beta_k} = \sum_{i=1}^K (X_{ik} - A_{ik}), \quad (2.10)$$

X_{ik} to i ta obserwacja i k ta zmienna.

gdzie sumowanie jest po $l \in \mathcal{R}(t_i)$, zaś czynnik

$$A_{ik} = \frac{\sum X_{ik} e^{X_l' \beta}}{\sum e^{X_l' \beta}} \quad (2.11)$$

to średnia z $X_{.k}$ (k -tych zmiennych) po skończonej populacji $\mathcal{R}(t_i)$, z wykorzystaniem *ważonej eksponencjalnie* formy próbkowania.

Z kolei drugie pochodne cząstkowe, jak podaje [9], mają postać dla $k_1, k_2 = 1, \dots, p$

$$\mathcal{J}_{k_1 k_2}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta_{k_1} \partial \beta_{k_2}} = \sum_{i=1}^K C_{ik_1 k_2}(\beta), \quad (2.12)$$

gdzie

$$C_{ik_1 k_2}(\beta) = \frac{\sum X_{ik_1} X_{ik_2} e^{X_l' \beta}}{\sum e^{X_l' \beta}} - A_{ik_1}(\beta) A_{ik_2}(\beta) \quad (2.13)$$

to kowariancja pomiędzy $X_{.k_1}$ (k_1 -tymi zmiennymi) a $X_{.k_2}$ (k_2 -tymi zmiennymi) przy tej formie ważonego próbkowania.

Estymator największej wiarygodności β można uzyskać poprzez przyrównanie (2.10) do 0, a numerycznie poprzez iteracyjne wykorzystanie (2.10) oraz (2.12) w algorytmie spadku gradientu rzędu II nazywanego również Algorytmem Raphsona-Newtona, który jest opisany w podrozdziale ???. Jest to tradycyjne i szeroko stosowane podejście do estymacji współczynników w modelu proporcjonalnych hazardów Coxa. Niniejsza praca skupia się na wykorzystaniu metody estymacji współczynników jaką jest metoda stochastycznego spadku wzdłuż gradientu, która jest szerzej opisana w następnym rozdziale w podrozdziale 3.2.

Rozdział 3

Numeryczne metody estymacji

Przez **numerykę** rozumie się dziedzinę matematyki zajmującą się rozwiązywaniem przybliżonych zagadnień algebraicznych. Odkąd zjawiska przyrodnicze zaczęto opisywać przy użyciu formalizmu matematycznego, pojawiła się potrzeba rozwiązywania zadań analizy matematycznej czy algebry. Dopóki były one nieskomplikowane, dawały się rozwiązywać analitycznie, tzn. z użyciem pewnych przekształceń algebraicznych prowadzących do otrzymywania rozwiązań ścisłych danych problemów. Z czasem jednak, przy powstawaniu coraz to bardziej skomplikowanych teorii opisujących zjawiska, problemy te stawały się na tyle złożone, iż ich rozwiązywanie ściśle było albo bardzo czasochłonne albo też zgoła niemożliwe. Numeryka pozwalała znajdować przybliżone rozwiązania z żadaną dokładnością. Ich podstawową zaletą była ogólność tak formułowanych algorytmów, tzn. w ramach danego zagadnienia nie miało znaczenia czy było ono proste czy też bardzo skomplikowane (najwyżej wiązało się z większym nakładem pracy obliczeniowej). Natomiast wadą była czasochłonność. Stąd prawdziwy renesans metod numerycznych nastąpił wraz z powszechnym użyciem w pracy naukowej maszyn cyfrowych, a w szczególności mikrokomputerów [22]. Dziś dziesiątki żmudnych dla człowieka operacji arytmetycznych wykonuje komputer, jednak złożoność obliczeniowa algorytmów uczących i modeli statystycznych stała się krytycznym czynnikiem ograniczającym w sytuacjach, gdy rozważane są duże zbiory danych. Te ograniczenia spowodowały, że w uczeniu maszynowym i modelowaniu statystycznym wielkiej skali zaczęto wykorzystywać algorytmy **stochastycznego spadku gradientu**. W poniższym rozdziale przedstawione są klasyczne algorytmy spadku wzdłuż gradientu Cauchy’ego oraz Raphsona-Newtona. Następnie omówiony jest algorytm stochastycznego spadku wzdłuż gradientu, którego wykorzystanie do estymacji współczynników w modelu Coxa jest kluczowym celem tej pracy. Algorytm stochastycznego spadku gradientu to metoda optymalizacji wzdłuż spadku gradientu wykorzystywana w sytuacjach, gdy rozważaną funkcję można zapisać jako sumę różniczkowalnych składników. Ponadto przedstawiono również zalety algorytmów stochastycznego spadku gradientu, które przemawiają za atrakcyjnością i popularnością tego typu rozwiązania. Ostatecznie przedyskutowano asymptotyczną efektywność estymatorów uzyskanych dzięki jednemu przejściu po zbiorze, zwanym *epoką*. Definicje i pojęcia w tym rozdziale pochodzą z [4], [6], [20] i [14].

3.1. Algorytmy spadku wzdłuż gradientu

Poniższy rozdział przedstawia popularne iteracyjne algorytmy wyznaczania przybliżonej wartości miejsca zerowego funkcji oraz rozważaną w pracy metodę stochastycznego spadku gradientu. Szukanie miejsc zerowych funkcji jest przydatne w problemach optymalizacyjnych, gdy celem jest znalezienie pierwiastka pochodnych badanej funkcji. Dodatkowo takie algorytmy wykorzystywane są do rozwiązywania (nieliniowych) układów równań. Metody iteracyjne składają się zazwyczaj z k kroków bądź są zatrzymywane, gdy osiągnięty zostanie warunek stopu, czyli gdy odległość pomiędzy kolejnymi przybliżeniami jest dość mała $\|w_{k+1} - w_k\| < \epsilon$ lub wartość gradientu funkcji w wyznaczonym punkcie jest bliska wektorowemu zerowemu $\|\nabla_Q(\mathbf{w}_k)\| \leq \epsilon$ (test stacjonarności), gdzie ϵ to zadana z góry precyzja. Metoda stochastycznego spadku wzdłuż gradientu zakłada, że minimalizowaną funkcję $Q(w)$ można przedstawić jako różniczkowalną sumę jej składników $Q(w) = \sum_{i=1}^n Q_i(w)$. W poniższych algorytmach α_k oznacza długość kroku algorytmu.

Metoda spadku wzdłuż gradientu I (Cauchy'ego)

Minimalizacja funkcji $Q(w)$:

- Zaczynamy od wybranego rozwiązania startowego, np. $w_0 = 0$.
- Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wyznaczamy gradient w punkcie w_{k-1} , $\nabla_Q(w_{k-1})$.
 - Robimy krok wzdłuż negatywnego gradientu:

$$w_k = w_{k-1} - \alpha_k \nabla_Q(w_{k-1}).$$

Metoda spadku wzdłuż gradientu II (Newtona-Raphsona)

Minimalizacja funkcji $Q(w)$:

- Zaczynamy od wybranego rozwiązania startowego, np. $w_0 = 0$.
- Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wyznaczamy gradient w punkcie w_{k-1} , $\nabla_Q(w_{k-1})$ i odwrotność Hessianu $(D_Q^2(w_{k-1}))^{-1}$.
 - Robimy krok wzdłuż negatywnego gradientu z zadany krok przez Hessian:

$$w_k = w_{k-1} - (D_Q^2(w_{k-1}))^{-1} \nabla_Q(w_{k-1}).$$

Metoda stochastycznego spadku wzdłuż gradientu I

Minimalizacja funkcji $Q(w)$:

- Zaczynamy od wybranego rozwiązania startowego, np. $w_0 = 0$.
- Dla $k = 1, 2, \dots$ aż do zbieżności
 - Wylosuj $i \in \{1, \dots, n\}$
 - Wyznaczamy gradient funkcji Q_i w punkcie w_{k-1} , $\nabla_{Q_i}(w_{k-1})$.
 - Robimy krok wzdłuż negatywnego gradientu:

$$w_k = w_{k-1} - \alpha_k \nabla_{Q_i}(w_{k-1}).$$

3.2. Algorytm stochastycznego spadku wzdłuż gradientu I

Stochastyczny spadek gradientu to popularny algorytm wykorzystywany do estymacji współczynników w szerokiej gamie modeli uczenia maszynowego takich jak maszyny wektorów podporządkowanych (*ang. Support Vector Machines*), regresja logistyczna czy modele graficzne [11]. W połączeniu z algorytmem propagacji wstecznej jest standardowym algorytmem w trenowaniu sztucznych sieci neuronowych. Algorytm stochastycznego spadku gradientu był używany już od 1960 przy estymacji współczynników w modelu regresji liniowej, oryginalnie znanym pod nazwą *ADALINE* [35]. Kolejnym znanym popularnym algorytmem wykorzystującym stochastyczny spadek gradientu jest filtr adaptacyjny najmniejszych średnich kwadratów [36] (*ang. least mean squares (LMS) adaptive filter*), który również został wynaleziony w 1960 przez Bernarda Widrowa, który jest także twórcą *ADALINE*.

Idea algorytmu stochastycznego spadku gradientu jest następująca: zamiast obliczać gradient na całej funkcji L , w danym kroku oblicz gradient tylko na pojedynczym elemencie ℓ_i . Nazwa *stochastyczny* bierze się stąd, iż oryginalnie wybiera się element ℓ_i losowo. W praktyce zwykle przechodzi się po całym zbiorze danych w losowej kolejności lub, o ile to możliwe, w kolejności chronologicznej obserwacji.

3.2.1. Właściwości stochastycznego spadku wzdłuż gradientu

Zbieżność algorytmu stochastycznego spadku gradientu była badana szeroko badana w literaturze aproksymacji stochastycznych. Aby uzyskać zbieżność zazwyczaj wymaga się aby ciąg kroków algorytmu α_k był malejący i spełniał poniższe warunki $\sum_k \alpha_k = \infty$ oraz $\sum_k \alpha_k^2 < \infty$ [4]. Twierdzenie Robbinsa-Siegmunda [30] przy łagodnych warunkach zapewnia zbieżność prawie na pewno [5], nawet w sytuacjach gdy optymalizowana funkcja nie jest wszędzie różniczkowalna.

Prędkość zbieżności stochastycznego spadku gradientu jest w rzeczywistości ograniczana przez zgrubną (*ang. noisy*) aproksymację prawdziwego gradientu. Gdy długości kroków algorytmu maleją zbyt wolno, wariancja estymatorów parametrów w_k maleje równie wolno. Gdy kroki algorytmu maleją zbyt szybko, oczekiwane estymatory parametrów w_k potrzebują więcej czasu by osiągnąć optimum [4]. Pod pewnymi warunkami regularności [24], najlepsza prędkość zbieżności jest uzyskana dla kroków algorytmu $\alpha_k \sim k^{-1}$.

Jak wykazano w [10] pod pewnymi odpowiednimi warunkami regularności, gdy zainicjowany współczynnik początkowy w_0 jest wystarczająco blisko optimum i krok algorytmu jest odpowiednio mały, algorytm stochastycznego spadku gradientu osiąga liniową zbieżność. Oznacza to, iż przy spełnieniu założeń metody, odległości pomiędzy kolejnymi przybliżeniami a minimum funkcji \mathbf{w}^* maleją liniowo: $\|\mathbf{w}^* - \mathbf{w}_{k+1}\| \leq c \|\mathbf{w}^* - \mathbf{w}_k\|$. Zbieżność wymaga często przejścia parokrotnie po całym zbiorze danych. Zwykle nie losuje się obserwacji, ale przechodzi się po zbiorze danych w losowej kolejności. Dodatkowo wady i zalety algorytmu wymienione są poniżej. Zalety zdecydowanie przewyższają wady.

Zalety

- **Szybkość**: obliczenie gradientu wymaga wzięcia tylko jednej obserwacji.
- **Skalowalność**: cały zbiór danych nie musi nawet znajdować się w pamięci operacyjnej.
- **Prostota**: gradient funkcji Q_i daje bardzo prosty wzór na modyfikacje wag.

Wady

- **Wolna zbieżność:** czasem gradient stochastyczny zbiega wolno i wymaga wielu iteracji po zbiorze uczącym.
- **Problem z ustaleniem długości kroku k :** wyznaczenie k przez przeszukiwanie liniowe nie przynosi dobrych rezultatów, ponieważ optymalizujemy oryginalnej funkcji Q tylko jej jeden składnik Q_i .

3.3. Ewentualne obrazki obrazujące/porównujące algorytmy

Coś tego typu porównujące może algorytm Cauchy'ego, Raphsona-Newtona i SGD?

Rysunek 3.1: Porównanie algorytmów spadku gradientu.

Rozdział 4

Estymacja w modelu Coxa metodą stochastycznego spadku gradientu

Rozdział 5

Analiza danych genomicznych

5.1. Opis i pobranie danych

5.2. Analiza

Dodatek A

Wykorzystane narzędzia

Dodatek B

Kody w R

Dodatek C

Dokumentacja pakietu RTCGA

R documentation

of all in ‘RTCGA/man’

May 5, 2015

R topics documented:

RTCGA-package	1
availableDataSets	2
availableDates	3
checkDataSetsAvailability	3
checkGenesNamesAvailability	4
downloadTCGA	5
infoTCGA	6
mergeTCGA	6
read.clinical	7
Index	8

RTCGA-package	<i>The Cancer Genome Atlas data integration</i>
---------------	-------------------------------------------------

Description

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes. The key is to understand genomics to improve cancer care. RTCGA package offers download and integration of the variety and volume of TCGA data using patient barcode key, what enables easier data possession. This may have an beneficial influence on impact on development of science and improvement of patients’ treatment. Furthermore, RTCGA package transforms TCGA data to form which is convenient to use in R statistical package. Those data transformations can be a part of statistical analysis pipeline which can be more reproducible with RTCGA

Details

For more detailed information visit **RTCGA** wiki on [Github](#).

Author(s)

Marcin Kosinski [aut, cre] < m.p.kosinski@gmail.com >
Przemysław Biecek [aut] < przemyslaw.biecek@gmail.com >

See Also

Other RTCGA: [availableDataSets](#); [availableDates](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

availableDataSets	<i>TCGA datasets' names</i>
-------------------	-----------------------------

Description

Enables to check TCGA datasets' names for current release date and cohort.

Usage

```
availableDataSets(cancerType, date = NULL)
```

Arguments

cancerType	A character of length 1 containing abbreviation (Cohort code) of types of cancers to check for available datasets' names on http://gdac.broadinstitute.org/ .
date	A NULL or character specifying from which date datasets' names should be checked. By default (date = NULL) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format "YYYY-MM-DD".

Value

A vector of available datasets' names to pass to the [downloadTCGA](#) function.

See Also

Other RTCGA: [RTCGA-package](#); [availableDates](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Examples

```
## Not run:
availableDataSets( "BRCA" )
availableDataSets( "OV", availableDates()[5] ) # error

## End(Not run)
```

availableDates	<i>TCGA datasets' releases dates</i>
----------------	--------------------------------------

Description

Enables to check dates of TCGA datasets' releases.

Usage

```
availableDates()
```

Value

A vector of available dates to pass to the [downloadTCGA](#) function.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Examples

```
## Not run:
availableDates()

## End(Not run)
```

checkDataSetsAvailability	<i>TCGA datasets' names availability</i>
---------------------------	------------------------------------------

Description

Enables to check TCGA datasets' names availability for current release date and cancer type.

Usage

```
checkDataSetsAvailability(cancerTypes, pattern = "Merge_Clinical.Level_1",
  date = NULL)
```

Arguments

cancerTypes	A character vector containing abbreviation (Cohort code) of types of cancers to check for availability of datasets' name on http://gdac.broadinstitute.org/ .
pattern	A character vector of length 1 containing a part of a dataset's name to be checked for availability for current date parameter. By default phrase "Merge_Clinical.Level_1" is checked.
date	A NULL or character specifying from which date datasets' names should be checked for availability. By default (date = NULL) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format "YYYY-MM-DD".

Value

A vector of available datasets names to pass to the [downloadTCGA](#) function.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Examples

```
## Not run:
checkDataSetsAvailability( "BRCA" )
checkDataSetsAvailability( c("BRCA", "OV") )
checkDataSetsAvailability( "BRCA", "Mutation_Packager_Calls.Level" )

## End(Not run)
```

checkGenesNamesAvailability

TCGA genes' names and availability in Merge_rnaseqv2__... dataset.

Description

availableGenesNames returns all available genes' names from genes' expressions dataset, where checkGenesNamesAvailability checks whether genes specified in genes are available in Merge_rnaseqv2__illumina (genes' expressions) dataset.

Usage

```
checkGenesNamesAvailability(rnaseqDir, genes)

availableGenesNames(rnaseqDir)
```

Arguments

rnaseqDir	A directory to a cancerType.rnaseqv2__illumina.rnaseqv2__unc_edu__Level_3__RSEM file
genes	A character - which genes to check for availability in a dataset.

Value

A vector containing genes' names that matched existing names.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Examples

```
## Not run:
  checkGenesNamesAvailability( rnaseqDir, "TP53" )

## End(Not run)
```

downloadTCGA	<i>Download TCGA data</i>
--------------	---------------------------

Description

Enables to download TCGA data from specified dates of releases of concrete Cohorts of cancer types. Pass a name of required dataset to the `dataSet` parameter. By default the Merged Clinical `dataSet` is downloaded (value `dataSet = "Merge_Clinical.Level_1"`) from the newest available date of release.

Usage

```
downloadTCGA(cancerTypes, dataSet = "Merge_Clinical.Level_1", destDir,
  date = NULL)
```

Arguments

<code>cancerTypes</code>	A character vector containing abbreviations (Cohort code) of types of cancers to download from http://gdac.broadinstitute.org/ .
<code>dataSet</code>	A part of the name of <code>dataSet</code> to be downloaded from http://gdac.broadinstitute.org/runs/ . By default the Merged Clinical <code>dataSet</code> is downloaded (value <code>dataSet = "Merge_Clinical.Level_1"</code>). Available datasets' names can be checked using availableDataSets function.
<code>destDir</code>	A character specifying a directory into which <code>dataSets</code> will be downloaded.
<code>date</code>	A NULL or character specifying from which date <code>dataSets</code> should be downloaded. By default (<code>date = NULL</code>) the newest available date is used. All available dates can be checked on http://gdac.broadinstitute.org/runs/ or by using availableDates function. Required format "YYYY-MM-DD".

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [infoTCGA](#); [mergeTCGA](#); [read.clinical](#)

Examples

```
## Not run:

dir.create( "hre")

downloadTCGA( cancerTypes = "BRCA", dataSet = "miR_gene_expression",
  destDir = "hre/" )

downloadTCGA( cancerTypes = c("BRCA", "OV"), destDir = "hre/" )

## End(Not run)
```

infoTCGA

Information about cohorts from TCGA project

Description

Function restores codes and counts for each cohort from TCGA project.

Usage

```
infoTCGA()
```

Value

A list with a tabular information from <http://gdac.broadinstitute.org/>.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [availableGenesNames](#), [checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [mergeTCGA](#); [read.clinical](#)

mergeTCGA

Merge Clinical data with genes' Mutations and Expressions data

Description

mergeTCGA enables to

Usage

```
mergeTCGA(clinicalDir, rnaseqDir = NULL, mutationDir = NULL, genes)
```

Arguments

clinicalDir A directory to a cancerType.clin.merged.txt file. cancerType might be BRCA, OV etc. Can be checked using [infoTCGA](#) function.

rnaseqDir A directory to a cancerType.rnaseqv2__illuminahisec_rnaseqv2__unc_edu__Level_3__RSEM file, which is a set with gene's Expressions.

mutationDir A directory to a Mutation_Packager_Calls.Level1 folder where are genes' Mutations files.

genes For rnaseqDir - which genes' expressions to merge with clinical data in clinicalDir. For mutationDir which gene's mutations to merge with clinical data in clinicalDir.

Value

A cancerType.clin.merged.txt file is updated with newline containing informations about genes passed to genes argument.

Note

Original cancerType.clin.merged.txt file will be changed after performing merge operation.
Only one of rnaseqDir and mutationDir can be used at a time.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [availableGenesNames](#),
[checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [read.clinical](#)

read.clinical	<i>Read from txt fo,e</i>
---------------	---------------------------

Description

TODO

Usage

```
read.clinical(clinicalDir, ...)
```

Arguments

clinicalDir	A directory to a cancerType.clin.merged.txt file. cancerType might be BRCA, OV etc.
-------------	-------------------------------------------------------------------------------------

Value

A data.frame with clinical data.

See Also

Other RTCGA: [RTCGA-package](#); [availableDataSets](#); [availableDates](#); [availableGenesNames](#),
[checkGenesNamesAvailability](#); [checkDataSetsAvailability](#); [downloadTCGA](#); [infoTCGA](#); [mergeTCGA](#)

Index

availableDataSets, [2](#), [2](#), [3–7](#)
availableDates, [2](#), [3](#), [3](#), [4–7](#)
availableGenesNames, [2–7](#)
availableGenesNames
 (checkGenesNamesAvailability),
 [4](#)

checkDataSetsAvailability, [2](#), [3](#), [3](#), [4–7](#)
checkGenesNamesAvailability, [2–4](#), [4](#), [5–7](#)

downloadTCGA, [2–4](#), [5](#), [6](#), [7](#)

infoTCGA, [2–6](#), [6](#), [7](#)

mergeTCGA, [2–6](#), [6](#), [7](#)

read.clinical, [2–7](#), [7](#)
RTCGA-package, [1](#)

Literatura

- [1] Aldrich J., (1997) *R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922*, Statistical Science 1997, Vol. 12, No. 3, 162-176.
- [2] Asselain B., Mould R. F., (2010) *Methodology of the Cox proportional hazards model*, Journal of Oncology 2010, volume 60, Number 5, 403–409.
- [3] Biecek P., (2011) *Przewodnik po pakiecie R*, Rozprawa doktorska, Oficyna Wydawnicza GiS, wydanie II.
- [4] Bottou L., (2010) *Large-Scale Machine Learning with Stochastic Gradient Descent*.
- [5] Bottou L., (1998), *Online Learning and Stochastic Approximations*.
- [6] Bottou L., (2012) *Stochastic Gradient Descent Tricks*.
- [7] Burzykowski T., (2015?) *Notatki do przedmiotu Biostatystyka*, <https://e.mini.pw.edu.pl/sites/default/files/biostatystyka.pdf>.
- [8] Cox D. R. (1962) *Renewal Theory*. Methuen Monograph on Applied Probability & Statistics, London: Methuen.
- [9] Cox D. R., (1972) *Regression models and life-tables (with discussion)*, Journal of the Royal Statistical Society Series B 34:187-220..
- [10] Dennis J. E. Jr., Schnabel R. B., (1983), *Numerical Methods For Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall.
- [11] Finkel J. R., Kleeman A., Manning C. D., (2008), *Efficient, Feature-based, Conditional Random Field Parsing*, Proc. Annual Meeting of the ACL.
- [12] Fisher R. A., (1912) *An absolute criterion for fitting frequency curves*.
- [13] Fisher R. A., (1922) *On the mathematical foundations of theoretical statistics*, Philos. Trans. Roy. Soc. London Ser. A 222 309-368.
- [14] Fortuna Z., Macukow B., Wąsowski J., (2006), *Metody Numeryczne*, Wydawnictwa Naukowo-Techniczne.
- [15] Gauss C. F., (1809) *Theoria Motus Corporum Coelestium*.
- [16] Gągolewski M., (2014) *Programowanie w języku R*, Wydawnictwo Naukowe PWN.
- [17] Hald A., (1949) *Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point*, Skandinavisk Aktuarietidskrift, 119-134.
- [18] Hutchinson J. B., (1928) *The Application of the "Method of Maximum Likelihood" to the Estimation of Linkage*, Genetics. 1929 Nov; 14(6): 519–537.

- [19] Kenward M. G., Lesaffre E. and Molenberghs G., (1994) *An Application of Maximum Likelihood and Generalized Estimating Equations to the Analysis of Ordinal Data from a Longitudinal Study with Cases Missing at Random*, Biometrics Vol. 50, No. 4 (Dec., 1994), pp. 945-953.
- [20] Kotłowski W., (2012), Notatki do przedmiotu *Techniki Optymalizacji* prowadzonego na Politechnice Poznańskiej,
<http://www.cs.put.poznan.pl/wkotlowski/teaching/wyklad3b.pdf>
- [21] Legendre A. M., (1804) *Nouvelles méthodes pour la détermination des orbites des comètes*.
- [22] Milewski S., (2006) Konspekt do przedmiotu *Metody Numeryczne* prowadzonego na Politechnice Krakowskiej,
http://15.pk.edu.pl/images/skrypty/Metody_numeryczne_1
- [23] Millar R. B., (2011) *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB, chapter 6. Some Widely Used Applications of Maximum Likelihood*, John Wiley & Sons, Ltd.
- [24] Murata N., (1998), *A Statistical Study of On-line Learning*. In *Online Learning and Neural Networks*, Cambridge University Press.
- [25] Niemi W., (2011) Skrypt do przedmiotu *Statystyka* prowadzonego na Uniwersytecie Warszawskim,
<http://www-users.mat.umk.pl/~wniem/Statystyka/Statystyka.pdf>
- [26] Norwegian Multicentre Study Group, (1981) *Timolol-induced reduction in mortality and reinfarction*, The New England Journal of Medicine; 304: 801-7.
- [27] Panchenko D., (2006), Notatki do otwartego kursu MIT *Statistics for Applications, Lecture 2: Maximum Likelihood Estimators.*,
<http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/>
- [28] Panchenko D., (2006), Notatki do otwartego kursu MIT *Statistics for Applications, Lecture 3: Properties of MLE: consistency, asymptotic normality. Fisher information.*,
<http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/>
- [29] R Core Team, (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wiedeń , ISBN 3-900051-07-0,
<http://www.R-project.org/>.
- [30] Robbins H. E., Siegmund. D. O., (1971), *A convergence theorem for non negative almost supermartingales and some applications*, In Proc. Sympos. Optimizing Methods in Statistics, pages 233–257, Ohio State University. Academic Press, New York.
- [31] Rydlewski J., (2009) *Estymatory Największej Wiarogodności w Uogólnionych Modelach Regresji Nieliniowej*, Rozprawa doktorska.
- [32] Sokołowski A., (2010) *Jak rozumieć i wykonywać analizę przeżycia*
http://www.statsoft.pl/Portals/0/Downloads/Jak_rozumiec_i_wykonac_analize_przezycia.pdf
- [33] Statistics Views, (2014) *"I would like to think of myself as a scientist, who happens largely to specialise in the use of statistics"*– An interview with Sir David Cox.
- [34] Therneau T. M., Grambsch P. M., (2000), *Modeling Survival Data: Extending the Cox Model*, Springer.

-
- [35] Widrow B, (1960), *An adaptive "ADALINE" neuron using chemical "memistors"*, Technical Report No. 1553-2, Stanford University.
 - [36] Widrow B, Stearns S. D., (1985), *Adaptive Signal Processing*, Prentice Hall.
 - [37] Wikipedia, encyklopedia wolnego dostępu wikipedia.pl
 - [38] Woodcock S, (2014), Notatki do otwartego kursu Uniwersytetu Simona Frasera *ECON 837, Lecture 11 Asymptotic Properties of Maximum Likelihood Estimators*, <http://www.sfu.ca/~swoodcoc/teaching/sp2014/econ837/11.mle.pdf>
 - [39] Zieliński R., (1990) *Siedem wykładów wprowadzających do statystyki matematycznej*, Warszawa, Wydawnictwo Naukowe PWN.

Marcin Piotr Kosiński
Nr albumu 265361

Warszawa, 2 września 2015

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Estymacja w modelu Coxa metodą stochastycznego spadku gradientu z przykładami zastosowań w analizie danych z The Cancer Genome Atlas”, której promotorem jest prof. ndzw. dr hab. inż Przemysław Biecek wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

.....

Marcin Piotr Kosiński