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

Spis rysunków

Wprowadzenie

+++ Zlagodzenie zalozen o proporcjonalnych hazardach Przykład leukemii, który wykorzystał Cox [1] jako ilustrację modelu, podobnie jak przykład nieoperacyjnego raka piersi w III stopniu zaawansowania [13], nie są szczególnie dobre, ale założenie o proporcjonalności ryzyka jest do zaakceptowania. Jednakże w przypadku danych onkologicznych ze znaczącą liczbą długoletnich przeżyć, założenie stałości ryzyka względnego w czasie może już nie być rozsądne. Taka sytuację zaobserwowali Gore & Pocock [14] dla chorych na raka piersi, których wyniki odległe opublikował Langlands i inni [15]. Stwierdzili oni, że dla tych danych założenie proporcjonalności ryzyka nie było spełnione. Ponadto w komentarzu Gore zauważył, że stopień zaawansowania, który ma początkowo znaczenie, traci je w czasie i po 10 latach obserwacji roczna śmiertelność jest już od niego niezależna. Problem niespełnienia założenia proporcjonalności można próbować rozwiązać na różne sposoby. Można, na przykład, uwzględnić w modelu czynniki prognostyczne zależne od czasu, wtedy: lambda(t, Z(t)) = lambda0(t)exp[Z(t)'] Iloczyn exp[Z(t)beta'] będąc funkcją czasu, nie jest już stały i współczynnik proporcjonalności także się zmienia. To prowadzi do mniej eleganckiego, ale być może bardziej realistycznego spojrzenia na historię choroby, w której wpływ czynników prognostycznych zmienia się wraz z czasem obserwacji. +++

+++ 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 najszerzej 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 [5], 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 [44].

Definicja 0.1. ***Model statystyczny*** *określamy przez podanie rodziny* $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *rozkładów prawdopodobieństwa na przestrzeni próbkowej* Ω *oraz zmiennej losowej* $X : \Omega \to \mathcal{X}$*, 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. ***Przestrzenią 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 próba 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 [62].

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 $X = (X_1, \ldots, X_n)$, to odwzorowanie mierzalne $T : \mathcal{X} \to \mathcal{R}$.*

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 [44], gdy rozmiar próbki $X_1, \ldots, X_n$ 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 opisane 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 $\hat{g}(X_1, \ldots, X_n)$ wielkości $g(\theta)$ jest* ***nieobciążony****, jeśli dla każdego n*

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

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

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

dla każdego $\varepsilon > 0$.

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

$$\mathbb{P}_\theta\Big(\lim_{n\to\infty} \hat{g}(X_1, \ldots, X_n) = g(\theta)\Big) = 1.$$

Zgodność (mocna zgodność) znaczy tyle, że

$$\hat{g}(X_1, \ldots, X_n) \to g(\theta), \qquad (n \to \infty)$$

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

$\xrightarrow{d}$ *oznacza zbieżność wg rozkładu.*

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

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

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

Inaczej mówiąc, estymator jest asymptotycznie normalny, gdy:

$$\lim_{n\to\infty} \mathbb{P}_\theta\Big(\frac{\sqrt{n}}{\sigma(\theta)}(\hat{g}(X_1, \ldots, X_n) - g(\theta) \leq a\Big) = \Phi(a),$$

gdzie $\Phi(x)$ to dystrybuanta 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

$$\mathbb{P}_\theta\Big(\frac{\sqrt{n}}{\sigma(\theta)}(\hat{g}(X_1, \ldots, X_n) - g(\theta) \leq a\Big) - \Phi(a) = f(n) \in \eth(\frac{1}{\sqrt{n}}).$$

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

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

1.2. Metoda największej wiarogodności

Metodę największej wiarogodności wprowadził R. A. Fisher w 1922 r. [21], dla której po raz pierwszy procedurę numeryczną zaproponował już w 1912 r. [20]. O burzliwym procesie powstawania metody, o zmianach w jej uzasadnieniu, o koncepcjach, które powstały w obrębie tej metody takich jak parametr, statystyka, wiarogodność, 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 [38], [23], była rozwijana i szeroko stosowana później przez wielu statystyków, i wciąż znajduje obszerne zastosowania w wielu obszarach estymacji statystycznej, np. [27], [31], [41].

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

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

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

którą rozważamy jako funkcję parametru θ *przy ustalonych wartościach obserwacji* $x_1, \ldots, x_n$*, gdzie*

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

Oznacza to, że wiarogodność 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 wiarogodności*** *parametru* θ*, oznaczanym ENW(*θ*), nazywamy wartość parametru, w której funkcja wiarogodności 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 wiarogodności, supremum zastępują wartością największą [51], [63], [46].

1.3. Asymptotyczne własności estymatora największej wiarogodności

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 [47] i [63].

Zgodność estymatora największej wiarogodności

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

Definicja 1.9. ***Funkcja log-wiarogodności*** *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 wiarogodności, 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

$\mathbb{E}_{\theta_0}$ oznacza wartość oczekiwaną względem rozkładu parametryzowanego przez θ_0.

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

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 wiarogodności* $ENW(\theta)$ *jest zgodny, tzn.*

$$ENW(\theta) \to \theta \quad dla \quad n \to \infty.$$

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 $X_1, \ldots, X_n$), 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

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

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

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 wiarogodności.

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

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

gdzie odpowiednio

$$\mathcal{I}(\theta) = \int \Big(\frac{\partial}{\partial\theta}\log f_\theta(x)\Big)^2 f_\theta(x)dx \qquad \textit{dla zmiennej ciągłej;}$$
$$\mathcal{I}(\theta) = \sum_x \Big(\frac{\partial}{\partial\theta}\log f_\theta(x)\Big)^2 \mathbb{P}_\theta(X = x) \;\; \textit{dla zmiennej dyskretnej.}$$

W dowodzie asymptotycznej normalności estymatora największej wiarogodności 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 modułem 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 $\int \ldots dx$.

Wprowadzając takie założenia, otrzymano przydatne właściwości Informacji Fishera.

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

$$(i)\quad \mathbb{E}_\theta\frac{\partial}{\partial\theta}\log f_\theta(x) = 0,$$
$$(ii)\quad \mathcal{I}(\theta) = \mathbb{V}ar_\theta\Big(\frac{\partial}{\partial\theta}\log f_\theta(X)\Big),$$
$$(iii)\quad \mathcal{I}(\theta) = -\mathbb{E}_\theta\Big(\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\Big).$$

Dowód tego stwierdzenia można znaleźć w [44].

Patrząc na postać pochodnej funkcji log-wiarogodności

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

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 bliski do θ_0, a co za tym idzie, dużo ciężej 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 wiarogodności jest asymptotycznie normalny,*

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

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

$$\sqrt{n}(ENW(\theta) - \theta_0) = -\frac{\sqrt{n}\ell_n'(\theta_0)}{\ell_n''(\theta_1)}. \tag{1.2}$$

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

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

a to można wstawić do licznika w równaniu (1.2)

$$\begin{aligned}\sqrt{n}\ell_n'(\theta_0) &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta_0) - 0\Big) \\ &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta_0) - \mathbb{E}_{\theta_0}\ell'(\theta_0)\Big) \to \mathcal{N}\Big(0, \mathbb{V}ar_{\theta_0}(\ell'(\theta_0))\Big),\end{aligned} \tag{1.4}$$

gdzie zbieżność wynika z Centralnego Twierdzenia Granicznego.

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

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

z Prawa Wielkich Liczb.

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

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

z punktu (*iii*) ze Stwierdzenia (1.3).

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

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

Ostatecznie wariancja

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

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 [13], [56], [2] i [9].

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*

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

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

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

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

W tych definicjach T^* oznacza czas do wystąpienia zdarzenia. Zakłada się, że wewnątrz każdej grupy $j = 1, 2, \ldots, m$ wyznaczonej przez poziomy zmiennych objaśniających, czasy $T^*_{j,i}$ dla $i = 1, \ldots, n_j$ 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 [12], w której analizuje się awaryjność elementów przemysłowych.

Model proporcjonalnych hazardów Coxa [13] 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 [45].

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

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

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. ***Cenzuorwanie 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*

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

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*

$$g(t; \theta, \phi) \equiv g(t; \phi), \tag{2.4}$$

gdzie $g(t; \theta, \phi)$ jest funkcją gęstości dla cenzurowań C_i wyrażonych jako niezależne zmienne losowe o jednakowym rozkładzie, zaś prawdzie 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, ponieaż nie zależy od parametrów od których zależny jest hazard.

Model proporcjonalnych hazardów Coxa oparty jest na założeniach:

i) Współczynniki modelu $\beta_k, k = 1, \cdots, p$ 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 - 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 wiarogodności, którą wprowadził Cox w 1972 r. [13].

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

$$\frac{e^{X_i'\beta}}{\sum\limits_{l\in\mathscr{R}(t_i)} e^{X_l'\beta}}, \tag{2.5}$$

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

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

$$L(\beta,\varphi) = L_p(\beta)\cdot L^*(\beta,\varphi), \tag{2.6}$$

gdzie, dla $\lambda(t)$ wprowadzonego w definicji (2.2)

$$L_p(\beta) = \prod_{i=1}^{n} f(t_i;\beta)^{\delta_i} S(t_i;\beta)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(t_i;\beta)^{\delta_i} S(t_i;\beta) \tag{2.7}$$

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

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

$$L_p(\beta) = \prod_{i=1}^{K} \frac{e^{X_i'\beta}}{\sum\limits_{l=1}^{n} Y_l(t_i) e^{X_l'\beta}}, \tag{2.8}$$

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 wiarogodności 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 wiarogodności, 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 wiarogodności 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 wiarogodności nie jest funkcją wiarogodności w sensie bycia proporcjonalną do prawdopodobieństwa obserwowanego zbioru, można ją traktować jako funkcję wiarogodności dla celów asymptotycznego wnioskowania o współczynnikach modelu, zobacz [**?**].

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-wiarogodności**

$$\ell_p(\beta) = \sum_{i=1}^{K} X_i'\beta - \sum_{i=1}^{K} \log\Big(\sum_{l\in\mathscr{R}(t_i)} e^{X_l'\beta}\Big). \tag{2.9}$$

Analityczne obliczenia dają p-wymiarowy wektor pochodnych, dla $k = 1, \ldots, p$

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

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

gdzie czynnik

$$A_{ik} = \frac{\sum\limits_{l\in\mathscr{R}(t_i)} X_{lk} e^{X_l'\beta}}{\sum\limits_{l\in\mathscr{R}(t_i)} e^{X_l'\beta}} \tag{2.11}$$

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

Z kolei drugie pochodne cząstkowe, jak podaje [13], mają postać dla $k_1, k_2 = 1, \ldots, p$

$$\mathscr{I}_{k_1k_2}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial\beta_{k_1}\partial\beta_{k_2}} = \sum_{i=1}^{K} C_{ik_1k_2}(\beta), \tag{2.12}$$

gdzie

$$C_{ik_1k_2}(\beta) = \frac{\sum\limits_{l\in\mathscr{R}(t_i)} X_{lk_1}X_{lk_2}e^{X_l'\beta}}{\sum\limits_{l\in\mathscr{R}(t_i)} e^{X_l'\beta}} - A_{ik_1}(\beta)A_{ik_2}(\beta) \tag{2.13}$$

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

Estymator największej wiarogodności 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 3.1. 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.

2.4. Generowanie danych dla modelu Coxa

W celu skutecznej diagnostyki procesu estymacji w modelu Coxa, należy wiedzieć jak dane dla tego modelu można generować, aby ów model symulować dla znanych współczynników.

Poniższy rozdział przedstawia metodę odwrotnych prawdopodobieństw opisaną szerzej w [3], dzięki której można wygenerować czasy zdarzeń dla zadanej z góry funkcji hazardu i zmiennych objaśniających. Ta metoda posłuży w rozdziale 4 do wygenerowania danych w celu weryfikacji jakości procesu numerycznej estymacji współczynników modelu, w sytuacji gdy wykorzystywany jest algorytm stochastycznego spadku gradientu, który jest opisany w rozdziale 3.

Mówiąc o funkcji przeżycia warto wprowadzić skumulowaną funkcję hazardu.

Definicja 2.7. ***Skumulowaną funkcją hazardu*** *nazywa się funkcję spełniającą zależność*

$$H(t) = \int_0^t \lambda(u)du, \tag{2.14}$$

gdzie $\lambda(u)$ to pewna funkcja hazardu, o której mówi definicja (2.2).

Wtedy dla zdefiniowanej w (2.3) funkcji hazardu funkcja przeżycia i jej dopełnienie (dystrybuanta) dla modelu Coxa proporcjonalnych hazardów wygląda następująco

$$S(t|x) = e^{-H_0(t)\cdot e^{x'\beta}} \tag{2.15}$$

$$F(t|x) = 1 - e^{-H_0(t)\cdot e^{x'\beta}} \tag{2.16}$$

gdzie $H_0(t)$ to bazowa skumulowana funkcja hazardu.

Niech Y będzie zmienną losową o dystrybuancie zadanej w (2.16), wtedy jak wykazano w [42] zmienna losowa $U = F(Y)$ pochodzi z rozkładu jednostajnego $U \sim \mathcal{U}([0,1])$. Zachodzi to również dla zmiennej losowej $1-U \sim \mathcal{U}([0,1])$. Dodatkowo, niech T będzie czasem przeżycia w modelu Coxa (definicja 2.3), wtedy z (2.16) wynika

$$U = e^{-H_0(T)\cdot e^{x'\beta}} \sim \mathcal{U}([0,1]). \tag{2.17}$$

Jeżeli $\lambda_0(t) > 0$ dla każdego t, to $H_0(t)$ jest dodatnia i można mówić o jej odwrotności, zaś czas przeżycia T dla modelu Coxa może być wyrażony przez

$$T = H_0^{-1}(-\log(U)\cdot e^{-x'\beta}), \tag{2.18}$$

gdzie $U \sim \mathcal{U}([0,1])$.

Przykład symulacji dla rozkładu Weibulla

Równanie (2.18) jest odpowiednie do generowania czasów do zdarzenia w modelu Coxa, gdy potrafi się odpowiednio generować zmienne z rozkładu $\mathcal{U}([0,1])$. Jest to możliwe w większości pakietów statystycznych. Poniżej przedstawiony jest kod w języku $\mathcal{R}$ [49], dzięki któremu możliwe jest generowanie czasów zdarzeń pochodzących z rozkładu Weibulla [14], dla którego funkcja hazardu ma postać

$$\lambda_0(t) = \lambda\rho t^{\rho-1}, \tag{2.19}$$

gdzie $\lambda > 0$ to parametr skali, zaś $\rho > 0$ to parametr kształtu.

Z (2.14) wynika, że bazowa skumulowana funkcja hazardu dla rozkładu Weibulla wynosi

$$H_0(t) = \int_0^t \lambda \rho u^{\rho-1} du = \lambda t^{\rho}, \tag{2.20}$$

zaś jej funkcja przeciwna to

$$H_0^{-1}(t) = (\lambda^{-1} t)^{-\rho}. \tag{2.21}$$

Wtedy podstawiając (2.21) do (2.18) można otrzymać

$$T = (\lambda^{-1} \cdot (-\log(U)) \cdot e^{-x'\beta})^{-\rho} = \Big(-\frac{\log(U)}{\lambda e^{x'\beta}} \Big)^{-\rho}. \tag{2.22}$$

Wynik ten oznacza, że T czyli odpowiadające czasy zdarzeń pochodzą z warunkowego rozkładu Weibulla (warunkowanego przez x) o parametrach kształtu ρ i skali $\lambda e^{x'\beta}$.

Dzięki tym rozważaniom, możliwe było stworzenie kodu generującego czasy i indykatory zdarzeń dla zadanych z góry współczynników modelu i zmiennych objaśniających. Poniższy kod i jego rezultat posłużą w rozdziale 4.2 do diagnostyki procesu estymacji w modelu Coxa w przypadku wykorzystania algorytmu stochastycznego spadku gradientu.

```
dataCox <- function(N, lambda, rho, x, beta, censRate){

  # real Weibull times
  u <- runif(N)
  Treal <- (- log(u) / (lambda * exp(x %*% beta)))^(1 / rho)

  # censoring times
  Censoring <- rexp(N, censRate)

  # follow-up times and event indicators
  time <- pmin(Treal, Censoring)
  status <- as.numeric(Treal <= Censoring)

  # data set
  data.frame(id=1:N, time=time, status=status, x=x)
}

x <- matrix(sample(0:1, size = 40, replace = TRUE), ncol = 2)

head(dataCox(20, 3, 2, x, beta = c(2,3), 5))
```

```
  id       time status x.1 x.2
1  1 0.01193626      0   0   1
2  2 0.03567485      0   1   1
3  3 0.13330012      1   0   1
4  4 0.04358821      1   1   1
5  5 0.03825366      1   1   1
6  6 0.29355955      1   1   0
```

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 ścisłe było albo bardzo czasochłonne albo też zgoła niemożliwe. Numeryka pozwalała znajdywać przybliżone rozwiązania z żądaną 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 [40]. 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 [5], [7], [36] i [22].

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 wektorowymi zerowemu $\| \nabla_Q(\mathbf{w_k}) \| \leqslant \epsilon$ (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 $k = 1, 2, \ldots$ 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 $k = 1, 2, \ldots$ aż do zbieżności
  - Wyznaczamy gradient w punkcie $w_{k-1}, \nabla_Q(w_{k-1})$ i odwrotność Hesjanu $(D^2_Q(w_{k-1}))^{-1}$.
  - Robimy krok wzdłuż negatywnego gradientu z zadanym krokiem przez Hesjan:

$$w_k = w_{k-1} - (D^2_Q(w_{k-1}))^{-1} \nabla_Q(w_{k-1}). \tag{3.1}$$

Metoda stochastycznego spadku wzdłuż gradientu I

Minimalizacja funkcji $Q(w)$:

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

$$w_k = w_{k-1} - \alpha_k \nabla_{Q_i}(w_{k-1}). \tag{3.2}$$

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 podpierających (*ang. Support Vector Machines*), regresja logistyczna czy modele graficzne [19]. 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 jako *ADALINE* [59]. Kolejnym algorytmem wykorzystującym stochastyczny spadek gradientu jest filtr adaptacyjny najmniejszych średnich kwadratów [61] (*ang. least mean squares (LMS) adaptive filter*), który został wynaleziony przez Bernarda Widrowa, twórcę *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.

Właściwości stochastycznego spadku wzdłuż gradientu

Zbieżność algorytmu stochastycznego spadku gradientu była 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 $\sum_k \alpha_k^2 < \infty$ [5]. Twierdzenie Robbinsa-Siegmunda [50] przy łagodnych warunkach zapewnia zbieżność prawie na pewno [6], nawet gdy optymalizowana funkcja nie jest wszędzie różniczkowalna.

Prędkość zbieżności stochastycznego spadku gradientu jest w rzeczywistości ograniczana przez zgrubną (*ang. noisy*) aproksymacje 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 [5]. Pod pewnymi warunkami regularności [43], najlepsza prędkość zbieżności jest uzyskana dla kroków algorytmu $\alpha_k \sim k^{-1}$.

Jak wykazano w [16] 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: $\| \mathbf{w}^* - \mathbf{w_{k+1}} \| \leqslant c \| \mathbf{w}^* - \mathbf{w_k} \|$. Zbieżność wymaga często przejścia parokrotnie po całym zbiorze danych. Wady i zalety algorytmu wymienione są poniżej. Zalety zdecydowanie przewyższają wady.

Zalety

- Szybkość kroku: 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 nie optymalizujemy oryginalnej funkcji Q tylko jej jeden składnik Q_i.

3.3. Porównanie algorytmów spadku wzdłuż gradientu

W niniejszym podrozdziale przedstawiono graficznie różnice w wyborze kolejnych punktów w trakcie optymalizacji między omawianymi w poprzedniej części pracy algorytmami spadku wzdłuż gradientu I (Cauchy'ego), spadku wzdłuż gradientu II (Newtona-Raphsona) oraz stochastycznego spadku wzdłuż gradientu I. W celu zobrazowania przykładu na dwuwymiarowym wykresie, postanowiono ograniczyć się do modelu z jedną zmienną objaśniającą i wyrazem wolnym. Do przykładu wybrano model regresji logistycznej, z racji na prostotę przedstawienia funkcji log-wiarogodności jako sumy różniczkowalnych składników.

Funkcja log-wiarogodności dla modelu regresji logistycznej, za [15] i [17], ma postać

$\beta = (\beta_1, \beta_2)$

$$\ell(\beta) = \sum_{i=1}^{N} \Big(y_i(\beta_1 + \beta_2 x_i) - \log(1 + \exp(\beta_1 + \beta_2 x_i)) \Big) = \sum_{i=1}^{N} Q_i(\beta_1, \beta_2), \tag{3.3}$$

$$Q_i(\beta_1, \beta_2) = y_i(\beta_1 + \beta_2 x_i) - \log(1 + \exp(\beta_1 + \beta_2 x_i)). \tag{3.4}$$

Dla tak skonstruowanej funkcji wiarogodności, współrzędne gradientu to odpowiednio

$$\frac{\partial \ell(\beta)}{\partial \beta_1} = \sum_{i=1}^{N} \Big(y_i - \pi_i(\beta) \Big), \qquad \frac{\partial \ell(\beta)}{\partial \beta_2} = \sum_{i=1}^{N} x_i \Big(y_i - \pi_i(\beta) \Big),$$

zaś macierz informacji wyraża się jak następuje

$$\mathscr{I}(\beta) = \begin{bmatrix} \sum\limits_{i=1}^{N} \pi_i(\beta)(1 - \pi_i(\beta)) & \sum\limits_{i=1}^{N} x_i \pi_i(\beta)(1 - \pi_i(\beta)) \\ \sum\limits_{i=1}^{N} x_i \pi_i(\beta)(1 - \pi_i(\beta)) & \sum\limits_{i=1}^{N} x_i^2 \pi_i(\beta)(1 - \pi_i(\beta)) \end{bmatrix}, \tag{3.5}$$

gdzie $\pi_i(\beta) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$, a N to liczba obserwacji.

Wtedy aktualizacja kandydata na miejsce zerowe w k-tym kroku algorytmu optymalizacyjnego dla kolejnych metod omówionych w rozdziale (3.1) wyraża się poniższymi wzorami

Metoda spadku wzdłuż gradientu I (Cauchy'ego)

$\alpha_k \in \mathbb{R}$

$$\beta_k = \beta_{k-1} + \alpha_k \cdot \Big(\sum_{i=1}^{N} (y_i - \pi_i(\beta_{k-1})), \sum_{i=1}^{N} x_i (y_i - \pi_i(\beta_{k-1})) \Big).$$

Metoda spadku wzdłuż gradientu II (Newtona-Raphsona)

$$\beta_k = \beta_{k-1} + \mathscr{I}(\beta_{k-1})^{-1} \cdot \Big(\sum_{i=1}^{N} (y_i - \pi_i(\beta_{k-1})), \sum_{i=1}^{N} x_i (y_i - \pi_i(\beta_{k-1})) \Big),$$

gdzie $\mathscr{I}(\beta_{k-1})$ zdefiniowane jest we wzorze (3.5).

Metoda stochastycznego spadku wzdłuż gradientu I

$$\beta_k = \beta_{k-1} + \alpha_k \cdot \Big(y_i - \pi_i(\beta_{k-1}), x_i (y_i - \pi_i(\beta_{k-1})) \Big),$$

dla wylosowanego w danym kroku i.

Ponieważ algorytmy te znajdują minimum funkcji, a docelowo szukane jest maksimum, stąd wykorzystano przeciwieństwo funkcji log-wiarogodności, dlatego w wyżej wymienionych wzorach zmieniono znaki przed pochodnymi na przeciwne.

Symulacje trajektorii zbieżności algorytmów

Poniższymi wywołaniami kodów z pakietu $\mathcal{R}$ [49] można wygenerować 1000 obserwacji z rozkładu jednostajnego i na ich podstawie wygenerować 1000 zmiennych z rozkładu dwupunktowego o takim rozkładzie prawdopodobieństwa sukcesu, by rzeczywiste współczynniki w modelu regresji logistycznej dla tych zmiennych wynosiły odpowiednio: 2 dla wyrazu wolnego oraz 3 dla zmiennej objaśniającej z rozkładu normalnego.

```
x <- runif(1000)
z <- 2 + 3*x
pr <- 1/(1+exp(-z))
y <- rbinom(1000,1,pr)
```

Dla tak sztucznie zasymulowanych danych przygotowano funkcję `logitGD()`, dzięki której można śledzić wartości ekstremum w danym kroku kolejnych omawianych algorytmów. Kody funkcji `logitGD()` oraz funkcji `graphSGD` tworzącej wykresy porównujące trajektorie zbieżności dla różnych algorytmów spadku gradientu dostępne są w Dodatku A.1. Wyniki poniższych wywołań funkcji `graphSGD` zostały umieszczone na Rysunkach (3.2) - (3.5).

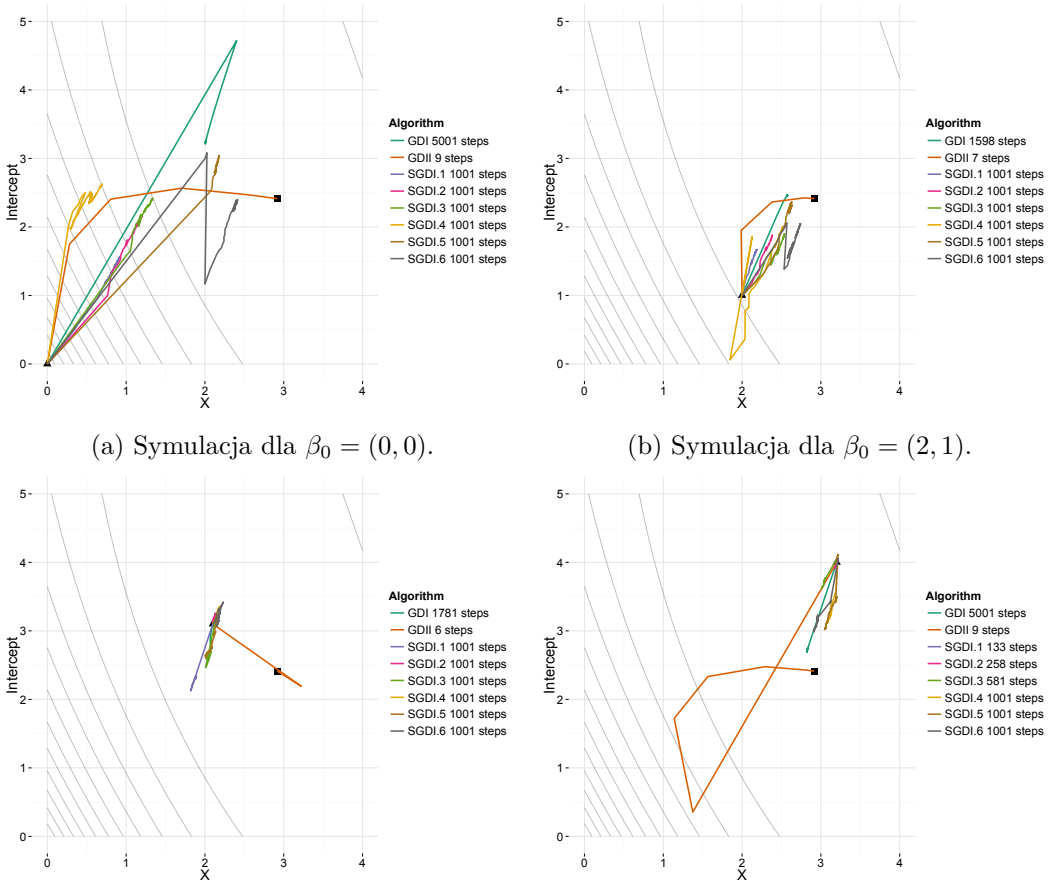
```
graphSGD(c(0,0), y, x, 4561);graphSGD(c(0,0), y, x, 456)
graphSGD(c(2,1), y, x, 4561);graphSGD(c(2,1), y, x, 456)
graphSGD(c(1,0), y, x, 4561);graphSGD(c(1,0), y, x, 456)
graphSGD(c(2.1,3.1), y, x, 4561);graphSGD(c(2.1,3.1), y, x, 456)
```

Na każdym z wykresów na osi OX ukazano współczynnik dla zmiennej objaśniającej w prostym modelu regresji logistycznej z jedną zmienną objaśniającą w kolejnych krokach poszczególnych algorytmów. Na osi OY zaznaczono współczynnik wyrazu wolnego w tym modelu. Punkt startowym zaznaczono na wykresie czarnym trójkątem. Czarnym kwadratem zaznaczono ekstremum funkcji log-wiarogodności w tym modelu, wyliczone dzięki funkcji `glm` [26], która do estymacji używa algorytmu Fishera (*ang. Fisher's Scoring algorithm* [39], [30]) używającego obserwowanej macierzy Informacji Fishera (definicja (1.10)) w miejscu Hesjanu w algorytmie Newtona-Raphsona (równanie (3.1)). Trajektoria zbieżności do minimum dla odrębnych algorytmów zaznaczono oddzielnymi kolorami, które są wyjaśnione na legendzie wykresu, na której dodatkowo wpisano liczbę kroków wymaganą przez algorytm do zbieżności. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.

W trakcie każdej symulacji postawiono pewne warunki konieczne do zbieżności. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, gdzie przez pierwszy krok rozumiano start z punktu startowego. Dodatkowo warunek stopu ustalono na $\epsilon = 0.000001$ oraz ustalono ciąg odpowiadający długościom kroków w algorytmie spadku gradientu rzędu I na $\alpha_k = \frac{1}{100k}$. Symulacje powtórzono czterokrotnie dla czterech różnych punktów startu algorytmów $\beta_0 = (0,0), (2,1), (1,0), (2.1, 3.1)$.

Ponieważ każda trajektoria w procesie estymacji metodą stochastycznego gradientu dla tych samych parametrów zbieżności jest inna, z racji na losową kolejność obserwacji, toteż dla każdego z czterech ustalonych punktów startowych zdecydowano się zobrazować symulację dwukrotnie, aby móc przedstawić stochastyczny aspekt tej metody. Symulacja nr 1 bazowała na losowym wzięciu punktów do algorytmów stochastycznego spadku gradientu, gdy ziarno losowania było ustawione na 4561, zaś dla symulacji nr 2 ziarno ustawiono na 456.

Ponieważ na Rysunkach (3.2) - (3.5) każda z symulacji ma inny zakres osi na wykresie, postanowiono na Rysunku (3.1) przygotować zestawienie podobnych symulacji przedstawiając trajektorie zbieżności na wspólnym zakresie osi. Miało to na celu uwypuklić skalę różnic w długościach kroków algorytmów w zależności od odległości punktu startowego od rzeczywistego ekstremum. Dodatkowo na osiach zaznaczono warstwice funkcji log-wiarogodności dla modelu regresji logistycznej (równanie (3.3)).



(a) Symulacja dla $\beta_0 = (0, 0)$.

(b) Symulacja dla $\beta_0 = (2, 1)$.

(c) Symulacja dla $\beta_0 = (2.1, 3.1)$.

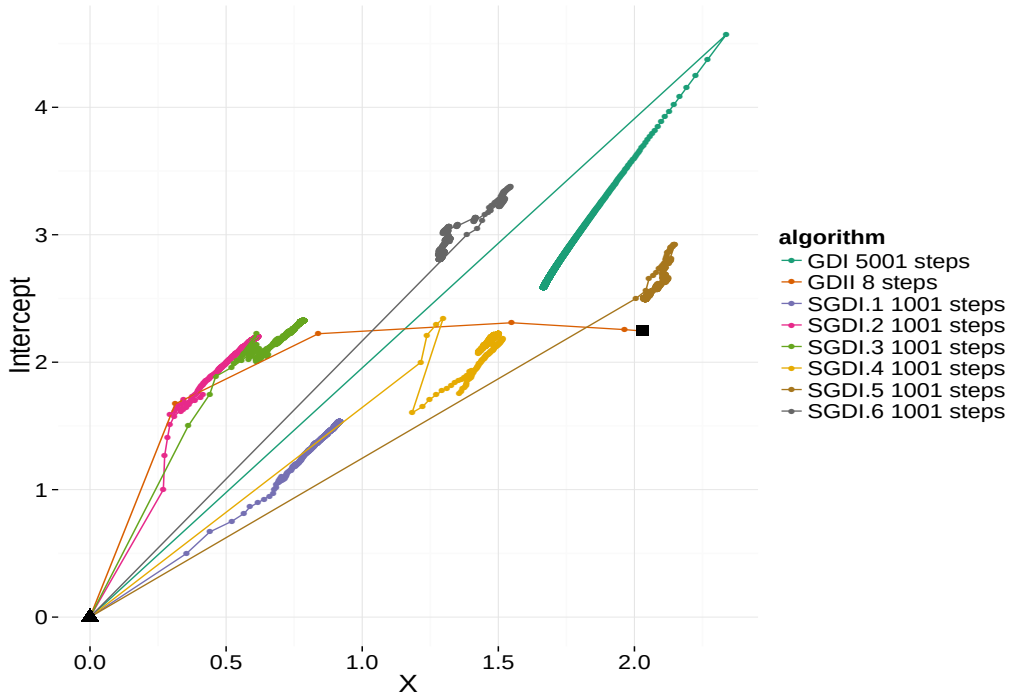
(d) Symulacja dla $\beta_0 = (3.2, 4)$.

Rysunek 3.1: Wykresy o wspólnym zakresie osi z zaznaczeniem warstwic funkcji log-wiarogodności dla modelu regresji logistycznej (równanie (3.3)). Wykresy przedstawiają porównanie algorytmów spadku gradientu. Początkowe dane losowano z innego ziarna niż na wykresach ukazanych na Rysunkach (3.2) - (3.5). Wykresy przedstawiają ścieżki zbieżności w kolejnych krokach algorytmów spadku gradientu. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, zaś warunek stopu ustalono na $\epsilon = 10^{-6}$. Trójkątem zaznaczono punkt startowy, a kwadratem wyestymowane rozwiązanie przy pomocy funkcji `glm` [26]. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.

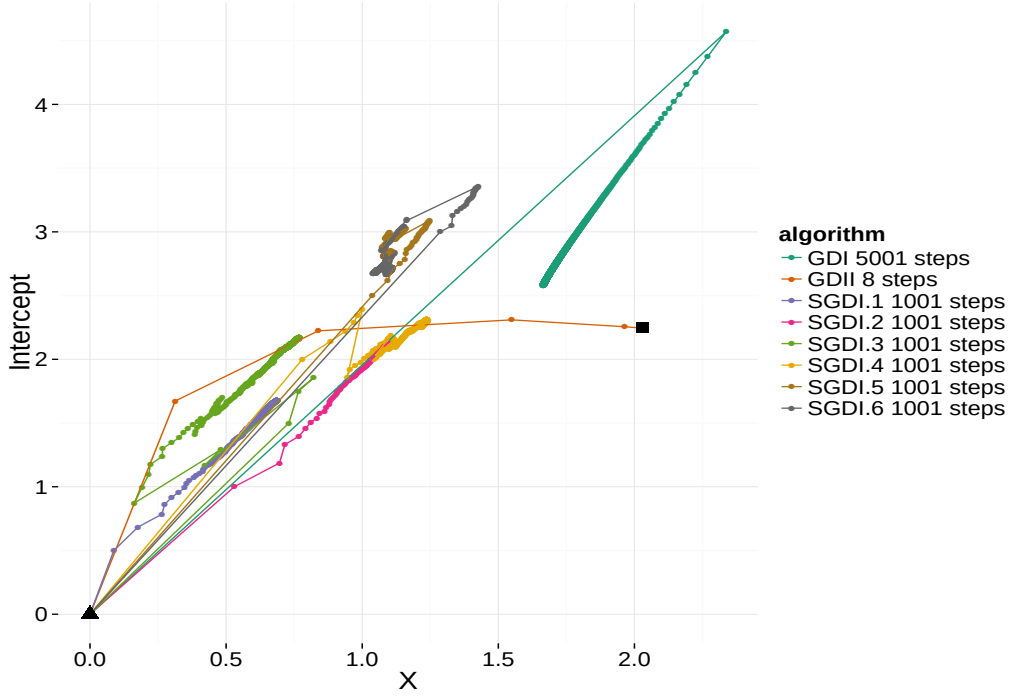
Na kolejnych wykresach na Rysunkach (3.2) - (3.5) osie dopasowane są do obecnej symulacji. Ze względu na pomniejszenie zakresu osi, nie zdecydowano się na ukazanie warstwic.

Podsumowanie symulacji

W trackie symulacji badano zachowanie trajektorii zbieżności w kolejnych krokach algorytmów spadku gradientu. Algorytm spadku gradientu rzędu II zbiegał do rozwiązania wyznaczonego przez funkcję `glm` [26]. Na Rysunkach (3.4) i (3.2) widać, zwłaszcza dla Symulacji nr 2, że ścieżki wyznaczane przez algorytmy stochastycznego spadku gradientu nie pokrywają się ze ścieżką wyznaczoną przez algorytm spadku gradientu rzędu II, a raczej pokrywają się ze ścieżką algorytmu spadku gradientu rzędu I. W tych sytuacjach widać, że warianty stochastycznego spadku gradientu zmierzają bezpośrednio do rozwiązania w przypadkach gdy pierwsze kroki algorytmy są dłuższe, czyli gdy ustali się ciąg odpowiadający długościom kroków algorytmu na $\alpha_{ki} = \frac{i}{k}$ i $i >= 4$. Rysunek (3.5) obrazuje ciekawy przypadek, w którym stochastyczne warianty algorytmów rozpoczynają estymację w kierunku przeciwnym do rzeczywistego optimum, by dopiero po kilku obserwacjach porzucić ten kierunek i szukać rozwiązania w okolicach zbliżonych do prawdziwego minimum. W tej sytuacji znowu widać, że im dłuższe proporcjonalnie kolejne kroki algorytmu, tym szanse na powrót na właściwe tory dla algorytmów stochastycznych jest większa. Algorytmy o indeksach $i = 1, 2$ nie posunęły się z optymalizacją za daleko od punktu startowego i nie zdążyły zawrócić na właściwy kierunek, zanim wyczerpały się wszystkie dostępne obserwacje. Najgorzej sprawuje się stochastyczny spadek gradientu na Rysunku (3.3), gdzie widać na powiększeniu, że algorytmy stochastyczne (a nawet spadek gradientu rzędu I) błądzą w kierunku jedynie zbliżonym do rzeczywistego kierunku, na którym znaleźć można optimum, jednak może to wynikać z małych różnic w wartości funkcji log-wiarogodności na obszarze błądzenia w poszukiwaniu minimum. Z Rysunku (3.1) można wywnioskować, że im dłuższy dystans od punktu początkowego do prawdziwego optimum, tym trajektoria zbiegania algorytmów są dłuższe (wykres a). Można też z niego odczytać, że istnieją sytuacje, w których optymalizacja sobie nie radzi i algorytmy, poza spadkiem gradientu rzędu II, stoją w miejscu (wykres c) oraz że w pewnych przypadkach (wykres d) nawet spadek gradientu rzędu II potrafi błądzić zanim ostatecznie dotrze do minimum.

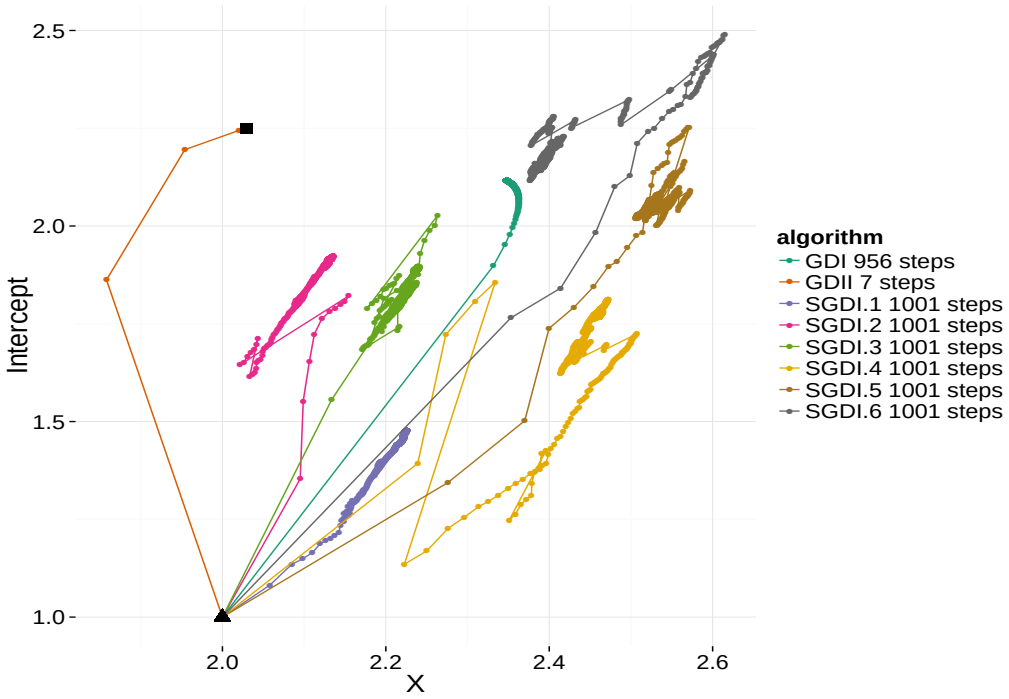


(a) Symulacja dla $\beta_0 = (0, 0)$ nr 1 dla ziarna losowania ustalonego na 4561.

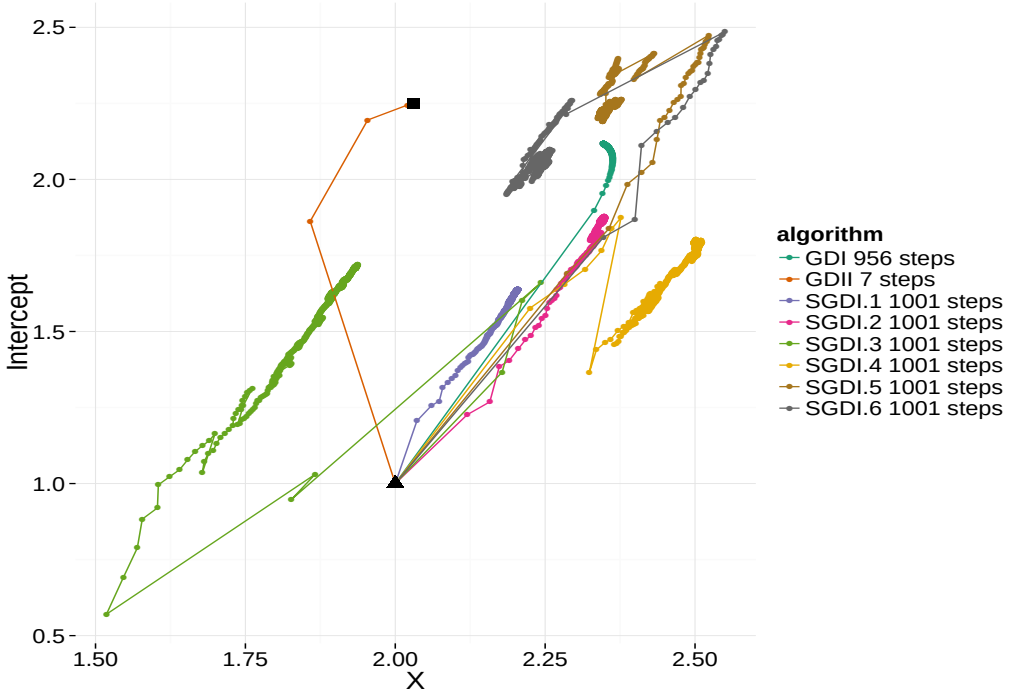


(b) Symulacja dla $\beta_0 = (0, 0)$ nr 2 dla ziarna losowania ustalonego na 456.

Rysunek 3.2: Porównanie algorytmów spadku gradientu dla punktu startowego $\beta_0 = (0, 0)$. Wykres przedstawia ścieżki zbieżności w kolejnych krokach algorytmów spadku gradientu. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, zaś warunek stopu ustalono na $\epsilon = 10^{-6}$. Trójkątem zaznaczono punkt startowy, a kwadratem wyestymowane rozwiązanie przy pomocy funkcji `glm` [26]. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.

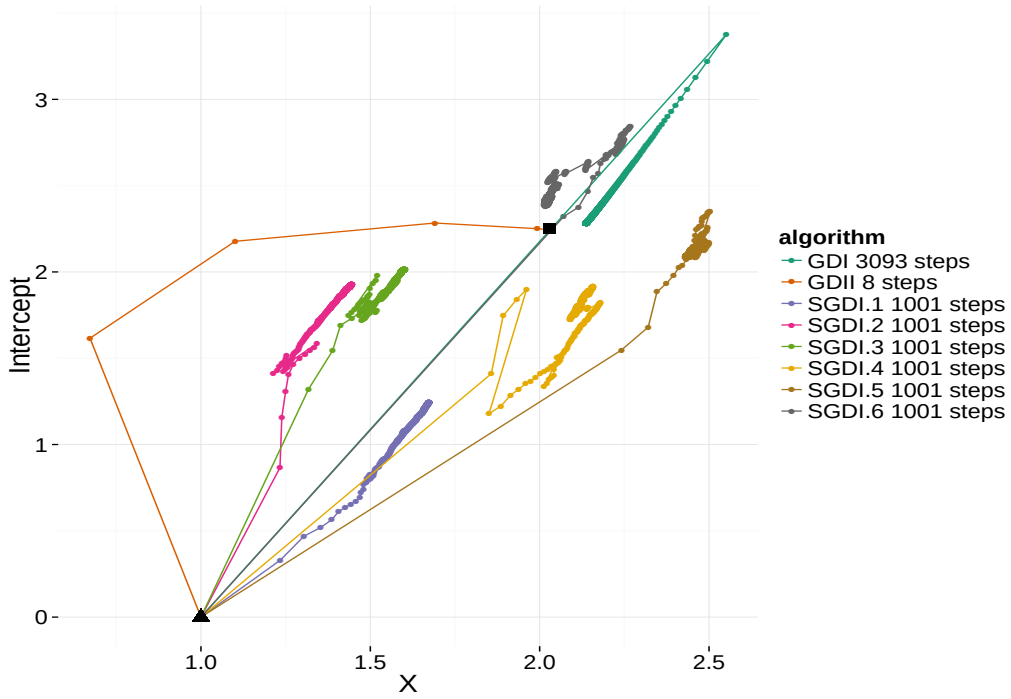


(a) Symulacja dla $\beta_0 = (2, 1)$ nr 1 dla ziarna losowania ustalonego na 4561.

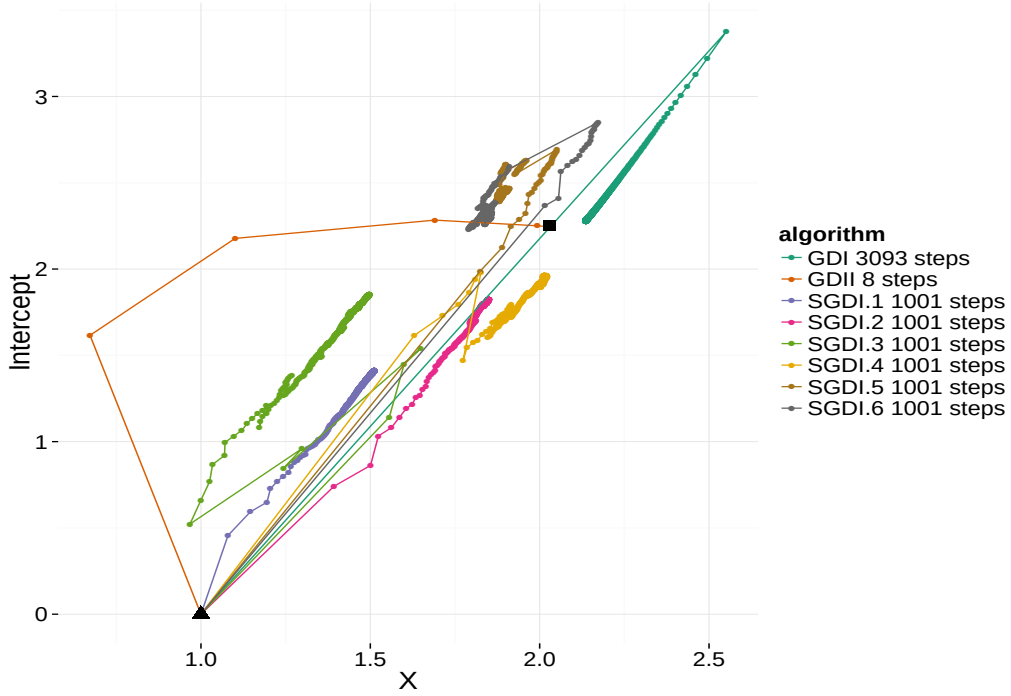


(b) Symulacja dla $\beta_0 = (2, 1)$ nr 2 dla ziarna losowania ustalonego na 456.

Rysunek 3.3: Porównanie algorytmów spadku gradientu dla punktu startowego $\beta_0 = (2, 1)$. Wykres przedstawia ścieżki zbieżności w kolejnych krokach algorytmów spadku gradientu. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, zaś warunek stopu ustalono na $\epsilon = 10^{-6}$. Trójkątem zaznaczono punkt startowy, a kwadratem wyestymowane rozwiązanie przy pomocy funkcji `glm` [26]. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.

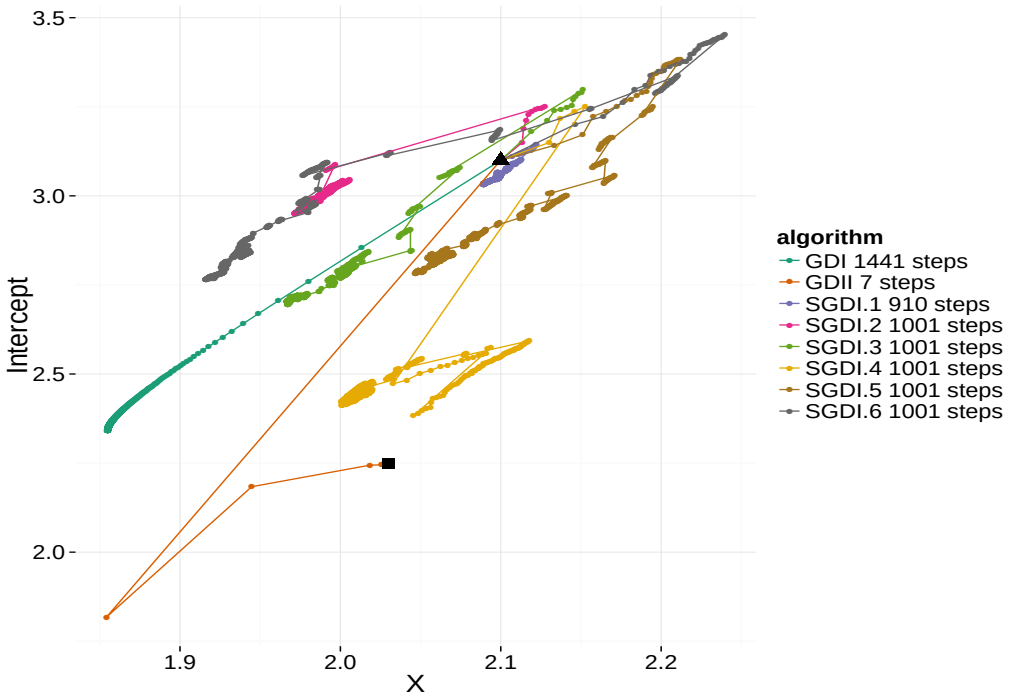


(a) Symulacja dla $\beta_0 = (1, 0)$ nr 1 dla ziarna losowania ustalonego na 4561.

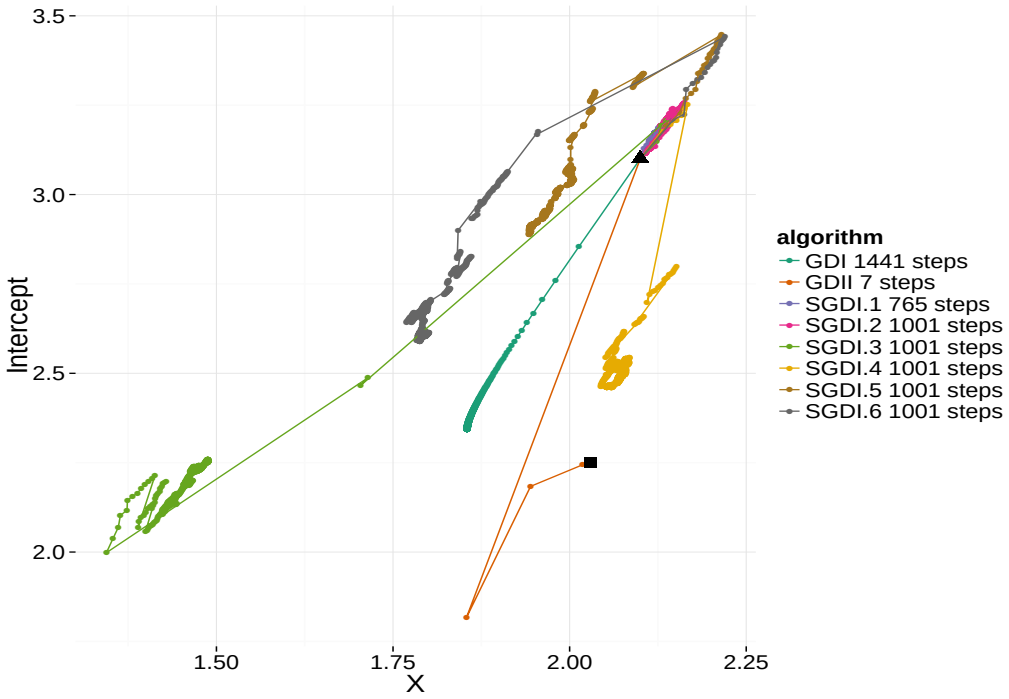


(b) Symulacja dla $\beta_0 = (1, 0)$ nr 2 dla ziarna losowania ustalonego na 456.

Rysunek 3.4: Porównanie algorytmów spadku gradientu dla punktu startowego $\beta_0 = (1, 0)$. Wykres przedstawia ścieżki zbieżności w kolejnych krokach algorytmów spadku gradientu. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, zaś warunek stopu ustalono na $\epsilon = 10^{-6}$. Trójkątem zaznaczono punkt startowy, a kwadratem wyestymowane rozwiązanie przy pomocy funkcji `glm` [26]. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.



(a) Symulacja dla $\beta_0 = (2.1, 3.1)$ nr 1 dla ziarna losowania ustalonego na 4561.



(b) Symulacja dla $\beta_0 = (2.1, 3.1)$ nr 2 dla ziarna losowania ustalonego na 456.

Rysunek 3.5: Porównanie algorytmów spadku gradientu dla punktu startowego $\beta_0 = (2.1, 3.1)$. Wykres przedstawia ścieżki zbieżności w kolejnych krokach algorytmów spadku gradientu. Ustalono maksymalną liczbę iteracji na 5001 dla algorytmów spadku gradientu i 1001 dla wariantów stochastycznych, zaś warunek stopu ustalono na $\epsilon = 10^{-6}$. Trójkątem zaznaczono punkt startowy, a kwadratem wyestymowane rozwiązanie przy pomocy funkcji `glm` [26]. Przez `GDI` oznaczono trajektoria dla algorytmu spadku gradientu rzędu I, przez `GDII` oznaczono trajektoria dla algorytmu spadku gradientu rzędu II, zaś przez `SGD.i` oznaczono 6 różnych trajektorii dla algorytmów stochastycznego spadku gradientu dla różnych ciągów odpowiadających długości kroku algorytmu. Indeks i odpowiada ciągowi wybranemu do wyznaczania długości kroku algorytmu na zasadzie $\alpha_{ki} = \frac{i}{k}$.

Rozdział 4

Estymacja w modelu Coxa metodą stochastycznego spadku gradientu

Poniższy rozdział przedstawia implementację oraz zastosowanie metody stochastycznego spadku gradientu do estymacji współczynników w modelu proporcjonalnych hazardów Coxa. Jest to główny cel pracy. Poniższe rozważania odnośnie podejścia do stosowania tej metody w tym konkretnym modelu nie są oparte na żadnej literaturze ze względu na jej brak.

W czasie powstawania pracy nawiązano jedynie nieznaczną wymianę informacji z pracownikami *Harvard Laboratory for Applied Statistical Methodology & Data Science*, dzięki której dowiedziano się że podjęte zostały kroki w kierunku stworzenia podwalin pod teorię do omawianego zagadnienia, jednak ewentualne publikacje nie zostały jeszcze dokończone. Przedsmak niedokończonej implementacji algorytmu przez pracowników wyżej wymienionego laboratorium można znaleźć w [54].

W procesie estymacji współczynników w omawianym modelu wykorzystano pochodne cząstkowe częściowej funkcji log-wiarogodności (2.10)

$$U_k(\beta) = \frac{\partial \ell_k(\beta)}{\partial \beta_k} = \sum_{i=1}^{K} \Big(X_{ik} - \frac{\sum\limits_{l \in \mathscr{R}(t_i)} X_{lk} e^{X_l' \beta}}{\sum\limits_{l \in \mathscr{R}(t_i)} e^{X_l' \beta}} \Big) = \sum_{i=1}^{n} Y_i \Big(X_{ik} - \frac{\sum\limits_{l \in \mathscr{R}(t_i)} X_{lk} e^{X_l' \beta}}{\sum\limits_{l \in \mathscr{R}(t_i)} e^{X_l' \beta}} \Big) = \sum_{i=1}^{n} U_{k_i}(\beta),$$

(dla $Y_i = 1$, gdy obserwacja nie była cenzurowana i $Y_i = 0$, gdy obserwacja była cenzurowana; K to liczba zdarzeń; n to liczba obserwacji) oraz wzór (3.2), którego wersja z adekwatnymi oznaczeniami dla powyższej funkcji wygląda następująco

$$\beta_{k_{j+1}} = \beta_{k_j} - \alpha_j U_{k_i}(\beta_{k_j}), \tag{4.1}$$

gdzie j oznacza krok algorytmu, i iteruje składniki U_k, α_j to długość j-tego kroku algorytmu, zaś U_k to k-ta pochodna cząstkowa gradientu U oraz β_k to k-ta współrzędna estymowanego wektora współczynników β o rozmiarze p, czyli $k = 1, \ldots, p$.

Własna implementacja algorytmu w języku $\mathcal{R}$ znajduje się w podrozdziale (4.2).

4.1. Założenia i obserwacje

Skupiając się na informatycznym aspekcie algorytmu można stwierdzić, że idea stochastycznego spadku gradientu polega na losowaniu składnika optymalizowanej funkcji. Jednak ze statystycznego punktu widzenia, metoda stochastycznego spadku gradientu opiera się o losowanie indeksu obserwacji ze zbioru, z którego uczony jest algorytm, zanim postanowi się w jakikolwiek sposób przedstawić konkretną funkcję wiarogodności. Zatem w celu estymacji w oparciu o stochastyczny spadek gradientu w modelu Coxa, konstruując metodę, należy najpierw losować obserwacje a następnie dopiero wyznaczać formę optymalizowanej funkcji częściowej log-wiarogodności.

Dla wielu modeli opierających się o funkcje wiarogodności te dwa punkty widzenia są równoważne, jednak dla modelu Coxa nie, gdyż niektóre składniki w funkcji log-wiarogodności są zależne od poprzednich obserwacji. W przypadku modelu ADALINE sformułowanego jak w [60], opartego na minimalizowaniu funkcji kosztu w postaci błędu najmniejszych kwadratów, w [7] podano postać funkcji straty oraz równanie algorytmu stochastycznego spadku gradientu jak poniżej

$$Q_{adaline} = \frac{1}{2}(y - w'\Phi(x))^2, \quad \Phi(x) \in \mathbb{R}^d, y \in \{-1, 1\},$$

$$w \leftarrow w + \alpha_k(y_k - w'\Phi(x_t))\Phi(x_t),$$

dla których widać, że w kolejnych krokach algorytmu t wystarczy tylko jedna obserwacja $z_t = (x_t, y_t)$ aby poprawić oszacowanie parametru w.

W modelu proporcjonalnych hazardów Coxa jest to bardziej skomplikowane. Dla zadanej z góry funkcji hazardu, częściowa funkcja wiarogodności odpowiada prawdopodobieństwu tego, że obserwowane zdarzenia zdarzyłyby się dokładnie w tej kolejności w jakiej się pojawiły. To prawdopodobieństwo zależy od wszystkich obserwacji w zbiorze. Niemożliwe jest wyliczenie tego poprzez obliczenie wartości funkcji częściowej wiarogodności oddzielnie dla obserwacji o numerach od 1 do 5 i oddzielnie dla obserwacji od numerach 6 do 10, a następnie przemnożeniu wyników przez siebie. Z tej przyczyny niemożliwe jest losowanie czynników optymalizowanej funkcji przy użyciu metody stochastycznego spadku gradientu do estymacji współczynników w tym modelu. Alternatywnym podejściem do tego problemu mogłoby być pamiętanie wartości licznika i mianownika dla składników częściowej funkcji log-wiarogodności dla wszystkich zaobserwowanych czasów zdarzeń i poprawianie odpowiednich składników z wykorzystaniem świeżo zaobserwowanych obserwacji. Taki zabieg jest pamięciowo oszczędniejszy niż pamiętanie wszystkich obserwacji. **Możliwe jest też losowanie podzbioru obserwacji a następnie konstruowanie funkcji wiarogodności dla zaobserwowanego zredukowanego zbioru. Właśnie ta metoda zostanie opisana w dalszej części pracy.** Taki sposób wprowadzania obserwacji do estymacji można wykorzystać w sytuacjach, gdy mamy do czynienia z nieskończonym napływem nowych obserwacji a interesują nas oszacowania estymowanych parametrów modelu $\beta_k, k = 1, \ldots, p$ dla obecnie zaobserwowanych i wykorzystanych obserwacji. Proces ten ma dwie zalety: nie dość, że dla nowych obserwacji model jest w stanie na bazie obecnych oszacowań parametrów dokonać predykcji proporcji hazardów, to dodatkowo po każdej porcji obserwacji aktualizuje parametry modelu.

Ponieważ omawiane algorytmy rozwiązują problem minimalizacji badanej funkcji, zaś celem estymacji w modelu Coxa jest znalezienie parametrów modelu maksymalizujących funkcję częściowej log-wiarogodności, zatem wzięcie do minimalizacji funkcji z przeciwnym znakiem doprowadzi do wykorzystania metod znajdujących minimum do znalezienia maksimum.

Zakładając, że dla j-ego kroku algorytmu i k-tej pochodnej cząstkowej dysponuje się zaobserwowanym podzbiorem $\mathcal{B}$, częściową funkcję wiarogodności dla zaobserwowanego podzbioru obserwacji $\mathcal{B}$ wykorzystywaną do minimalizacji można zapisać następująco

$$-U_k^{\mathcal{B}}(\beta_j) = -\sum_{i \in \mathcal{B}_{\text{ind}}} U_{k_i}^{\mathcal{B}}(\beta_j) = -\sum_{i \in \mathcal{B}_{\text{ind}}} Y_i \Big(X_{ik} - \frac{\sum\limits_{l \in \mathscr{R}_{\mathcal{B}}(t_i)} X_{lk} e^{X_l' \beta_j}}{\sum\limits_{l \in \mathscr{R}_{\mathcal{B}}(t_i)} e^{X_l' \beta_j}} \Big), \tag{4.2}$$

gdzie indeksy obserwacji należące do zbioru $\mathcal{B}$ definiuje się jako $\mathcal{B}_{\text{ind}} = \{i : X_i \in \mathcal{B}\}$, zaś $\mathscr{R}_{\mathcal{B}}(t_i)$ to zbiór ryzyka dla podzbioru $\mathcal{B}$ w czasie t_i.

Postać powyższej funkcji nasuwa pewne obserwacje. Dla danego kroku algorytmu, obecnie wykorzystywany zaobserwowany podzbiór obserwacji $\mathcal{B}$

- *nie powinien składać się jedynie z obserwacji cenzurowanych.*

  Dla podzbioru zawierającego jedynie takie obserwacje wartość pochodnej cząstkowej funkcji log-wiarogodności jest równa zero, ze względu na czynniki Y_i, które dla obserwacji cenzurowanych są równe zero. Zerowa wartość pochodnej cząstkowej funkcji log-wiarogodności doprowadzi do niezmienienia się optymalizowanych parametrów we wzorze (4.1), a co za tym idzie, doprowadzi do przerwania optymalizacji.

- *nie powinien składać się jedynie z jednej obserwacji.*

  W takim przypadku wartość pochodnej cząstkowej funkcji log-wiarogodności jest również równa zero, bez względu na to czy obserwacja była cenzurowana czy nie. Zerowa wartość pochodnej cząstkowej funkcji log-wiarogodności doprowadzi do niezmienienia się optymalizowanych parametrów we wzorze (4.1), a co za tym idzie, doprowadzi do przerwania optymalizacji.

- *nie powinien zawierać tylko jednej obserwacji niecenzurowanej w zbiorze, która dodatkowo ma największy czas bycia pod obserwacją.*

  W tej sytuacji jedyny niezerowy czynnik w całej pochodnej cząstkowej funkcji log-wiarogodności zajdzie tylko dla obserwacji niecenzurowanej, dla której zbiór ryzyka będzie zawierał tylko nią, co da ostatecznie zerową wartość pochodnej cząstkowej optymalizowanej funkcji log-wiarogodności, co doprowadzi do przerwania optymalizacji.

Dodatkowo nie dochodzi do wykorzystania obserwacji cenzurowanych, gdy:

- *zaobserwowany zbiór zawiera obserwacje cenzurowane, które wszystkie mają czasy obserwacji krótsze niż najmniejszy czas obserwacji dla obserwacji niecenzurowanej.*

  Obserwacje cenzurowane niosą ze sobą mniej informacji, jednak teoria analizy przeżycia stara się je maksymalnie wykorzystać, stąd uwzględnia się ich udział w zbiorze ryzyka przy estymacji fragmentów funkcji log-wiarogodności dla obserwacji niecenzurowanych. W sytuacji, gdy wszystkie obserwacje cenzurowane mają czas obserwacji krótszy niż najmniejszy czas obserwacji dla obserwacji niecenzurowanej w podzbiorze, nie dochodzi do wykorzystania obserwacji cenzurowanych w jakikolwiek sposób przy estymacji współczynników w danym kroku algorytmu.

Mając na względzie te uwagi, w sytuacji gdy zaobserwuje się podzbiór, który spełnia jeden z wyżej wymienionych warunków, można zamiast wykorzystywać ten zbiór do estymacji można go dołączyć do kolejnego podzbioru, który ma być zaobserwowany.

4.2. Implementacja

Omówiona w tym rozdziale metoda estymacji metodą stochastycznego spadku gradientu dla modelu proporcjonalnych hazardów Coxa została zaimplementowana w języku $\mathcal{R}$ [49] i jest dostępna w specjalnie przygotowanym pakiecie o nazwie `coxphSGD()`, który można pobrać z internetu i zainstalować poleceniem

```
devtools::install_github("MarcinKosinski/Cox-SGD")
```

Dokumentacja wraz z opisem argumentów w języku angielskim funkcji `coxphSGD()`, która estymuje współczynniki w modelu proporcjonalnych hazardów Coxa metodą stochastycznego spadku gradientu, dostępna jest w Dodatku A. Starano zachować jednorodność kolejności i nazewnictwa parametrów z funkcją `coxph()` z pakietu `survival` [56], [57].

Implementacja algorytmu estymacji w modelu Coxa metodą stochastycznego spadku gradientu opiera się na poniższym pseudo-kodzie i zakłada, że kolejne podzbiory $\mathcal{B}$ dostarczane są jako kolejne elementy listy.

Estymacja w modelu Coxa metodą stochastycznego spadku gradientu

```
                    # wstępna inicjalizacja parametrów
eps = 1e-5                                  # warunek stopu.

n = length(data)                            # data jest listą ramek danych.

diff = eps + 1                              # różnice w oszacowaniach parametrów
                                            # między kolejnymi krokami.

learningRates = function(x) 1/x             # długości kroku algorytmu.

beta_old = numeric(0, length = k)           # punkt startowy dlugosci k,
                                            # gdzie k to liczba zmiennych
                                            # objaśniających w modelu.

                                # estymacja
i = 1                                       # iterator kroku algorytmu
while(i <= n & diff > eps) do               # do zbieżności lub wyczerpania zbiorów
  batch = data[[i]]

  beta_new = beta_old - learningRates(i) * U_Batch(batch)
                                            # U_Batch to częściowa funkcja
                                            # log-wiarogdności dla zaobserwowanego
                                            # zbioru 'batch'

  diff = euclidean_dist(beta_new, beta_old) # odległość euklidesowa

  beta_old = beta_new

  i = i + 1
end while
return beta_new
```

Docelowa implementacja w języku $\mathcal{R}$ znajduje się poniżej.

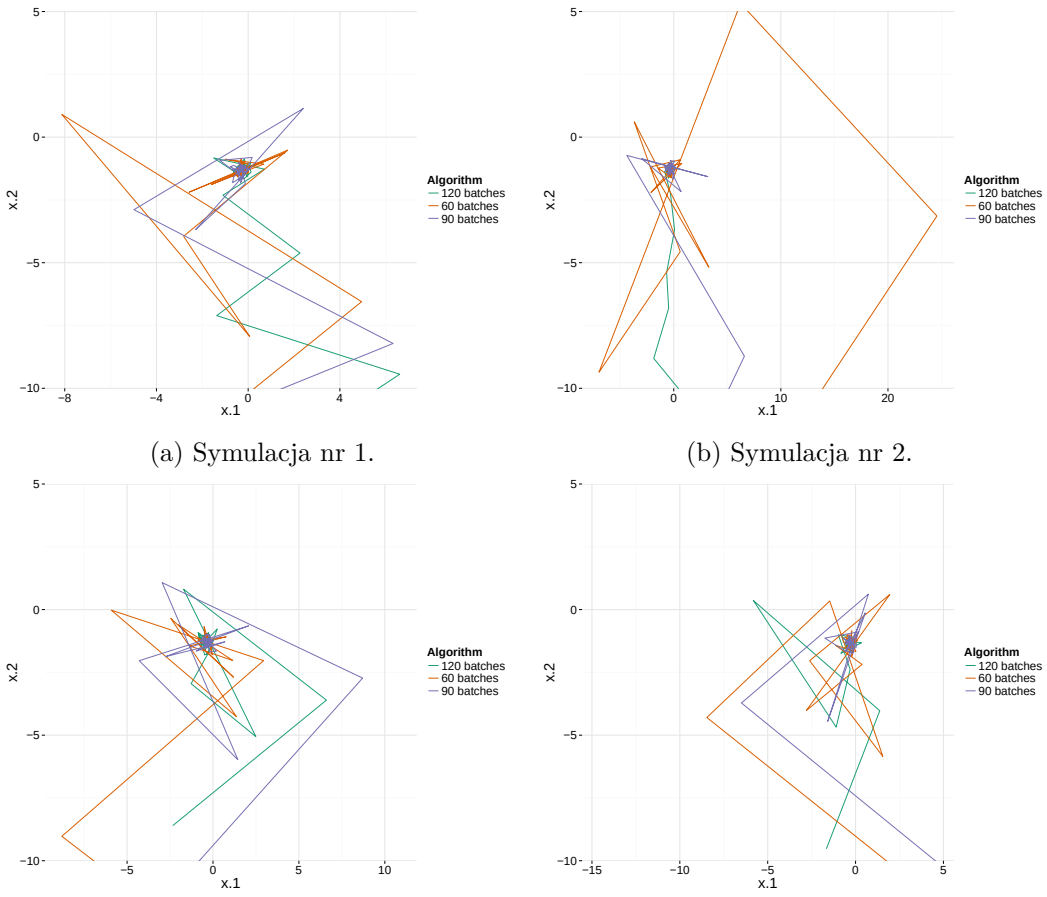
```
coxphSGD <- function(formula, data, learningRates = function(x){1/x},
                     beta_0 = 0, epsilon = 1e-5 ) {
  checkArguments(formula, data, learningRates,
                 beta_0, epsilon) -> beta_old # check arguments
  n <- length(data)
  diff <- epsilon + 1
  i <- 1
  beta_new <- list() # steps are saved in a list so that they can
                     # be tracked in the future
  # estimate
  while(i <= n & diff > epsilon) {
    beta_new[[i]] <- coxphSGD_batch(formula = formula, data = data[[i]],
                      learningRate = learningRates(i), beta = beta_old)

    diff <- sqrt(sum((beta_new[[i]] - beta_old)^2))
    beta_old <- beta_new[[i]]
    i <- i + 1
  }
  # return results
  list(Call = match.call(), coefficients = beta_new, epsilon = epsilon,
       learningRates = learningRates, steps = i)
}

coxphSGD_batch <- function(formula, data, learningRate, beta){
  # collect times, status, variables and reorder samples
  batchData <- prepareBatch(formula = formula, data = data)
  # calculate the log-likelihood for this batch sample
  partial_sum <- list()
  for(k in 1:nrow(batchData)) {
    # risk set for current time/observation
    risk_set <- batchData %>% filter(times <= batchData$times[k])

    nominator <- apply(risk_set[, -c(1,2)], MARGIN = 1, function(element){
      element * exp(element * beta)
    }) %>% rowSums()

    denominator <- apply(risk_set[, -c(1,2)], MARGIN = 1, function(element){
      exp(element * beta)
    }) %>% rowSums()

    partial_sum[[k]] <-
      batchData[k, "event"] * (batchData[k, -c(1,2)] - nominator/denominator)
  }
  do.call(rbind, partial_sum) %>%
    colSums() -> U_batch

  return(beta + learningRate * U_batch)
}
```

4.3. Symulacje estymacji w modelu Coxa

Dzięki przygotowanej w rozdziale 2.4 funkcji `dataCox()` możliwe będzie symulacje zbadanie zachowania się współczynników w modelu Coxa w trakcie estymacji metodą stochastycznego spadku gradientu, której algorytm został zaimplementowany w rozdziale 4.2. Poniższe wywołanie zwraca ramkę danych o 10 tysiącach wierszy, wraz z ze sztucznie wygenerowanymi czasami z rozkładu Weibulla o parametrach $\lambda = 3$ oraz $\rho = 2$. Czasy cenzurowania generowane są z rozkładu wykładniczego o parametrze 5 i odpowiadają współczynnikom modelu $\beta = (1, 3)$. W rezultacie otrzymać można czasy przeżycia oraz status zdarzenia bądź cenzurowania. Funkcja `vizCoxSGD()` opisana w Dodatku A.2 trzykrotnie dzieli wygenerowany zbiór danych odpowiednio na 60, 90 i 120 podzbiorów ustawiając obserwacje w losowej kolejności, a następnie wykorzystuję funkcję `coxphSGD()` do wyznaczenia wartości współczynników w kolejnych krokach estymacji funkcji log-wiarogodności dla modelu Coxa przy pomocy metody stochastycznego spadku gradientu.

```
x <- matrix(sample(0:1, size = 20000, replace = TRUE), ncol = 2)
dCox <- dataCox(10^4, lambda = 3, rho = 2, x, beta = c(1,3), censRate = 5)
vizCoxSGD(dCox)
```

Efekt czterokrotnego wywołania funkcji `vizCoxSGD()` widoczny jest na Rysunku (4.1).



(a) Symulacja nr 1.

(b) Symulacja nr 2.

(c) Symulacja nr 3.

(d) Symulacja nr 4.

Rysunek 4.1: Porównanie estymacji w modelu Coxa metodą stochastycznego spadku gradientu dla różnych podziałów zbioru początkowego na podzbiory.

Rozdział 5

Analiza danych genomicznych

Tke key is to understand genomics
to improve cancer care.
The Cancer Genome Atlas Study

Niniejszy rozdział poświęcony jest analizie danych genomicznych. W podrozdziale 5.1 przedstawiono pokrótce genetyczne podstawy nowotworzenia, które szerzej opisane są w cytowanej literaturze. W podrozdziale 5.2 opisano schemat *The Cancer Genome Atlas* (TCGA), badania z którego zaczerpnięto dane do analizy biostatystycznej. Analiza przeżycia polegająca na zastosowaniu modelu Coxa, dla którego estymacja współczynników odbywa się metodą stochastycznego spadku gradientu (opisaną w rozdziale 4) zaprezentowana została w podrozdziale 5.3, zaś komentarz do wyników można znaleźć w podrozdziale 5.4.

W ramach analizy starano się zweryfikować siłę zagadnienia wpływu mutacji genów na przeżywalność pacjentów. Celem analizy było uzyskanie odpowiedzi na pytanie czy istnieją geny, które zmutowane powodują, że ryzyko wystąpienia zdarzenia jakim jest zgon w wyniku choroby nowotworowej jest większe w grupach pacjentów bez mutacji w danym genie. Aby zweryfikować hipotezę na tak obszernych danych zdecydowano się zastosować metodę strumieniowej estymacj współczynników w modelu Coxa przy pomocy stochastycznego spadku gradientu, której stworzenie było głównym założeniem pracy, gdyż tradycyjna metoda zaimplementowana w funkcji `coxph` w pakiecie `survival` [57] nie jest przystosowana do tak obfitych zbiorów danych jak te z TCGA. Z racji na zbyt dużą ilość genów (przekraczającą kilkanaście tysięcy), które mogły być zmutowane i wzięte do analizy jako zmienna objaśniająca czas do zdarzenia, niemożliwe byłoby wyestymowanie współczynników modelu standardową metodą.

Wielką zaletą modelu Coxa z estymacją przy użyciu metody stochastycznego spadku gradientu jest możliwość zastosowania go do zbiorów danych o rozmiarze zmiennych znacznie przekraczającym wymiar obserwacji. Zaprezentowana w rozdziale 4.2 implementacja pozwala w skończonym czasie znaleźć współczynniki modelu odpowiadające za wpływ danych genów na śmiertelność pacjentów w chorobach nowotworowych. Jest to nie tylko zdobycz obliczeniowa ale także ważne narzędzie w analizach biostatystycznych danych dużej skali, które z powodzeniem może zostać wykorzystywane do analizy ludzkiego genomu, tak by w przyszłości poszerzać zrozumienie procesów mutacji zachodzących w ludzkim organizmie dzięki którym możliwe byłoby skuteczniejsze leczenie chorób nowotworowych.

5.1. Genetyczne podstawy nowotworzenia

Choroby nowotworowe stanowią drugą, po chorobach serca, przyczynę zachorowań i zgonów na całym świecie. Wyniki opracowane przez Polską Unię Onkologiczna wskazują na tendencję wzrostową liczby zachorowań na nowotwory i rokują na utrzymanie się jej do 2020 [48].

Według współczesnej definicji **nowotwór** jest chorobą cyklu komórkowego i oznacza [18]

nieprawidłową tkankę, która powstała z jednej komórki i rośnie jako następstwo zaburzeń dynamizmu i prawidłowego przebiegu cyklu komórkowego oraz zaburzeń różnicowania się komórki i komunikacji wewnątrzkomórkowej, międzykomórkowej i pozakomórkowej jej klonalnego potomstwa.

Wyniki badań nad transformacją nowotworową wykazały, że nowotwory powstają jako wynik wielu nielegalnych mutacji w DNA komórki somatycznej, które poprzez kumulację wywołują utratę kontroli nad proliferacją (rozwojem), wzrostem i różnicowaniem [37].

Proces tworzenia nowotworu (**karcynogeneza**) jest wieloczynnikowy i wielostopniowy, a zmiany w nim nasilają się w miarę pogłębiania się niestabilności genetycznej [18]. Transformacja nowotworowa następuje w wyniku zmian powstałych w obrębie czterech różnych kategorii genów, które wpływają na proliferację i różnicowanie komórek [48]. Lista genów wpływających na proliferację i różnicowanie komórek z opisem [29] znajduje się poniżej

- *geny regulujące naprawę uszkodzonego DNA* - mechanizmy szybkiej naprawy DNA zapobiegają przed mutacjami genów odpowiedzialnych m. in. za proliferację i rożnicowanie się komórek. Geny biorące udział w naprawie DNA nie są onkogenne, natomiast mutacje w ich obrębie mogą ułatwić transformację nowotworową oraz podwyższają ryzyko utrwalenia się zmian w pozostałych grupach i dlatego mają podstawowe znaczenie dla integracji genomu.

- *geny supresorowe* (antyonkogeny) - geny działające hamująco na procesy proliferacji komórkowej bądź stabilizująco na procesy utrzymujące stabilność genetyczną komórki,

- *protoonkogeny* - potencjalnie zdolne do wyzwolenia procesu transformacji nowotworowej. Uwarunkowana mutacją zmiana ich ekspresji sprawia, że przekształcają się w onkogeny, czyli geny bezpośrednio aktywujące transformację nowotworową,

- *geny regulujące apoptozę* (naturalny proces zaprogramowanej śmierci komórki w organizmach wielokomórkowych) - zahamowanie procesu apoptozy wydłuża okres przeżycia komórek, zwiększając tym samym liczebność populacji komórek narażonych na działanie karcynogenów i prawdopodobieństwo wystąpienia mutacji w komórce.

Zmiany genetyczne w komórkach zachodzą pod wpływem działania czynników mutagennych, do których można zaliczyć: promieniowanie UV (czynniki fizyczne), substancje obecne w dymie papierosowym i spalinach samochodowych, azbest i niektóre metale ciężkie (czynniki chemiczne), wirusy, toksyny bakteryjne i pasożytnicze oraz błędy podczas replikacji czy pośrednie produkty przemiany materii tj. hormony i wolne rodniki (czynniki biologiczne) [28].

Jedną z podstawowych cech komórek nowotworowych jest zdolność do endogennej produkcji sygnałów mitogennych bez udziału zewnętrznych czynników wzrostu, dlatego choroby nowotworowe są tak niebezpieczne a ich zwalczanie tak potrzebne.

5.2. Projekt The Cancer Genome Atlas

Do analizy wykorzystano dane udostępniane w ramach *The Cancer Genome Atlas* [8]. The Cancer Genome Atlas (TCGA) to kompleksowy projekt o skoordynowanych wysiłkach, mający na celu przyśpieszenie zrozumienia molekularnych podstaw raka. Ma się to odbywać poprzez stosowanie technologii analizy danych dużej skali do udostępnianych danych genomicznych i zsekwencjonowanego genomu tkanek nowotworowych. TCGA to inicjatywa *National Cancer Institute* (NCI) oraz *National Human Genome Research Institute* (NHGRI), czyli 2 z pośród 27 instytutów i centrów Narodowego Instytutu Zdrowia w Departamencie Zdrowia i Opieki Społecznej Stanów Zjednoczonych.

Jak podaje [58], w biologii jest znane ponad 200 różnych form nowotworu i jeszcze więcej ich podtypów. Każdy z nich wywołany jest błędami w DNA, które sprawiają, że komórki rozrastają się w sposób niekontrolowany. Identyfikując zmiany w kompletnym zbiorze DNA (genomie) dla każdego nowotworu i zrozumiawszy jak te zmiany wpływają na jego rozrost, możliwe będzie stworzenie podstaw pod poprawne zapobieganie nowotworom, wczesne ich wykrywanie i leczenie.

Nadrzędnym celem TCGA jest poprawić naszą zdolność do diagnozowania, leczenia i profilaktyki raka. TCGA aby osiągnąć ten cel w sposób naukowo rygorystyczny, początkowo jako projekt pilotażowy, opracowało i przetestowało strukturę badawczą konieczną do systematycznego badania całego spektrum zmian genomu zawartych w ponad 20 rodzajach raka.

Dzięki projektowi TCGA społeczność naukowców walczących z rakiem może korzystać z danych o zsekwencjonowanych tkankach nowotworowych zebranych przez instytucje takie jak *Cancer Genomics Hub* (CGHub) czy *Genome Data Analysis Centers* (GDACs). Te i wiele więcej instytucji opisanych szerzej jest w [58]. TCGA Genome Data Analysis Centers składają się z 7 instytucji, a jedna z nich *Broad Institute, Cambridge* udostępnia dane oraz wyniki swoich analiz poprzez portal Firehose Broad GDAC (`http://gdac.broadinstitute.org/`). Portal udostępnia dane dla 38 kohort związanych z występującym typem nowotworu. Więcej na temat kohort i danych zawartych w TCGA można znaleźć w [10], [11], [55] czy [58].

Dane z TCGA pobrano przy użyciu pakietu `RTCGA` [33] i umieszczono w pakietach

- `RTCGA.clinical` [34], zawierającym dane kliniczne o pacjentach,
- `RTCGA.mutations` [35], zawierającym dane o występujących w genach mutacjach

oraz zostały one wykorzystane w następnym rozdziale do sprawdzenia czy występowanie mutacji w danym genie ma wpływ na przeżywalność pacjentów dotkniętych nowotworem.

5.3. Analiza

5.4. Wnioski

Dodatek A

Wykorzystane narzędzia, dokumentacja i kody pakietu $\mathcal{R}$ użyte w pracy

A.1. Implementacje optymalizacji w regresji logistycznej

A.2. Model proporcjonalnych hazardów Coxa

```
checkArguments <- function(formula, data, learningRates,
                           beta_0, epsilon) {
  assert_that(is.list(data) & length(data) > 0)
  assert_that(length(unique(unlist(lapply(data, ncol)))) == 1)
  # + check names and types for every variables
  assert_that(is.function(learningRates))
  assert_that(is.numeric(epsilon))
  assert_that(is.numeric(beta_0))

    # check length of the start parameter
  if (length(beta_0) == 1) {
    beta_0 <- rep(beta_0, as.character(formula)[3] %>%
                  strsplit("\\+") %>%
                  unlist %>%
                  length)
  }
  return(beta_0)
}
```

coxphSGD

November 2, 2015

---

coxphSGD	*Stochastic Gradient Descent log-likelihood estimation in Cox proportional hazards model*

Description

Function coxphSGD estimates coefficients using stochastic gradient descent algorithm in Cox proportional hazards model.

Usage

```
coxphSGD(formula, data, learningRates = function(x) {
    1/x
}, beta_0 = 0, epsilon = 1e-05)
```

Arguments

formula	a formula object, with the response on the left of a ~ operator, and the terms on the right. The response must be a survival object as returned by the Surv function.
data	a list of batch data.frames in which to interpret the variables named in the formula. See Details.
learningRates	a function specifing how to define learning rates in steps of the algorithm. By default the f(t)=1/t is used, where t is the number of algorithm's step.
beta_0	a numeric vector (if of length 1 then will be replicated) of length equal to the number of variables after using formula in the model.matrix function
epsilon	a numeric value with the stop condition of the estimation algorithm.

Details

A data argument should be a list of data.frames, where in every batch data.frame there is the same structure and naming convention for explanatory and survival (times, censoring) variables. See Examples.

Note

If one of the conditions is fullfiled (j denotes the step number)

- $||\beta_{j+1} - \beta_j|| <$ `epsilon` parameter for any j
- $j >$ `#batches`

the estimation process is stopped.

Examples

```
library(survival)
## Not run:
coxphSGD(Surv(time, status) ~ ph.ecog + age, data = list(lung[1:50, ], lung[51:100,
    ], lung[101:150, ], lung[151:228, ]))

## End(Not run)
```

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] Bender R., Augustin T., Blettner Maria, (2005) *Generating survival times to simulate Cox proportional hazards models*, Statistics in Medicine, Volume 24, Issue 11, 1713–1723.

[4] Biecek P., (2011), *Przewodnik po pakiecie R*, Rozprawa doktorska, Oficyna Wydawnicza GiS, wydanie II.

[5] Bottou L., (2010), *Large-Scale Machine Learning with Stochastic Gradient Descent.*

[6] Bottou L., (1998), *Online Learning and Stochastic Approximations.*

[7] Bottou L., (2012), *Stochastic Gradient Descent Tricks.*

[8] Broad Institute TCGA Genome Data Analysis Center (2014): Firehose stddata 2015 06 01 run. Broad Institute of MIT and Harvard. DOI:10.7908/C1251HBG

[9] Burzykowski T., (2015?), *Notatki do przedmiotu Biostatystyka*, `https://e.mini.pw.edu.pl/sites/default/files/biostatystyka.pdf`.

[10] Chin L., Hahn W.C., Getz G., Meyerson M., (2011), *Making sense of cancer genomic data. Genes and Development*, 25(6): 534-555.

[11] Chin L., Andersen J.N., Futreal P.A., (2011), *Cancer genomics: from discovery science to personalized medicine*, Nature Medicine, 17(3): 297-303.

[12] Cox D. R. (1962), *Renewal Theory. Methuen Monograph on Applied Probability & Statistics*, London: Methuen.

[13] Cox D. R., (1972), *Regression models and life-tables (with discussion)*, Journal of the Royal Statistical Society Series B 34:187-220.

[14] Collett D., (1994), *Modelling Survival Data in Medical Research*, Chapman and Hall: London.

[15] Czepiel S. A., *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, `http://czep.net/stat/mlelr.pdf`.

[16] Dennis J. E. Jr., Schnabel R. B., (1983), *Numerical Methods For Un- constrained Optimization and Nonlinear Equations*, Prentice-Hall.

[17] Dobson A. J., (2002), *An Introduction to Generalized Linear Models*, Wydanie II, Chapman & Hall/CRC.

[18] Domagała W., (2007), *Molekularne podstawy karcynogenezy i ścieżki sygnałowe niektórych nowotworów ośrodkowego układu nerwowego*, Polski Przegląd Neurologiczny, tom 3: 127-141.

[19] Finkel J. R., Kleeman A., Manning C. D., (2008), *Efficient, Feature-based, Conditional Random Field Parsing*, Proc. Annual Meeting of the ACL.

[20] Fisher R. A., (1912) *An absolute criterion for fitting frequency curves.*

[21] Fisher R. A., (1922) *On the mathematical foundations of theoretical statistics*, Philos. Trans. Roy. Soc. London Ser. A 222 309-368.

[22] Fortuna Z., Macukow B., Wąsowski J., (2006), *Metody Numeryczne*, Wydawnictwa Naukowo-Techniczne.

[23] Gauss C. F., (1809,) *Theoria Motus Corporum Coelestium.*

[24] Gągolewski M., (2014), *Programowanie w języku R*, Wydawnictwo Naukowe PWN.

[25] Hald A., (1949), *Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point*, Skandinavisk Aktuarietidskrift, 119-134.

[26] Hastie T. J., Pregibon D., (1992), *Generalized linear models. Chapter 6 of Statistical Models in S*, eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

[27] Hutchinson J. B., (1928), *The Application of the "Method of Maximum Likelihood" to the Estimation of Linkage*, Genetics. 1929 Nov; 14(6): 519–537.

[28] Janik-Papis K., Błasiak J., (2010), *Molekularne wyznaczniki raka piersi. Inicjacja i promocja - część I*, Nowotwory - Journal of Oncology, 60 (3): 236-247.

[29] Janik-Papis K., Błasiak J., (2010), *Molekularne wyznaczniki raka piersi. Progresja i nowi kandydaci - część II*, Nowotwory - Journal of Oncology, 60 (4): 341-350.

[30] Jennrich R. I., Sampson P. F., (1976), *Newton-Raphson and related algorithms for maximum likelihood variance component estimation*, Technometrics, 18, 11-17.

[31] 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.

[32] Kosiński M., (2015), *coxphSGD: Stochastic Gradient Descent log-likelihood estimation in Cox proportional hazards model*, R package version 0.0.3, `https://github.com/MarcinKosinski/Cox-SGD`.

[33] Kosiński M., Biecek P., (2015), *RTCGA: The Cancer Genome Atlas Data Integration*, R package version 1.1.7, `https://github.com/RTCGA`.

[34] Kosiński M., (2015), *RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project*, R package version 20150821.1.2, `https://bioconductor.org/packages/release/data/experiment/html/RTCGA.clinical.html`

[35] Kosiński M., (2015), *RTCGA.mutations: Clinical datasets from The Cancer Genome Atlas Project*, R package version 20150821.1.1, `https://bioconductor.org/packages/release/data/experiment/html/RTCGA.mutations.html`

[36] Kotłowski W., (2012), Notatki do przedmiotu *Techniki Optymalizacji* prowadzonego na Politechnice Poznańskiej, `http://www.cs.put.poznan.pl/wkotlowski/teaching/wyklad3b.pdf`

[37] Kozłowska J., Łaczmańska I., (2010), *Niestabilność genetyczna - jej znaczenie w procesie powstawania nowotworów oraz diagnostyka laboratoryjna*, Nowotwory - Journal of Oncology, 60 (6): 548-553.

[38] Legendre A. M., (1804), *Nouvelles mèthods pour la dètermination des orbites des comètes.*

[39] Longford N. T., (1987), *A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects*, Biometrika 74 (4): 817-827.

[40] Milewski S., (2006), Konspekt do przedmiotu *Metody Numeryczne* prowadzonego na Politechnice Krakowskiej,
`http://l5.pk.edu.pl/images/skrypty/Metody_numeryczne_1`

[41] 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.

[42] Mood A. M., Graybill F. A., Boes D. C., (1974), *Introduction to the Theory of Statistics*, McGraw-Hill: New York.

[43] Murata N., (1998), *A Statistical Study of On-line Learning. In Online Learning and Neural Networks*, Cambridge University Press.

[44] Niemiro W., (2011), Skrypt do przedmiotu *Statystyka* prowadzonego na Uniwersytecie Warszawskim,
`http://www-users.mat.umk.pl/~wniem/Statystyka/Statystyka.pdf`

[45] Norwegian Multicentre Study Group, (1981), *Timolol-induced reduction in mortality and reinfarction*, The New England Journal of Medicine; 304: 801-7.

[46] 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/`

[47] 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/`

[48] Podsiadły K., (2011), *Genetyczne podstawy nowotworzenia*, `www.e-biotechnologia.pl/Artykuly/Genetyczne-podstawy-nowotworzenia`.

[49] 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/`.

[50] 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.

[51] Rydlewski J., (2009), *Estymatory Największej Wiarogodności w Uogólnionych Modelach Regresji Nieliniowej*, Rozprawa doktorska.

[52] Sokołowski A., (2010), *Jak rozumieć i wykonywać analizę przeżycia http://www.statsoft.pl/Portals/0/Downloads/Jak_rozumiec_i_wykonac_analize_przezycia.pdf*

[53] 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.*

[54] Tran D., Lan T., Toulis P., (2015), *sgd: Stochastic Gradient Descent for Scalable Estimation. R package version 0.1*, `https://github.com/airoldilab/sgd`.

[55] *The future of cancer genomics*, Nature Medicine, (2015), 21(2): 99.

[56] Therneau T. M., Grambsch P. M., (2000), *Modeling Survival Data: Extending the Cox Model*, Springer.

[57] Therneau T. M., (2015), *A Package for Survival Analysis in S. version 2.38*, `http://CRAN.R-project.org/package=survival`.

[58] Tomczak K., Czerwińska P., Wiznerowicz M., (2015), *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge*, Contemporary Oncology. 19(1A): A68-A77.

[59] Widrow B., (1960), *An adaptive "ADALINE" neuron using chemical "memistors"*, Technical Report No. 1553-2, Stanford University.

[60] Widrow B., Ho M.E., (1960), *Adaptive switching circuits*, In: IRE WESCON Conv. Record, Part 4. pp. 96-104.

[61] Widrow B., Stearns S. D., (1985), *Adaptive Signal Processing, Prentice Hall.*

[62] Wikipedia, encyklopedia wolnego dostępu `wikipedia.pl`

[63] 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`

[64] Zieliński R., (1990), *Siedem wykładów wprowadzających do statystyki matematycznej*, Warszawa, Wydawnictwo Naukowe PWN.

Marcin Piotr Kosiński
Nr albumu 265361

Warszawa, 25 listopada 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