Uogólnione modele liniowe – materiały do wykładu

Agnieszka Piliszek

Wydział Matematyki i Nauk Informacyjnych

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja — Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Spis treści

Rozdział 1

Uogólnione Modele Liniowe: wprowadzenie

All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.
Box, Draper [Empirical Model-Building and Response Surfaces. Wiley, New York (1987), s. 424]

1.1 Podstawowe pojęcia i notacja

Będziemy rozważać **zmienne losowe** Y_i (*zmienna objaśniana*) i $X_{i1}, \ldots, X_{ip}$ dla $i = 1, \ldots, N$ (*zmienne objaśniające, predyktory*), $N \in \mathbb{N}$, $p \in \mathbb{N}$. Zbiór liczb naturalnych $\mathbb{N} = \{1, 2, 3, \ldots\}$. W razie potrzeby posługiwać się będziemy symbolem $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Obserwacje pochodzące z odpowiednich zmiennych losowych będziemy oznaczać małymi literami: $(y_i, x_{i1}, \ldots, x_{ip})$ jest wektorem obserwacji z $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$.

Model regresyjny liniowy w parametrach (ang. *Regression model linear in the parameters*)

Założenia: $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$ – homoskedastyczność, $\mathbb{C}\mathrm{ov}(\varepsilon_i, \varepsilon_j) = 0$ dla $j \neq i$.
Model:

$$Y_i = f\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right) + \varepsilon_i, \quad \beta_j \in \mathbb{R},\ j = 1, \ldots, p,\ \ i = 1, \ldots, N.$$

Inna postać:

$$\mu_i = f\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right),$$

gdzie $\mu_i := \mathbb{E}(Y_i)$.

Jeśli $f \equiv \mathrm{id}$, to mamy model liniowy; o zmiennych losowych Y_i zakłada się, że są z rozkładu normalnego: $\mathbf{Y}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ (składowa losowa modelu). Jeśli nie, to mamy **uogólniony model liniowy** (*generalized linear model*), w skrócie GLM. Wtedy wprowadzamy pojęcie **funkcji łączącej** (ang. *link function*) g, takiej że

$$\mu_i = g^{-1}\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right) \tag{1.1}$$

lub równoważnie

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}.$$

O funkcji łączącej g zakłada się zwykle, że jest **różniczkowalna, monotoniczna, odwracalna**.

Składowa losowa modelu: Y_i, $i = 1, \ldots, n$, pochodzą z określonego rozkłady z rodziny wykładniczej (EDM). Definicję EDM podamy później. Ważnymi przykładami rozkładów z tej rodziny są: normalny, gamma, Poissona.

Notacja macierzowa

$\mathbf{X}_i^T = (1, X_{i1}, \ldots, X_{ip})$

$\underline{\boldsymbol{\beta}}_i^T = (\beta_0, \beta_1, \ldots, \beta_p)$

$Y_i = \mathbf{X}_i^T \cdot \underline{\boldsymbol{\beta}}_i + \varepsilon$

Wtedy

$$\mathbf{Y} = f\left(\mathbf{X}^T \underline{\boldsymbol{\beta}}\right) + \underline{\varepsilon},$$

gdzie $\mathbf{X}$ jest macierzą eksperymentu i

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{n1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1p} & \cdots & X_{np} \end{bmatrix}.$$

GLM:

$$g(\underline{\mu}) = \mathbf{X}^T \underline{\boldsymbol{\beta}},$$

gdzie $\underline{\mu} = (\mu_1, \ldots, \mu_n)$.

Analogiczne oznaczenia wprowadzamy dla obserwacji: $\underline{\mathbf{x}}$, $\underline{\mathbf{x}}_i$, $\underline{\mathbf{y}}$.

1.2 Uwagi ogólne

Zmienne objaśniające dzielą się na:

- ilościowe (*quantitative*)
- jakościowe (faktory, *qualitative*) przykład indykatorowego oznaczania faktorów;

– Zmienne jakościowe, które są nieuporządkowane: nie ma żadnej kolejności: czarny $\not\geq$ biały – zmienne **nominalne**;

– Zmienne jakościowe, które są uporządkowane – **zmienne porządkowe** (*ordinal*); np. wykształcenie: szkoła podstawowa<szkoła średnia<studia wyższe lic<mgr<dr<... – można je traktować dwojako: jak indykatorowe (i zapomnieć o uporządkowaniu) albo jak ilościowe, np. reprezentowane za pomocą liczb $1, 2, 3, 4, \ldots$.

Interpretacja: nie indywidualnie tylko podpopulacyjnie.

Przykład: nauka matematyki

Niech y_i – wynik testu z matematyki $i$tego studenta, x_{i1} – liczba lat edukacji matematycznej $i$tego studenta. Rozważamy model

$$\mu_i = \beta_0 + \beta_1 x_{i1}.$$

- Błędna interpretacja: „jeśli ten model jest prawdziwy to wzrost o 1 rok edukacji danego studenta powoduje wzrost o β_1 wyniku z testu";
- Poprawna interpretacja: „jeśli ten model jest prawdziwy, to jeśli porównamy populację studentów którzy uczyli się matematyki x lat z populacją studentów, którzy uczyli się $x + 1$ lat, to różnica średnich wyników będzie wynosiła β_1".

Błędna interpretacja, która się tutaj pojawiła, zakładała związek przyczynowo–skutkowy. Nasz model nie uwzględnia takiego związku. (Są jednak modele, które badają przyczynowość, patrz rozdział 6.)

Rozdział 2

Modelowanie proporcji – model logistyczny

Dane binarne to dane, w których zmienna odpowiedzi przyjmuje dwie wartości, oznaczane (zwykle) jako 0 i 1. Zwykle 1 odpowiada pewnemu wyróżnieniu, np. posiadaniu jakiejś cechy czy odniesieniu sukcesu (bycie młodym, bycie chorym itp.), ale nie musi tak być – możemy przecież rozważać dwie opcje, które są tak samo ważne (np. bycie kobietą i bycie mężczyzną). Spójrzmy na losową składową GLM dla takich danych.

Niech Y będzie binarną zmienną odpowiedzi. Niech $\pi := \mathbb{P}(Y = 1) = \mu$, a zatem $1 - \mu = 1 - \pi = \mathbb{P}(Y = 0)$. Zauważmy, że $\mathbb{P}(Y = y) = \pi^y (1 - \pi)^{1-y}$, $y \in \{0, 1\}$. Ponieważ $\mu = \mathbb{E}Y = \pi$, to można rozkład binarny traktować jak rodzinę rozkładów parametryzowanych przez $\mu \in (0, 1)$:

$$P(y, \mu) = \mu^y (1 - \mu)^{1-y}, \quad y \in \{0, 1\}.$$

Wówczas $\mathbb{V}\mathrm{ar}(Y) = \mu(1 - \mu)$.

Możliwa jest też inna parametryzacja, która istnieje dla wszystkich rozkładów z rodziny wykładniczej (patrz rozdział 4).

Parametryzacja kanoniczna

Zauważmy, że

$$P(y, \mu) = \exp\left\{y \log\left(\tfrac{\mu}{1-\mu}\right) + \log(1 - \mu)\right\} =: \exp\{y\,\theta - \kappa(\theta)\} \tag{2.1}$$

gdzie

$$\theta = \log\left(\tfrac{\mu}{1-\mu}\right) \quad \Leftrightarrow \quad \mu = \tfrac{e^\theta}{1+\mathbf{e}^\theta} = \tfrac{1}{1+e^{-\theta}}$$
$$\kappa(\theta) = -\log(1 - \mu) = \log(e^\theta + 1)$$

W jaki sposób dobrać składową systematyczną? Poza spełnianiem warunków technicznych, o których powiemy później, funkcja łącząca musi być sensowna. Zobaczmy co to oznacza dla rozkładu dwupunktowego i jakie funkcje łączące są (w tym sensie) dopuszczalne.

Niech $\tilde{Y}$ będzie (ukrytą) **ciągłą** zm. los. taką, że $\{Y = 1\} = \{\tilde{Y} > \theta\}$, (równoważnie $\tilde{Y}^{-1}[(\theta, \infty)] = Y^{-1}[\{1\}]$) gdzie $\theta \in \mathbb{R}$ jest nieznaną wartością krytyczną. Rozważmy model

$$\tilde{Y} = \mathbf{X}^T \underline{\boldsymbol{\beta}} + \varepsilon,$$

gdzie $-\varepsilon$ jest zmienną losową o dystrybuancie F odpowiadającą za szum losowy. Zakładamy, że ε i $\mathbf{X}$ są niezależne. Wówczas

$$\pi(\underline{\mathbf{x}}) = \mathbb{P}(Y = 1 \,|\, \underline{\mathbf{x}}) = \mathbb{P}(\tilde{Y} \geq \theta \,|\, \underline{\mathbf{x}}) = \mathbb{P}(\varepsilon \geq \theta - \mathbf{X}^T\beta \,|\, \underline{\mathbf{x}}) = \mathbb{P}(\varepsilon \geq \theta - \underline{\mathbf{x}}^T\beta \,|\, \underline{\mathbf{x}}) = F(\underline{\mathbf{x}}^T\tilde{\beta}),$$

gdzie $\tilde{\beta} = [-\theta, 0, \ldots, 0]^T + \beta$. (Zwróćmy uwagę na przedostatnią równość, która wynika z niezależności ε i $\mathbf{X}$.) Zauważmy, że dla dowolnej ciągłej dystrybuanty F:

(a) $\pi(\underline{\mathbf{x}}) \in [0; 1]$,

(b) θ (nieznana wartość krytyczna) *znika* w $\tilde{\beta}$.

Szeroko stosowanymi przykładami F są:

- dystrybuanta rozkładu normalnego $N(0, 1)$ – model probitowy,
- $F(x) = \frac{e^x}{1+e^x}$ – model logistyczny,
- $F(x) = \exp(-\exp(x))$ – model loglog,
- $F(x) = 1 - \exp(-\exp(x))$ – model conplementary loglog.

Zgodnie z (1.1) $g = F^{-1}$.

2.1 Dane zgrupowane – model dwumianowy

Rozważmy dane:

i	Y	X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
3	y_3	x_{31}	x_{32}	...	x_{3p}
..	..	..	..	..	..
j	y_j	x_{j1}	x_{j2}	...	x_{jp}
..	..	..	..	..	..
n	y_n	x_{n1}	x_{n2}	...	x_{np}

gdzie $Y(\omega) \in \{0,1\}$ oraz $\pi = \mathbb{P}(Y=1) = 1 - \mathbb{P}(Y=0)$.

Może się zdarzyć, że dla różnych $k \in \{1,\ldots,n\}$ wektory $(x_{j1}, x_{j2}, \ldots, x_{jp})$ i $(x_{k1}, x_{k2}, \ldots, x_{kp})$ będą takie same. Wówczas warto[1] zgrupować dane ze względu na wektor $\underline{x} = (x_1, \ldots, x_p)$. Zbiór indeksów wpadających do $i$tej grupy oznaczmy przez A_i, $i = 1, \ldots, N$. W ten sposób otrzymujemy przekształcone dane:

i	m_i	Z	X_1	X_2	...	X_p
1	m_1	z_1	x_{11}	x_{12}	...	x_{1p}
2	m_2	z_2	x_{21}	x_{22}	...	x_{2p}
3	m_3	z_3	x_{31}	x_{32}	...	x_{3p}
..	..	..	..	..	..	..
j	m_j	z_j	x_{j1}	x_{j2}	...	x_{jp}
..	..	..	..	..	..	..
N	m_N	z_N	x_{N1}	x_{N2}	...	x_{Np}

gdzie $m_j = |A_j|$ jest licznością $j$tej grupy, a $z_j = \sum_{i \in A_j} y_i$ jest łączną liczbą sukcesów w $j$tej grupie. Oczywiście, $m_1 + \ldots + m_N = n$.

Proporcje sukcesów defniujemy jako $F := Z/m$, gdzie $Z \sim \text{bin}(m, \pi)$. Z własności wartości oczekiwanej $\mathbb{E}F = \pi = \mathbb{E}Y$. Natomiast wariancja

$$\mathbb{V}\text{ar}(F) = \frac{1}{m^2}\,\mathbb{V}\text{ar}(Z) = \frac{\pi(1-\pi)}{m} < \mathbb{V}\text{ar}(Y).$$

Przykład

Mamy dziesięciu pacjentów z 2 różnych szpitali (X_1), których poddajemy trzem różnym terapiom (X_2). Wynik leczenia kodujemy za pomocą zmiennej Y: $\{y_i = 1\}$ oznacza, że $i$ty pacjent wyzdrowiał, a $\{y_i = 0\}$ oznacza, że nie wyzdrowiał.

[1] Chociażby ze względu na pamięć i szybkość obliczeń!

Przykład

i	Y	X_1	X_2
1	0	1	1
2	0	1	2
3	0	2	1
4	0	1	2
5	1	1	2
6	1	1	1
7	1	1	2
8	1	2	2
9	1	1	3
10	1	2	3

Dane te możemy zgrupować i zapisać w mniejszej tabeli:

i	m_i	Z_i	F_i	X_1	X_2
1	2	1	0.5	1	1
2	4	2	0.6	1	2
3	1	0	0	2	1
4	1	1	1	2	2
5	1	1	1	1	3
6	1	1	1	2	3

2.2 Estymator największej wiarogodności

2.2.1 *Score function* – funkcja oceny

Score-function to pochodna funkcji log-wiarogodności po wektorze $\underline{\boldsymbol{\beta}}$:

$$S(\underline{\boldsymbol{\beta}}) := \frac{\partial \log P(y,\, \mu)}{\partial \underline{\boldsymbol{\beta}}}.$$

Im większa jest jej wartość, tym możemy mówić o lepszej estymacji. (Pytanie do czytelnika: dlaczego?)

Stwierdzenie

A modelu logistycznym

$$S(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} [(y_i - \mu_i)x_{i1}, \ldots, (y_i - \mu_i)x_{ip}]^T = \sum_{i=1}^{n} \underline{\mathbf{x}}_i \cdot (y_i - \mu_i).$$

Poniżej znajduje się rozumowanie dowodzące prawdziwości powyższego stwierdzenia.

Dla funkcji prawdopodobieństwa postaci danej w (2.1), log–wiarogodność (*log-likelihood*) wynosi

$$\log P(y, \mu) = y\,\theta - \kappa(\theta).$$

Zauważmy, że $\kappa'(\theta) = e^{\theta}/\left(1 + e^{\theta}\right)$. Stąd

$$\frac{\partial \log P(y,\, \mu)}{\partial \theta} = y - \mu.$$

Natomiast (z reguły łańcuchowej)

$$\frac{\partial \log P(y,\, \mu)}{\partial \mu} = \frac{\partial \log P(y,\, \mu)}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} = (y - \mu) \cdot \frac{1 - \mu}{\mu} \frac{1 - \mu + \mu}{(1 - \mu)^2} = (y - \mu) \cdot \tfrac{1}{\mu(1-\mu)} = (y - \mu) \frac{1}{\mathbb{V}\mathrm{ar}Y}.$$

Uwaga

To nie jest przypadek!

Dla próby : $y_1, \ldots, y_n$.

Przypomnienie: $\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}} = \theta = \log\left(\frac{\mu_i}{1-\mu_i}\right)$. Równoważnie $\mu_i = h(\theta_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}}$.

Wiarogodność:

$$L(\underline{\boldsymbol{\beta}}) := \prod_{i=1}^{n} P(y_i,\, \mu_i) = \prod i = 1^n \exp\left\{y\,\theta_i - \kappa(\theta_i)\right\},$$

gdzie $\kappa(\theta_i) = \log(e^{\theta_i} + 1)$.

Log-wiarogodność:

$$l(\underline{\boldsymbol{\beta}}) := \log L(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \log P(y_i,\, \mu_i) = \sum_{i=1}^{n} y_i\,\theta_i - \kappa(\theta_i).$$

Score-functions:

$$S(\beta_j) = \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \log P(y_i,\, \mu_i)}{\partial \beta_j}.$$

Natomiast

$$\frac{\partial \log P(y_i,\, \mu_i)}{\partial \beta_j} = \frac{\partial \log P(y_i,\, \mu_i)}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} = (y - \mu) \frac{1}{\mu(1 - \mu)} \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j}.$$

Ponieważ

$$\frac{\partial \mu_i}{\partial \theta} = \frac{e^{\theta}}{(1 + e^{\theta})^2} = \mu_i(1 - \mu_i)$$

oraz

$$\frac{\partial \theta_i}{\partial \beta_j} = x_j$$

to

$$S(\beta_j) = \frac{\partial \log P(y_i,\, \mu_i)}{\partial \beta_j} = (y_i - \mu_i)\, x_{ij}.$$

2.2.2 Estymator największej wiarogodności

Estymatorem największej wiarogodności (MLE) $\underline{\boldsymbol{\beta}}$ jest $\hat{\underline{\boldsymbol{\beta}}}$ taki, że

$$S(\hat{\underline{\boldsymbol{\beta}}}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \underline{\mathbf{x}}_i\, y_i = \sum_{i=1}^{n} \underline{\mathbf{x}}_i\, \mu_i.$$

(Z prawej strony równości μ jest funkcją $\hat{\underline{\boldsymbol{\beta}}}$.)

W terminach $\hat{\beta}_j$ mamy następujące *nieliniowe równanie estymacji*:

$$\sum_{i=1}^{n} \underline{\mathbf{x}}_i\, y_i = \sum_{i=1}^{n} \underline{\mathbf{x}}_i\, \frac{\exp(\underline{\mathbf{x}}_i^T \hat{\underline{\boldsymbol{\beta}}})}{1 + \exp(\underline{\mathbf{x}}_i^T \hat{\underline{\boldsymbol{\beta}}})}.$$

Uwaga 2.1 (Istnienie MLE)**.** *Estymator ML dla binarnych danych istnieje wtedy i tylko wtedy, gdy nie ma separacji lub quasi-separacji. Tzn.* **nie** *istnieje wektor $\theta \in \mathbb{R}^p$ taki, że $\underline{\mathbf{x}}_i^t \theta \geq 0$ gdy $y_i = 1$ i $\underline{\mathbf{x}}_i^t \theta \leq 0$ gdy $y_i = 0$. Zauważmy, że jeśli istnieje taki θ, to oznacza, że mamy hiperpłaszczyznę, która oddziela zera od jedynek (i w związku z tym nie musimy nic estymować, tylko klasyfikować za pomocą tej hiperpłaszczyzny).*

W przypadku istnienia MLE $\underline{\boldsymbol{\beta}}$ nie umiemy znaleźć $\hat{\underline{\boldsymbol{\beta}}}$ analitycznie. Dlatego potrzebne są numeryczne metody znajdowania przybliżonego $\hat{\underline{\boldsymbol{\beta}}}$.

2.2.3 Metoda Newtona-Raphsona

Jest to metoda iteracyjna szukania miejsca zerowego funkcji. Chcemy znaleźć $\hat{\zeta}$ takie, że $S(\hat{\zeta}) = 0$. Metoda Newtona–Raphsona polega na przybliżeniu funkcji (w naszym przypadku funkcji S) dwoma wyrazami jej rozwinięcia Taylora. Stąd otrzymujemy krok iteracyjny:

$$\hat{\zeta}^{(k+1)} = \hat{\zeta}^{(k)} + \mathcal{I}_{\text{obs}} \left(\hat{\zeta}^{(k)}\right)^{-1} \cdot S\left(\hat{\zeta}^{(k)}\right),$$

gdzie $\mathcal{I}_{\text{obs}}$ jest macierzą informacji obserwowanej w modelu logitowym i

$$\mathcal{I}_{\text{obs}}(\underline{\boldsymbol{\beta}}, y_i) = -\frac{\partial^2 l(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}} \partial \underline{\boldsymbol{\beta}}^T}.$$

Policzmy element macierzy $\mathcal{I}_{\text{obs}}$:

$$-\frac{\partial^2 l(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}}_j \partial \underline{\boldsymbol{\beta}}_k^T} = -\frac{\partial (y_i - \mu_i) x_{ij}}{\partial \beta_k} = x_{ij} \mu_i (1 - \mu_i) x_{ik} = x_{ij} x_{ik} \frac{\exp(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}})}{(1 + \exp(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}))^2}.$$

Stąd

$$\mathcal{I}_{\text{obs}}(\underline{\boldsymbol{\beta}}, y_i) = \mathbf{X}^T \cdot W \cdot \mathbf{X},$$

gdzie

$$W = \text{diag}\left(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n)\right).$$

ZAUWAŻMY, że $\mathcal{I}_{\text{obs}}$ nie zależy od $\underline{\mathbf{y}}$! Ogólnie nie musi tak być.

Mamy zatem

$$\underline{\hat{\boldsymbol{\beta}}}^{(k+1)} = \underline{\hat{\boldsymbol{\beta}}}^{(k)} + \left(\mathbf{X}^T \cdot W \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \left(\underline{\mathbf{y}} - \mu^{(r)}\right).$$

Przykład

Metoda N.-R. zastosowana do maksymalizacji funkcji log-wiarogodności bazującej na pojedyńczej obserwacji (tzn. $n = 1$) z rozkładu $\text{bin}(N, p)$ (N – ustalone, estymujemy p). $l(p) = y \log p + (N - y)\log(1 - p)$. Pochodne:

$$\frac{\partial l}{\partial p} = \frac{y - Np}{p(1-p)}, \qquad \frac{\partial^2 l}{\partial p^2} = -\left(\frac{y}{p^2} + \frac{N-y}{(1-p)^2}\right).$$

W tym przypadku $\mathcal{I}_{\text{obs}}$ zależy od y! Stąd

$$p^{(k+1)} = p^{(k)} + \left[\frac{y}{(p^{(k)})^2} + \frac{N-y}{(1-p^{(k)})^2}\right]^{-1} \frac{y - Np^{(k)}}{p^{(k)}(1-p^{(k)})}$$

Jeśli

- $p^{(0)} = \frac{1}{2}$, to $p^{(1)} = \frac{y}{n} = \hat{p}$. To jest dobrze, bo $\frac{\partial l}{\partial p}\Big|_{p=y/n} = 0$.
- $p^{(0)} \neq \frac{1}{2}$, to zbieżność po kilku iteracjach.

Spróbuj sam!

Napisz funkcję `NR(y, n, x0)` zwracającej argmax i max (np. w formie wektora) funkcji log-wiarogodności dla pojedyńczej obserwacji `y` z rozkładu $\text{bin}(n, p)$. Przetestuj dla kilku wybranych wartości `x0`, `y` i `n`. Czy nasuwają się jakieś wnioski?

2.2.4 Algorytm Fisher Scoring

Modyfikacje algortymu N.R. - zamiast macierzy informacji obserwowanej wstawiamy jej wartość oczekiwaną, tzn.

$$\mathcal{I}\left(\underline{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\mathcal{I}_{\text{obs}}(\underline{\boldsymbol{\beta}}, \mathbf{Y})\right).$$

Macierz ta nazywana jest *macierzą informacji oczekiwanej* lub *macierzą informacji Fishera.*
Krok iteracyjny jest teraz postaci

$$\hat{\zeta}^{(k+1)} = \hat{\zeta}^{(k)} + \mathcal{I}\left(\hat{\zeta}^{(k)}\right)^{-1} \cdot S\left(\hat{\zeta}^{(k)}\right),$$

Przykład c.d.

$$\mathcal{I}(\underline{\boldsymbol{\beta}}) = -\mathbb{E}\left(\frac{\partial^2 l}{\partial p^2}\right) = \frac{\mathbb{E}Y}{p^2} + \frac{N - \mathbb{E}Y}{(1-p)^2}.$$

Ponieważ $\mathbb{E}Y = Np$, to

$$\mathcal{I}(\underline{\boldsymbol{\beta}}) = \frac{N}{p} + \frac{N}{1-p} = \frac{N}{p(1-p)}.$$

Stąd krok iteracyjny jest następujący:

$$p^{(k+1)} = p^{(k)} + \left(\frac{N}{p^{(k)}(1-p^{(k)})}\right)^{-1} \cdot \frac{y - Np^{(k)}}{p^{(k)}(1-p^{(k)})} = \frac{y}{N}.$$

Jest lepiej niż w N.R.! Od razu dostajemy, że $p^{(r)} = \frac{y}{N}$ dla każdego $r \geq q$ (niezależnie od warunku początkowego).

W przypadku regresji logistycznej (modelu logitowego):

$$S(\underline{\boldsymbol{\beta}}) = \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}}} = \sum_{i=1}^{N} \underline{\mathbf{x}}_i \left(\underline{\mathbf{y}}_i - n_i \pi(x_i)\right)$$

2.3 Inne funkcje łączące

GLM dla proporcji: modelujemy π_i, jeśli $g(\pi_i) = \underline{\boldsymbol{\beta}}^T \underline{\mathbf{x}}_i$, równoważnie $\pi_i = g^{-1}(\underline{\boldsymbol{\beta}}^T \underline{\mathbf{x}}_i)$.
Za g możemy wziąć dowolną funkcję o wartościach w $[0, 1]$. Pożądane jest, aby g była monotoniczna i różniczkowalna. W modelu logistycznym

$$g^{-1}(\underline{\boldsymbol{\beta}}^T \underline{\mathbf{x}}_i) = \frac{e^{\underline{\boldsymbol{\beta}} \underline{\mathbf{x}}_i}}{1 + e^{\underline{\boldsymbol{\beta}} \underline{\mathbf{x}}_i}}.$$

Jest to dystrybuanta rozkładu logistycznego. Mówiąc ściślej, rozkład logistyczny o średniej μ i rozproszeniu $\tau > 0$ ma dystrybuantę

$$F(x) = \frac{\exp[(x-\mu)/\tau]}{1 + \exp[(x-\mu)/\tau]}, \qquad x \in \mathbb{R}.$$

Równie dobrze możemy wziąć inną dystrybunatę...

1. **Regresja probitowa**

$$g(\pi_i) = \Phi^{-1}(\pi_i),$$

gdzie Φ – dystrybuanta $\mathbb{N}(0,1)$ (można ogólniej wziąć $\mathbb{N}(\mu, \sigma^2)$.)

Tak samo jak logitowa regresja probitowa jest symetryczna wokół 1/2, tzn.

$$h(\pi(x)) = -h(1-\pi(x)).$$

W przypadku logitu:

$$\text{logit}(\pi(x)) = \log\frac{\pi(x)}{1-\pi(x)} = -\log\frac{1-\pi(x)}{\pi(x)} = -\text{logit}(1-\pi(x)).$$

2. **Complementary log-log**

$$g(\pi) = \log(-\log(1-\pi)).$$

3. **log-log**

$$g(\pi) = \log(-\log(\pi))$$

Czyli

$$\log(-\log(\pi(x)) = \alpha + \beta x \Rightarrow \pi(x) = \exp[-\exp(\alpha+\beta x)]$$

Zbiega do 0 szybciej niż do 1! (Użyteczny przy niesymetryczności.)

Ten model odpowiada dystrybuancje o rozkładzie Gumbella (jest to jedna z tzw. *extreme value distributions*).

Uwaga: Jeśli model *complementary log-log* zachodzi dla π to *log-log* zachodzi dla $1-\pi$.

2.4 Próbkowanie prospektywne i retrospektywne w kontekście regresji logistycznej

Dany jest model logistyczny:

$$Y_i \sim \text{bn}(1, \pi(\underline{\mathbf{x}}_i)),\ \pi(\underline{\mathbf{x}}_i) = \mathbb{P}(Y_i = 1|\underline{\mathbf{x}}_i) = 1 - \mathbb{P}(Y_i = 0|\underline{\mathbf{x}}_i), \quad i = 1, \ldots, N$$

oraz

$$\pi(\underline{\mathbf{x}}_i) = \pi(\underline{\mathbf{x}}_i; \underline{\boldsymbol{\beta}}) = \frac{\exp(\underline{\boldsymbol{\beta}}^T \underline{\mathbf{x}}_i)}{1+\exp(\underline{\boldsymbol{\beta}}^T \underline{\mathbf{x}}_i)} =: h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i),$$

gdzie $\underline{\boldsymbol{\beta}} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ i $\underline{\mathbf{x}}_i = (1, x_{i1}, \ldots, x_{ip})$.

Żeby zrozumieć ideę próbkowania retrospektywnego i prospektywnego rozważmy dane dotyczące metody karmienia nowonarodzonych dziewcząt i chłopców:

	Butelka	Pierś + butelka	Pierś
Chłopcy	77/458	19 / 147	47/494
Dziewczynki	48/384	16/127	31/464

W tabeli podana jest liczba dzieci chorych do (/) liczby dzieci zdrowych. Dane te mogły być zbierane na różne sposoby. **Prospektynie**, gdy wybieramy próbę z nowonarodzonych dziewczynek i chłopców, których rodzice wybrali konkretną metodę karmienia i patrzymy na ich zdrowie w pierwszym roku. Inna nazwa próbkowania prospektywnego to *cohort study* (badanie kohortowe).

W podejściu **retrospektywnym**, to stan zdrowia jest ustalony, a my patrzymy na predyktory. Np. patrzymy na chore dzieci, które przyszył do lekarza i zapisujemy ich płeć i sposób karmienia. Osobno dostajemy próbkę dzieci zdrowych (tzw. próba kontrolna) i zapisujemy ich płeć i sposób karmienia. Zakładamy przy tym, że to czy osobnik został objęty spisem jest niezależne od wartości predyktorów (czyli sposób karmienia i płeć nie mają wpływu na to czy ktoś przyjdzie do lekarz czy nie). Inna nazwa to *case–control study* (badanie kliniczno–kontrolne, studium przypadku).

W próbkowaniu prospektywnym predyktory (czyli $x$sy) są ustalone i obserwujemy wyniki (czyli $y$eki).

1. Podejście prospektywne: $(\underline{\mathbf{x}}_1, y_1, \ldots, \underline{\mathbf{x}}_n, y_n)$ – dane otrzymane przy próbie prospektywnej, tj. $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n$ – ustalone (nielosowe), $y_1, \ldots, y_n$ – niezależne (losowe). Wiarogodność wynosi wówczas

$$L_{\text{pros}}(\underline{\boldsymbol{\beta}}, \underline{\mathbf{x}}_1, y_1, \ldots, \underline{\mathbf{x}}_n, y_n) = \prod_{j=1}^{n} \left[\pi(\underline{\mathbf{x}}_j, \underline{\boldsymbol{\beta}})^{y_j} \left(1 - \pi(\underline{\mathbf{x}}_j, \underline{\boldsymbol{\beta}})\right)^{1-y_j}\right] = \prod_{j=1}^{n} h^{y_j}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_j)\bar{h}^{1-y_j}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}),$$

   gdzie $\bar{h}(x) = 1 - h(x)$.

2. Podejście retrospektywne: wybieramy próbę n–elementową z listy $(y_1, \underline{\mathbf{x}}_1), \ldots, (y_N, \underline{\mathbf{x}}_N)$ retrospektywnie, tzn. X próbkujemy z naszych zebranych danych takich że $Y = 1$ ('cases') i z takich, że $Y = 0$ ('control'). Dokładniej $X_1, X_2, \ldots, X_{n_0}$ niech będzie próbą z rozkładu $\mathbb{P}(x|Y = 0)$, a $X_1', \ldots, X_{n_1}'$ z rozkładu $\mathbb{P}(x|Y = 1)$.

Niech Z_i – indykator zdarzenia wyboru $i$tego elementu do próby, tzn. $Z_i = 1$ wttwg $i$ty element został wybrany do próby i $Z_i = 0$ wpp. Niech

$$\pi_1 = \mathbb{P}(Z_i = 1|Y_i = 1)$$

$$\pi_0 = \mathbb{P}(Z_i = 1|Y_i = 0).$$

Założenie : π_0 i π_1 nie zależą od $\underline{\mathbf{x}}_j$ i j! Wówczas

$$L_{\text{retro}}(\underline{\boldsymbol{\beta}}, \pi_0, \pi_1, y_1, \underline{\mathbf{x}}_1, \ldots, y_n, \underline{\mathbf{x}}_n) = \prod_{i=1}^{n} \mathbb{P}(Y_i = 1|\underline{\mathbf{x}}_i,\, Z_i = 1)^{y_i} \mathbb{P}(Y_i = 0|\underline{\mathbf{x}}_i,\, Z_i = 1)^{1-y_i}$$

Zauważmy, że

$$\begin{aligned}\mathbb{P}(Y=1|X=\underline{\mathbf{x}},Z=1) &= \frac{\mathbb{P}(Y=1,X=\underline{\mathbf{x}},Z=1)}{\mathbb{P}(X=\underline{\mathbf{x}},Z=1)}\\ &= \frac{\mathbb{P}(Z=1|Y=1,X=\underline{\mathbf{x}})\mathbb{P}(Y=1|X=\underline{\mathbf{x}})\mathbb{P}(X=\underline{\mathbf{x}})}{\mathbb{P}(Z=1|X=\underline{\mathbf{x}})\mathbb{P}(X=\underline{\mathbf{x}})}\\ &= \frac{\mathbb{P}(Z=1|Y=1,X=\underline{\mathbf{x}})\mathbb{P}(Y=1|X=\underline{\mathbf{x}})}{\mathbb{P}(Z=1|X=\underline{\mathbf{x}},Y=1)\mathbb{P}(Y=1|X=\underline{\mathbf{x}})+\mathbb{P}(Z=1|X=\underline{\mathbf{x}},Y=0)\mathbb{P}(Y=0|X=\underline{\mathbf{x}})}\\ &= (\star)\end{aligned}$$

Ponieważ Z nie zależy od X, to

$$\begin{aligned}(\star) &= \frac{\pi_1 h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}})}{\pi_1 h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}})+\pi_0 h(\underline{\boldsymbol{\beta}}x)}\\ &= \frac{h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}})}{h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}})+\pi_0/\pi_1 \bar{h}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}})}\end{aligned}$$

Oznaczmy $r := \pi_0/\pi_1$. Stąd

$$L_{\text{retro}}(\underline{\boldsymbol{\beta}},\pi_0,\pi_1,y_1,\underline{\mathbf{x}}_1,\ldots,y_n,\underline{\mathbf{x}}_n) = \prod_{i=1}^{n}\left(\frac{h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)}{h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)+r\bar{h}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)}\right)^{y_i}\cdot\left(\frac{r\bar{h}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)}{h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)+r\bar{h}(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i)}\right)^{1-y_i}.$$

W modelu logistycznym $h(x)=\frac{e^x}{1+e^x}$. Wtedy

$$\frac{h(a)}{h(a)+r(1-h(a))} = \frac{\frac{e^a}{1+e^a}}{\frac{e^a}{1+e^a}+r(1-\frac{e^a}{1+e^a})} = \frac{e^a}{e^a+r(1+e^a)-re^a} = \frac{e^a}{e^a+r} = \frac{\frac{1}{r}e^a}{1+\frac{1}{r}e^a} = \frac{e^{\tilde{a}}}{1+e^{\tilde{a}}},$$

gdzie $\tilde{a}=a+\log\frac{1}{r}$.

Zatem

$$L_{\text{retro}}() = \prod_{i=1}^{n} h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i-\log r)^{y_i}(1-h(\underline{\boldsymbol{\beta}}\underline{\mathbf{x}}_i-\log r))^{1-y_i}.$$

Wniosek: L_{prosp} różni się od L_{resp} tylko składnikiem β_0 przy założeniu modelu regresji logistycznej.

Rozdział 3

Zliczanie

3.1 Regresja Poissonowska

Regresja Poissonowska, tak jak omówiony wcześniej model dwumianowy, zlicza określone zdarzenia. Ponieważ opiera się ona na modelowaniu zmiennej odpowiedzi rozkładem Poissona, to ma sens o tyle, o ile spełnione są poniższe założenia:

(a) nie ma górnego ograniczenia na liczbę tego co zliczamy lub ograniczenie to jest bardzo duże (np. ograniczeniem na liczbę zachorowań jest liczba wszystkich osób w danej populacji, która jest bardzo duża);

(b) zliczane zdarzenia są (lub mogą być uznane za) niezależne.

Przypomnimy podstawowe fakty dotyczące rozkładu Poissona. Niech więc $Y \sim \text{Pois}(\lambda)$, gdzie $\lambda > 0$, tzn.

$$\mathbb{P}(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

Wówczas:

–> $\mathbb{E}Y = \lambda$,

–> $\mathbb{V}\text{ar}(Y) = \lambda$,

–> $\mathbb{E}e^{itY} = e^{-\lambda(1-e^{it})}$,

–> jeśli $X \sim \text{Pois}(\eta)$ oraz X i Y są niezależne, to $X + Y \sim \text{Pois}(\lambda + \eta)$.

Student/czytelnik każdą z tych własności powinien bez trudu udowodnić.

Wprowadźmy oznaczenie $P(y;\, \mu) := \mathbb{P}(Y = y)$, jeśli $Y \sim \text{Pois}(\mu)$.

Spróbuj sam!

Przekonaj się, że

$$P(y;\mu) = c(y)\,\exp\left\{y\,\theta - \kappa(\theta)\right\}, \qquad y \in \mathbb{N}_0,$$

gdzie $\theta = \ln(\mu)$, $\kappa(\theta) = e^{\theta} = \mu$.

Ile wynosi $c(y)$?

Model regresji poissonowskiej

Niech:

$(y_i, \underline{\mathbf{x}}_i)$ – n niezależnych obserwacji,

$\mu_i := \mathbb{E}(Y_i|x_i)$,

$Y_i|x_i \sim \text{Pois}(\mu_i)$ – składowa losowa modelu.

Składowa systemowa (czyli dalsze założenia modelu):

$$\mu_i = \exp(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}) \Leftrightarrow \log \mu_i = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}.$$

Zauważmy, że $\mu_i > 0$ dla dowolnej wartości $\underline{\mathbf{x}}_i \in \mathbb{R}^{p+1}$ dla każdego $i = 1, \ldots, n$.

Ważną zaletą regresji poissonowskiej jest łatwość interpretacji współczynników. Dzięki temu lepiej rozumiemy, co oznaczają konkretne wartości estymatorów. Spójrzmy:

$$\mu(\underline{\mathbf{x}}) = \exp(\underline{\mathbf{x}}^T \underline{\boldsymbol{\beta}}) = e^{\beta_0} \cdot e^{x_1\beta_1} \cdot \ldots \cdot e^{x_p\beta_p}$$

czyli β_j ma multiplikatywny wpływ na μ:

$$\frac{\mu(x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p)}{\mu(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_p)} = e^{\beta_j}$$

albo

$$\log \mu(x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p) - \log \mu(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_p) = \beta_j.$$

Widzimy, że zwiększenie wartości $j$tej zmiennej objaśniającej o jeden powoduje zmianę $\mu(\underline{\mathbf{x}})$ dokładnie e^{β_j} razy.

Inne możliwe funkcje łączące dla rozkładu Poissona:

- $h(x) = x^2$. Wtedy $\sqrt{\mu} = \underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}$. Jednak $\mu(\underline{\mathbf{x}}) = (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)^2$ jest trudne w interpretacji!
- $h(x) = |x|$, czyli $\mu(\underline{\mathbf{x}})) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p|$. Tutaj problemem jest nieróżniczkowalność w 0.

3.2 MLE dla regresji poissonowskiej

Ponieważ

$$P(y;\mu) = e^{-\mu}\frac{\mu^y}{y!} = \exp\{-\mu + y\log\mu - \log y!\},$$

to

$$l(\underline{\mathbf{y}},\mu) = l(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[y_i\log\mu_i - \mu_i - \log y_i!] = \sum_{i=1}^{n}\left[y_i\underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}} - \exp(\underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}}) - \log y_i!\right].$$

Dalej

$$s(\underline{\boldsymbol{\beta}}) = \frac{\partial l}{\partial\underline{\boldsymbol{\beta}}}(\underline{\boldsymbol{\beta}}) = X^T\underline{\mathbf{y}} - X^T\underline{\mu},$$

gdzie $\underline{\mu}^T = \underline{\mu}^T(\underline{\boldsymbol{\beta}}) = [\exp(\underline{\mathbf{x}}_1^T\underline{\boldsymbol{\beta}}), \exp(\underline{\mathbf{x}}_2^T\underline{\boldsymbol{\beta}}), \ldots, \exp(\underline{\mathbf{x}}_n^T\underline{\boldsymbol{\beta}})]$.

Szukamy $\hat{\underline{\boldsymbol{\beta}}}$ takiego, że:

$$s(\hat{\underline{\boldsymbol{\beta}}}) = 0 \qquad \Leftrightarrow \qquad \sum_{i=1}^{n}\underline{\mathbf{x}}_i y_i = \sum_{i=1}^{n}\underline{\mathbf{x}}_i\exp(\underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}}).$$

To jest nieliniowe równanie, więc nie ma rozwiązań analitycznych – konieczność stosowania optymalizacji numerycznej. Tak jak było w przypadku regresji logistycznej.

Macierz informacji obserwowanej wynosi:

$$\mathcal{I}(\underline{\boldsymbol{\beta}},\underline{\mathbf{y}}) = \mathbb{E}\left[-\frac{\partial^2 l(\underline{\boldsymbol{\beta}})}{\partial\underline{\boldsymbol{\beta}}\partial\underline{\boldsymbol{\beta}}^T}\right] = \sum_{i=1}^{n}\underline{\mathbf{x}}_i\underline{\mathbf{x}}_i^T\exp(\underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}}) = X^TWX,$$

gdzie $W = \mathrm{diag}(\exp(\underline{\mathbf{x}}_i^T\underline{\boldsymbol{\beta}})) = W(\underline{\boldsymbol{\beta}})$. Ponieważ, macierz informacji obserwowanej nie zależy od $\underline{\mathbf{y}}$, to macierz inf. Fishera $J(\underline{\boldsymbol{\beta}}) = I(\underline{\boldsymbol{\beta}},\underline{\mathbf{y}})$. Zatem algorytm Fisher Scoring jest tym samym, co algorytm Newtona-Raphsona, i:

$$\underline{\boldsymbol{\beta}}_{k+1} = \underline{\boldsymbol{\beta}}_k - \mathcal{I}^{-1}(\underline{\boldsymbol{\beta}}_k)\cdot s(\underline{\boldsymbol{\beta}}_k)$$

$$\begin{aligned}\underline{\boldsymbol{\beta}}_{k+1} &= \underline{\boldsymbol{\beta}}_k + (\mathbf{X}^TW(\underline{\boldsymbol{\beta}}_k)\mathbf{X})^{-1}\mathbf{X}^T(\underline{\mathbf{y}} - \hat{\underline{\mu}}(\underline{\boldsymbol{\beta}}_k)) \\ &= (\mathbf{X}^TW(\underline{\boldsymbol{\beta}}_k)\mathbf{X})^{-1}\mathbf{X}^TW(\underline{\boldsymbol{\beta}}_k)\left[\mathbf{X}\underline{\boldsymbol{\beta}}_k + W^{-1}(\underline{\boldsymbol{\beta}}_k)(y - \hat{\underline{\mu}}(\underline{\boldsymbol{\beta}}_k))\right].\end{aligned}$$

3.3 Overdispersion – nadwyżka rozproszenia

Dla rozkładu $Y \sim \mathrm{Pois}(\mu)$ wariancja $\mathbb{V}\mathrm{ar}(Y) = \mu$. Jednak w praktyce wariancja (warunkowa pod warunkiem obserwacji $\underline{\mathbf{x}}$) często przekracza μ (a czasem jest poniżej wartości oczekiwanej). To zjawisko nazywamy **nadwyżką rozproszenia** (*overdispersion*). Przeciwne zjawisko, czyli rozproszenie mniejsze niż wynikałoby to z modelu, to **niedobór rozproszenia** (*underdispersion*).

Przyczyny nadwyżki rozproszenia mogą być różne i czasem są trudne do wykrycia. Możliwe przyczyny to m.in.:

- niejednorodność danych, która nie jest uchwycona przez dany model,
- dodatnia korelacja zmiennych.

Uwaga!

Nadwyżka/niedobór rozproszenia wpływają na estymatory błędów standardowych!

Co można zrobić?

1. Modele kwaziwiarogodnościowe (*quasi-likelihood, quasi-log-likelihood*)

   W modelu regresji poissonowskiej $\mathbb{V}\mathrm{ar}(y) = \mu$.

   Teraz tak nie jest: $\mathbb{E}(y_i|x_i) = \mu = h(\underline{\mathbf{x}}_i^T \underline{\boldsymbol{\beta}}_i)$, ale $\mathbb{V}\mathrm{ar}(y_i|x_i) = \varphi V(\mu_i)$ (w reg. Poiss. $V(\mu) = \mu$ i $\varphi = 1$). Gdy $\varphi > 1$ – overdispersion, gdy $\varphi < 1$ – underdispersion (niedobór rozproszenia).

   Zauważmy, że parametr φ i funkcja V są wspólne dla wszystkich $i = 1, \ldots, n$. Nie zakładamy jednak postaci rozkładu $Y|x$, a jedynie związek między średnią a wariancją warunkową.

   Z ogólnej teorii GLMów wynika, że równania estymacyjne się nie zmieniają (φ się skraca), więc estymatory parametrów też się nie zmieniają. Natomiast

   $$\mathbb{C}\mathrm{ov}(\hat{\underline{\boldsymbol{\beta}}}) = \varphi\, F^{-1}(\underline{\boldsymbol{\beta}}).$$

   (Czyli błędy rosną, gdy φ rośnie.)

   Estymacja parametru rozproszenia $\boldsymbol{\varphi}$. Gdy $\varphi = 1$, to $\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_I}$ i jest statystyką χ^2.

   W przypadku nadwyżki/niedoboru rozproszenia i postulatu $\mathbb{V}\mathrm{ar}Y = \varphi\mu_i$, to powyższe χ^2 nie jest statystyką χ^2, ale χ^2/φ jest sumą kwadratów n wystandaryzowanych składników, więc (przy spełnieniu odpowiednich założeń modelu) asymptotycznie

   $$\frac{\chi^2}{\varphi} \sim \chi^2_{(n-p)}.$$

   W szczególności (ponieważ jeśli $Z \sim \chi^2(K)$, to $\mathbb{E}Z = K$):

   $$\mathbb{E}\left(\frac{\chi^2}{\varphi}\right) \approx n - p \quad \Rightarrow \varphi \approx \mathbb{E}\left(\frac{\chi^2}{n-p}\right)$$

   Zatem estymator momentowy φ ma postać

   $$\hat{\varphi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

   Postępowanie jest następujące:

   - dopasowujemy zwykły model poissonowski i estymujemy $\hat{\mu}$,

- estymujemy $\hat{\varphi}$,
- w celu uzyskania poprawnej macierzy kowariancji $\mathbb{C}\text{ov}(\hat{\underline{\beta}})$ mnożymy macierz uzyskaną z MLE przez $\hat{\varphi}$,
- błędy standardowe mnożymy przez $\sqrt{\hat{\varphi}}$.

2. Model ujemy dwumianowy (Gamma–Poisson model)

Ogólna uwaga:

Niech μ - zmienna losowa i niech $\theta := \mathbb{E}(\mu)$. Jeśli $Y|\mu \sim \text{Pois}(\mu)$, to $\mathbb{E}Y = \mathbb{E}\left[\mathbb{E}(Y|\mu)\right] = \mathbb{E}\mu = \theta$ oraz $\mathbb{V}\text{ar}(Y) = \mathbb{E}\left[\mathbb{V}\text{ar}(Y|\mu)\right] + \mathbb{V}\text{ar}\left[\mathbb{E}(Y|\mu)\right] = \mathbb{E}\mu + \mathbb{V}\text{ar}\mu = \theta + \mathbb{V}\text{ar}\mu > \theta$.

Rozkład Gamma

$Z \sim G(a,b),\ a,b>0$:
$f(z) = \frac{b^a}{\Gamma(a)} z^{a-1}e^{-bz}I_{(0,\infty)}(z)$,
$\mathbb{E}(Z) = \frac{a}{b}$,
$\mathbb{V}\text{ar}(Z) = \frac{1}{b}\mathbb{E}(Z) = \frac{a}{b^2}$.

Rozkład ujemny dwumianowy

$Z \sim \text{NB}(r,p),\ r>0, p\in(0,1)$:
$\mathbb{P}(Z=k) = \frac{\Gamma(r+k)}{\Gamma(r)k!}p^k(1-p)^r$,
dla $k = 0,1,2,\ldots$
$\mathbb{E}(Z) = \frac{rp}{1-p}$,
$\mathbb{V}\text{ar}(Z) = \frac{r}{1-p}\mathbb{E}(Z) = \frac{rp}{(1-p)^2}$.

Załóżmy, że $Y|\lambda \sim \text{Pois}(\lambda)$ oraz $\lambda \sim \Gamma\left(\nu, \frac{\nu}{\mu}\right)$. Wówczas

$$\begin{aligned}
\mathbb{P}(Y=y) &= \int_0^\infty \mathbb{P}(Y=y|\lambda=u) f_\lambda(u)du \\
&= \int_0^\infty e^{-u}\frac{u^k}{k!}u^{\nu-1}e^{-\frac{\nu}{\mu}u}\left(\frac{\nu}{\mu}\right)^\nu \frac{1}{\Gamma(\nu)}du \\
&= \left(\frac{\nu}{\mu}\right)^\nu \frac{1}{\Gamma(\nu)y!}\int_0^\infty e^{-\left(\frac{\nu}{\mu}+1\right)u}u^{\nu+y-1}du \\
&= \left(\frac{\nu}{\mu}\right)^\nu \frac{1}{\Gamma(\nu)y!}\frac{\Gamma(\nu+y)}{\left(\frac{\nu}{\mu}+1\right)^{\nu+y}} \\
&= \left(\frac{\nu}{\nu+\mu}\right)^\nu \left(\frac{\mu}{\nu+\mu}\right)^y \frac{\Gamma(\nu+y)}{\Gamma(\nu)y!}
\end{aligned}$$

dla $y \in \{0,1,2,\ldots\}$.

A zatem Y ma rozkład ujemny dwumianowy $\text{NB}\left(\nu, \frac{\mu}{\nu+\mu}\right)$ i $\mathbb{V}\text{ar}(Y) = \mu + \frac{\mu^2}{\nu} > \mu$, a więc model ten tłumaczy poissonowską nadwyżkę rozproszenia.

Zauważmy, że gdy ν jest małe, to nadwyżka rozproszenia jest duża, a gdy $\nu \to \infty$, to nadwyżka zbiega do 0, a więc $\mathbb{V}\text{ar}(Y) \to \mu$. Parametr $1/\nu$ zwany jest parametrem **rozproszenia**.

Uwaga

Jeśli ν jest ustalone, to rozkład ujemny dwumianowy należy do rodziny rozkładów wykładniczych.

Rozdział 4

Rodziny wykładnicze rozkładów prawdopodobieństwa

W GLM zakładamy, że zmienna odpowiedzi Y pochodzi z rozkładu, który należy do rodziny wykładniczej. Rodzina wykładnicza, to rodzina rozkładów, które mają gęstość specjalnej postaci. Dzięki tej specjalnej postaci spełniają własności i można dla nich wyprowadzać pewne ogólne wzory na różne rzeczy, np. związek średniej z wariancją, postać estymatora, dewiancję itp.

Definicja 4.1. *Naturalna rodzina wykładnicza (ang. natural exponential family NEF) to rodzina rozkładów p-stwa o gęstości (lub funkcji prawdopodobieństwa) postaci:*

$$f(y;\theta) = a(y)\exp\left\{\theta y - \kappa(\theta)\right\}, \quad \theta \in \Omega,$$

gdzie

- κ – *funkcja kumulanty (cumulant function),*
- θ – *parametr kanoniczny (canonical parameter),*
- Ω – *przestrzeń parametrów kanonicznych.*

Rodzina wykładnicza z parametrem rozproszenia

(ang. *exponential dispersion model family* EDM) to rodzina rozkładów p-stwa o gęstości (lub funkcji prawdopodobieństwa) postaci:

$$f(y;\theta,\varphi) = a(y,\varphi)\exp\left\{\frac{\theta y - \kappa(\theta)}{\varphi}\right\}, \quad \theta \in \Omega,$$

gdzie

- κ – funkcja kumulanty (*cumulant function*),
- θ – parametr kanoniczny (*canonical parameter*),
- Ω – przestrzeń parametrów kanonicznych,
- φ – parametr rozproszenia (*dispersion*).

Przykłady rozkładów z rodzin wykładniczych

1. rozkład Poissona: $Y \sim \text{Pois}(\mu)$, $|mu > 0$

   $$f(y;\mu) = \frac{\mu^y \exp(-\mu)}{y!} = \frac{1}{y!}\exp\{y\log\mu - \mu\}, \quad y \in \mathbb{N}_0$$

   więc $\theta = \log\mu \quad \Leftrightarrow \quad \mu = e^\theta$, $\kappa(\theta) = \mu$, $a(y) = \frac{1}{y!}$ – to jest NEF

   Kanoniczna przestrzeń parametrów: $\Omega = \mathbb{R}$ (tam gdzie jest zbieżna całka z gęstości, czyli gdzie gęstość jest gęstością).

2. rozkład normalny: $Z \sim \mathbb{N}(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$

   $$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}\exp\left\{\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2}\right\}, \quad y \in \mathbb{R}$$

   więc $\varphi = \sigma^2$, $\theta = \mu$, $\kappa(\theta) = \frac{\theta^2}{2}$, $a(y,\varphi) = \frac{1}{\sqrt{2\pi\varphi}}e^{-\frac{y^2}{2\varphi}}$.

3. rozkład Bernoulliego i rozkład dwumianowy: $Y \sim \text{b}(n,p)$, $n \in \mathbb{N}$, $p \in [0,1]$.

   $$f(z;n,p) = \binom{n}{z}p^z(1-p)^{n-z} = \binom{n}{z}\exp\left\{z\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}, \quad z \in \{0,1,\ldots,n\}.$$

   Jeśli $n = 1$ (czyli Bernoulli), to jest to element z NEF, wtedy $\theta = \log\left(\frac{p}{1-p}\right)$, $\kappa(\theta) = \log(1+\exp(\theta))$ (niech studenci sprawdzą sobie).

   Jeśli $n > 1$, to niech $Y := \frac{Z}{n}$.

   $$\mathbb{P}(Y = y) = \mathbb{P}(Z = ny) = \binom{n}{ny}\exp\left\{n\left(y\log(\frac{p}{1-p} + \log(1-p)\right)\right\},$$

   więc $\theta = \log\left(\frac{p}{1-p}\right)$, $\kappa(\theta) = \log(1+\exp(\theta))$, $\varphi = \frac{1}{n}$, $a(y,\varphi) = \binom{1/\varphi}{y/n}$.

4.1 Własności rozkładów z rodzin wykładniczych

Stwierdzenie 4.1. *Niech*

$$Y \sim f_Y(y, \theta, \varphi) = a(y, \varphi) \cdot \exp\left\{\frac{y\theta - \kappa(\theta)}{\varphi}\right\}.$$

Wtedy $\mathbb{E}(Y) = \kappa'(\theta) = \frac{d\kappa(\theta)}{d\theta}$, $\mathbb{V}\text{ar}Y = \varphi\kappa''(\theta) = \varphi\mu'(\theta)$.

Dowód. Mamy równość

$$1 = \int f_Y(y, \theta, \varphi)dy.$$

Różniczkujemy obustronnie po θ:

$$0 = \int \frac{d}{d\theta} f(y, \theta, \varphi)dy = \int c(y, \varphi) \exp\left\{\frac{1}{\varphi}(\theta y - \kappa(\theta))\right\} \cdot \frac{1}{\varphi}(y - \kappa'(\theta))dy = \frac{1}{\varphi}\mathbb{E}Y - \frac{\kappa'(\theta)}{\varphi}$$

Różniczkując jeszcze raz po θ:

$$0 = \int c(y, \varphi) \exp\left\{\frac{1}{\varphi}(\theta y - \kappa(\theta))\right\} \cdot \left[\frac{1}{\varphi^2}(y - \kappa'(\theta))^2 - \frac{1}{\varphi}\kappa''(\theta)\right] dy = \frac{1}{\varphi^2}\mathbb{V}\text{ar}(Y) - \frac{1}{\varphi}\kappa''(\theta).$$

▼

Niech $\tau(\theta) = \kappa'(\theta) = \mu = \mathbb{E}(Y)$. Wtedy $\tau'(\theta) > 0$ dla dowolnego $\theta \in \Omega$ (bo wariancja jest dodatnia), więc τ jest funkcją rosnącą i różnowartościową. Zatem definiuje jednoznaczne przekształcenie: $\tau : \Omega \to \mathcal{M} \subset \mathbb{R}$, gdzie $\mathcal{M}$ jest przestrzenią wartości oczekiwanej (*mean value space*).

Wariancję można więc wyrazić jako funkcję wartości oczekiwanej

Niech

$$V(\mu) := \frac{d\mu(\theta)}{d\theta}$$

będzie funkcją wariancji (*variance function*). Wtedy

$$\mathbb{V}\text{ar}(Y) = \varphi V(\mu).$$

Ważne: funkcja wariancji $V(\mu)$ jednoznacznie wyznacza rozkład (z dokładnością do parametrów) wewnątrz rodziny wykładniczej, ponieważ wyznacza $\kappa(\theta)$ z dokładnością do stałej addytywnej

Przykład 4.1. *$V(\mu) = \mu^2$, $\mu > 0$. Ponieważ $V(\mu) = d\mu/d\theta$, to $d\theta/d\mu = \mu^{-2}$, więc $\theta = -1/\mu$ (stała całkowania $= 0$). Dalej ponieważ $\mu = d\kappa(\theta)/d\theta$, to $\kappa(\theta) = -\log(-\theta) = \log\mu$. Zatem*

$$P(y) = a(y, \varphi) \exp\left\{\frac{y(-1/\mu) - \log\mu}{\varphi}\right\}.$$

I otrzymaliśmy rozkład gamma (z dowolnością parametrów).

Szczególnym przypadkiem jest $V(\mu) = 1$, tzn. nie ma zależności średniej i wariancji. Wtedy rozkład jest normalny.

Mamy zatem dwie równoważne reprezentacje rozkładów:

- przez funkcję kumulantową κ z parametryzacją przez parametr kanoniczny θ i parametr rozproszenia φ.
- przez funkcję wariancji V, która specyfikuje wariancję jako funkcję wartości oczekiwanej $\mu \in M$ i przez parametr φ.

Spróbuj sam!

Niech $V(\mu) = \mu^2$. Jaki to rozkład?

Rozwiązanie: $\mu'(\theta) = \mu(\theta) \Rightarrow \mu(\theta) = a\exp(\theta)$. Stąd $\theta = \log(\mu)$. Z kolei $\mu(\theta) = \kappa'(\theta)$, a więc $\kappa(\theta) = \exp(\theta)$. Jest to rozkład Poissona.

Rozdział 5

Uogólnione modele liniowe

Pionierami wprowadzającymi GLM byli John A. Nelder i P. McCullagh w 1983.

5.1 Funkcje łączące

Tak jak do tej pory, niech $(Y_i)_{i=1,\dots,N}$ będą zmiennymi objaśnianymi o rozkładzie z rodziny wykładniczej, $\mathbf{X}_i = (1, \mathbf{X}_{i1}, \dots, \mathbf{X}_{ip})$ dla $i = 1, \dots, N$ – zmiennymi objaśniającymi.

Definicja 5.1. *Funkcją łączącą w modelu GLM ($\mathbb{E}Y = \mu$) nazywamy funkcję g taką, że*

$$g(\mu) = \mathbf{X}^T \underline{\boldsymbol{\beta}} =: \eta.$$

Zakładamy, że g jest monotoniczna i różniczkowalna (na interesującej nas dziedzinie, np. $(0, \infty)$).

Parametr $\eta = \mathbf{X}^T \beta$ nazywamy *predykatorem liniowym.*

Uwaga 5.1. *Tylko parametr θ (a nie φ) związany jest ze zmiennymi objaśniającymi:*

$$g(\mu) = g(\kappa'(\theta)) = \mathbf{X}^T \underline{\boldsymbol{\beta}}.$$

Uwaga 5.2. *Możliwe są różne funkcje łączące.*

Definicja 5.2. Kanoniczna funkcja łącząca *(canonical link function) to taka funkcja łącząca g, że $g(\mu) = \theta$.*

Zauważmy, że $\mu = \kappa'(\theta)$, a zatem w przypadku kanonicznej funkcji łączącej $g^{-1} = \kappa'$. Wtedy też $\theta = \eta = g(\mu) = \mathbf{X}^T \underline{\boldsymbol{\beta}}$, więc predykator liniowy η modeluje właśnie parametr kanoniczny θ. Warto używać kanonicznej funkcji łączącej również ze względu na dobre własności matematyczne i uproszczenia we wzorach.

Przykłady kanonicznych funkcji łączących

1. $Y \sim \mathbb{N}(\mu, \sigma^2)$: $\theta = \mu$, więc $g = id$;
2. $Y \sim \text{Pois}(\mu)$: $\theta = \log \mu$, więc $g(\mu) = \log(\mu)$;
3. $Z \sim b(n,p)$, $Y = \frac{Z}{n}$: $\theta = \log\left(\frac{p}{1-p}\right)$, $\mu = \mathbb{E}Y = p$: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \text{logit}(\mu)$.

5.2 Estymacja największej wiarogodności dla rodzin wykładniczych

Niech $y_1, \ldots, y_n$ – niezależne obserwacje z rodziny wykładniczej. Wtedy

$$l(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{y\theta - \kappa(\theta)}{\varphi}$$

(pozwalamy sobie pominąć wyrazy, które nie zależą od $\underline{\boldsymbol{\beta}}$).

$$s(\underline{\boldsymbol{\beta}}) = \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}}}.$$

Odwrotność funkcji łączącej g oznaczymy przez h.

Stwierdzenie 5.1. *Score function (funkcja oceny) dla obserwacji $y_1, \ldots, y_n$ z rodziny wykładniczej ma postać*

(a) $s(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}\text{ar}(Y_i)} \underline{\mathbf{x}}_i \frac{\partial h(\eta_i)}{\partial \eta}$

(b) W przypadku kanonicznej funkcji łączącej

$$s(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \underline{\mathbf{x}}_i \frac{y_i - \mu_i}{\varphi_i}.$$

Dowód. Dla rozkładu z rodziny wykładniczej

$$l(\underline{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - \kappa(\theta_i)}{\varphi_i} \right).$$

Zatem

$$s(\underline{\boldsymbol{\beta}}) = \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \frac{\partial l_i(\underline{\boldsymbol{\beta}})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu} \frac{\partial h(\eta_i)}{\partial \eta} \frac{\partial \eta_i}{\partial \underline{\boldsymbol{\beta}}}.$$

Policzymy każdą z pochodnych cząstkowych pojawiających się w powyższym wzorze.

1. $\frac{\partial l_i(\underline{\boldsymbol{\beta}})}{\partial \theta} = \frac{y_i - \kappa'(\theta_i)}{\varphi_i} = \frac{y_i - \mu_i}{\varphi_i}$;
2. $\frac{\partial \theta_i(\mu)}{\partial \mu} = \left(\frac{\partial \mu(\theta_i)}{\partial \theta}\right)^{-1} = \left(\frac{\partial \kappa'(\theta_i)}{\partial \theta}\right)^{-1} = \frac{1}{\kappa''(\theta_i)} \frac{\varphi_i}{\varphi_i} = \frac{\varphi}{\mathbb{V}\text{ar}(Y_i)}$;

3. $\dfrac{\partial \eta_i(\underline{\boldsymbol{\beta}})}{\partial \underline{\boldsymbol{\beta}}} = \dfrac{\partial(\underline{\boldsymbol{\beta}}'\underline{\mathbf{x}}_i)}{\partial \underline{\boldsymbol{\beta}}} = \underline{\mathbf{x}}_i.$

Po wstawieniu do wzoru na $s(\underline{\boldsymbol{\beta}})$ otrzymujemy tezę.

(b)

Spróbuj sam!

Dla kanonicznej funkcji łączącej

$$\frac{\partial h(\eta)}{\partial \eta} = \frac{\mathbb{V}\mathrm{ar}(Y_i)}{\varphi_i} = \kappa''(\theta_i).$$

Notacja macierzowa. Wyprowadzone wzory warto zapisać za pomocą macierzy. Daje to nie tylko bardziej zwięzły zapis, ale też możliwość prostszej implementacji. Ponadto, pewne własności stają się bardziej widoczne.

W świetle ostatniego stwierdzenia możemy napisać, że

$$s(\underline{\boldsymbol{\beta}}) = \mathbf{X}' D \Sigma^{-1}(\underline{\mathbf{y}} - \mu),$$

gdzie $\mathbf{X}$ jest macierzą eksperymentu, $\Sigma = \mathrm{diag}(\mathbb{V}\mathrm{ar}(Y_1), \ldots, \mathbb{V}\mathrm{ar}(Y_n))$, $\mathcal{D} = \mathrm{diag}\left(\frac{\partial h(\eta_1)}{\partial \eta}, \ldots, \frac{\partial h(\eta_n)}{\partial \eta}\right)$.

Zdefiniujmy $W := D\Sigma^{-1}D^{-1}$. Wtedy

$$s(\underline{\boldsymbol{\beta}}) = \mathbf{X}' W D(\underline{\mathbf{y}} - \mu).$$

A jeśli h jest kanoniczną funkcją łączącą, to

$$s(\underline{\boldsymbol{\beta}}) = \mathbf{X}' \tilde{\Sigma}^{-1}(\underline{\mathbf{y}} - \mu),$$

gdzie $\tilde{\Sigma} = \mathrm{diag}(\varphi_1, \ldots, \varphi_n)$.

Znajdziemy postać macierzy informacji obserwowanej i macierzy informacji Fishera dla rodziny wykładniczej. Skorzystamy przy tym z faktu, że dla rodziny wykładniczej

$$\frac{\partial^2 l(\underline{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_k} = \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \beta_j} \frac{\partial l(\underline{\boldsymbol{\beta}})}{\partial \beta_k}.$$

Stwierdzenie 5.2. *Dla rodziny wykładniczej macierz informacji Fishera*

$$I(\underline{\boldsymbol{\beta}}) = -\mathbf{X}' W \mathbf{X},$$

gdzie $W = D\Sigma^{-1}D$, $D = \mathrm{diag}\left(\frac{\partial h(\eta_1)}{\partial \eta_1}, \ldots, \frac{\partial h(\eta_n)}{\partial \eta_n}\right)$, $\Sigma = \mathrm{diag}(\mathbb{V}\mathrm{ar}Y_1, \ldots, \mathbb{V}\mathrm{ar}Y_n)$.

Jeśli h jest kanoniczną funkcją łączącą, to

$$I(\underline{\boldsymbol{\beta}}) = \mathbf{X}^T W \mathbf{X},$$

gdzie $W = \mathrm{diag}\left(\frac{\mathbb{V}\mathrm{ar}Y_1}{\varphi_1^2}, \ldots, \frac{\mathbb{V}\mathrm{ar}Y_n}{\varphi_n^2}\right)$.

Dowód. Dowód sprowadza się do rachunków. ▼

Stwierdzenie 5.3. *Niech Y_j, $j = 1, \ldots, n$, będą zmiennymi losowymi takimi, że $Y_j \sim \mathrm{EDF}(\theta, \varphi, \kappa(\theta))$. Wówczas*

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \sim \mathrm{EDF}\left(\theta, \frac{\varphi}{n}, \kappa(\theta)\right).$$

Czyli zmienna losowa $\frac{1}{n}\sum_{j=1}^{n} Y_j$ ma gęstość postaci

$$f(y) = \tilde{c}\left(y, \frac{\varphi}{n}\right) \exp\left\{\frac{y\theta - \kappa(\theta)}{\varphi/n}\right\}.$$

Spróbuj dowieść sam!

Należy policzyć funkcję charakterystyczną Y_j, następnie funkcję charakterystyczną $\frac{1}{n}\sum_{j=1}^{n} Y_j$ i zorientować się, że to koniec.

Powyższa własność gwarantuje sensowność *grupowania* danych w GLM. Widzimy, że jeśli dane indywidualne zgrupujemy, to pozostaniemy w tej samej rodzinie EDF ze zmienionym parametrem rozproszenia. Odpowiada to intuicji statystyka.

Rozdział 6

Modele graficzne

Over the last decades, the collection and rapid diffusion of network data originating from a wide spectrum of scientific areas have created the need for new statistical theories and methodologies for modeling and analyzing large random graphs.
Lauritzen, Rinaldo, Sadeghi [Random Networks, Graphical Models, and Exchangeability, 2017]

Modelowanie zachowania dużych systemów zmiennych to aktualnie wielkie wyzwanie. Zmienne w systemach zależą od siebie i wpływają na siebie. Odnalezienie takich struktur zależności i niezależności jest bardzo cenne i pozwala znacznie (!) zmniejszyć złożoność obliczeniową i pamięciową sieci.

Modele graficznie między innymi temu służą, aby dać strukturę i narzędzia odpowiednie do modelowania sieci z wewnętrznym systemem warunkowych niezależności.

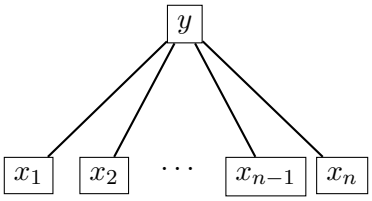
Przykład: warto zauważać warunkową niezależność!

Przypuśćmy, że mamy pewien model $p_\theta(y; x_1, \ldots, x_n)$ pojawiania się konkretnych słów w mailach spamowych. Wartość y mówi, czy dany mail jest spamem, a $x_1, \ldots, x_n$ mówią czy $i$te słowo (z ustalonego słownika) pojawiło się w tym mailu. Zatem mamy tu do czynienia z bardzo prostą sytuacją, ale... Wszystkich możliwych wartościowań y, x_1, $\ldots$, x_n jest 2^{n+1}. Jeśli n jest duże, a przecież jest!, to jest tego ZA dużo (i rośnie wykładniczo wraz z n). Jest to problem, który można rozwiązać zauważając lub zakładając warunkowe niezależności. Jeśli założyć, że występowanie słów pod warunkiem że mail jest spamem, jest niezależne, to

$$P(y; x_1, \ldots, x_n) = p(y) \prod_{i=1}^{n} p(x_i|y).$$

Każdy z parametrów $p(x_i|y)$ jest opisywany przez, tylko!, 4 czynniki. Zatem, suma sumarum, mamy $4(n+1)$ czynników, czyli $O(n)$.

Warunkowa niezależność pozwala się łatwo wyreprezentować na grafie. Ma przy tym tę zaletę, że jest intuicyjnie rozumiana przez szerokie grono odbiorców.



Ten graf opisuje sytuację przedstawioną powyżej. Brak krawędzi między wierzchołkami oznacza warunkową niezależność pod warunkiem (w tym przypadku) y. Możemy dopowiedzieć sobie: najpierw wybieramy czy dany mail jest spamem czy nie (y), a potem generujemy niezależnie słowa z rozkładu dla spamu/niespamu.

Inference: Modele graficzne mają też udostępniać narzędzia do odpowiedzi na zapytania (np. jakie jest prawdopodobieństwo, że mail jest spamem, jeśli widzę słowo "pigułka"). Typowe *inqueries* (zapytania):

- *marginal inference*: Jakie są prawdopodobieństwa brzegowe w modelu?

$$p(x_1) = \sum_{x_2} \sum_{x_3} \cdots \sum_{x_n} p(x_1, x_2, \ldots, x_n)$$

- *Maximum a posterioi inference*: jaka jest najbardziej prawdopodobna konfiguracja x-sów. Np.:

$$\max_{x_1,\ldots,x_n} p(x_1, \ldots, x_n, y = 1).$$

Choć powyższe zadania sa matematycznie jasne, to często okazują się NP-trudne.

Uczenie: Nasze ostatnie kluczowe zadanie odnosi się do dopasowania modelu do zbioru danych (którym może być np. duża liczba oznakowanych przykładów spamu). Przeglądając dane, możemy wnioskować o występujących w nich wzorcach (np. które słowa częściej znajdują się w spamie). Następnie, możemy je wykorzystać do przewidywania przyszłości (*inference*). Zobaczymy jednak, że uczenie się i wnioskowanie są również z natury rzeczy powiązane w bardziej subtelny sposób, ponieważ wnioskowanie okaże się kluczową podprogramą, którą będziemy wielokrotnie wywoływać w ramach algorytmów uczenia się. Temat uczenia się będzie również charakteryzował się ważnymi powiązaniami z teorią uczenia się obliczeniowego – która zajmuje się takimi zagadnieniami jak uogólnianie na podstawie ograniczonych danych i overfitting – jak również ze statystyką bayesowską – która mówi nam (między innymi) o tym, jak połączyć wcześniejszą wiedzę i zaobserwowane dowody w sposób pryncypialny.

6.1 Modele log–liniowe

Modele *log–liniowe* to modele dla wielowymiarowych dyskretnych danych zliczających ile było przypadków o poszczególnych cechach. Dane te są często reprezentowane w postaci **tablic kontyngencji** zwanych też **tabelami krzyżowymi**.

Przykład: zwyczaje wron i gawronów

Gatunek	liczba jaj	wielkość gniazda		
		Małe	średnie	Duże
Wrona	1	32	2	3
	2	12	20	12
	3	10	10	10
Gawron	1	7	20	3
	2	1	20	12
	3	2	10	50

6.1.1 Notacja

W dalszej części N będzie zwykle oznaczało liczbę obserwacji w zbiorze danych. Liczbę zmiennych (objaśniających, dyskretnych) – d. W przykładzie powyżej $N = 32+12+20+10+10+10+7+1+20+2+10+50 = 184$, a $d = 3$. Cały ten wektor zmiennych losowych zapisujemy: $\mathbf{X} = (X_v)_{v\in\Delta}$. *Poziomy* to możliwe wartości, które zmienna losowa może przyjąć (z angielskiego *levels*). Liczbę poziomów dla zmiennej X_v oznaczamy $|X_v|$. Przyjmujemy, że poziomy możemy oznaczyć: $1, 2, \ldots, |X_v|$ dla każdego v (choć w praktyce one coś znaczą). Wartość przyjmowana przez $\mathbf{X}$ ozn. $i = (i_1, \ldots, i_d)$ (zwana jest też komórką (ang. *cell*). Symbol $\mathcal{I}$ oznacza zbiór wszystkich możliwych do przyjęcia komórek. Zakładamy, podobnie jak w regresji Poissona i w modelu wielomianowych, że obserwacje są niezależne. Chcemy modelować prawdopodobieństwa $p(i) = \mathbb{P}(\mathbf{X} = i)$, $i \in \mathcal{I}$. Zatem łączne prawdopodobieństwo zastanych wartości $i^v,\ v = 1, \ldots, N$, to

$$p\left(i^v,\ v = 1, \ldots, N\right) = \prod_{v=1}^{N} p\left(i^v\right) = \prod_{i\in\mathcal{I}} p(i)^{n(i)}, \tag{6.1}$$

gdzie $n(i)$ jest liczbą wystąpień komórki i w całej próbie. (Więc jest też wartością w odpowiednim miejscu tabeli krzyżowej/kontyngencji.) Z drugiej strony

$$p\left(\{n(i)\}_{i\in\mathcal{I}}\right) = \frac{N!}{\prod_{i\in\mathcal{I}} \prod_{i\in\mathcal{I}} p(i)^{n(i)}}.$$

Należy upewnić się, że rozumie się dlaczego tak jest. To wyrażenie różni się od (6.1) tylko o stały współczynnik iloczynowy, który nie wpływa na zachowanie funkcji wiarogodności (a dokładnie na położenie jej ekstremów):

$$L(p) \propto \prod_{i \in \mathcal{I}} p(i)^{n(i)}.$$

Zrób to sam!

Pokazać (standardową metodą), że estymatorem największej wiarogodności p w opisanej wyżej sytuacji jest

$$\hat{p}(i) = \frac{n(i)}{N}, \quad i \in \mathcal{I}.$$

Pozyższy estymator jes ENW w sytuacji, w której nie ma żadnych dodatkowych restrykcji na rozkład p. Ten model nazywany jest **modelem wysyconym** (saturated).

Jakie to mogą być restrykcje? Mówimy o narzuceniu warunkowych niezależności poszczególnych zmiennych. Czyni się to, aby zaoszczędzić na obliczeniach, przestrzeni, strukturze.

Dalsze oznaczenia:

- $m(i)$ – wartość oczekiwana liczności danej komórki, więc $m(i) = np(i)$;
- $\hat{m}(i) := N\hat{p}(i)$ – wartości dopasowane $m(i)$;
- $i_A := (i_v)_{v \in A}$ – komórka marginalna zdefiniowana dla dowolnego $A \subset \Delta$.
- $n(i_A)$ i $p(i_A)$ - odpowiadające liczności i prawdopodobieństwa, czyli

$$p(j_A) = \sum_{\substack{i \in \mathcal{I}: \\ i_A = j_A}} p(i) \quad \text{i} \quad n(j_A) = \sum_{\substack{i \in \mathcal{I}: \\ i_A = j_A}} n(i)$$

6.1.2 Model log–liniowy hierarchiczny

Ogólna postać modelu log–liniowego:

$$\log p(i) = u + \sum_{A \subset \Delta} u^A(i_A).$$

Niektóre współczynniki u^A mogą być równe zer – to są ograniczenia związane z warunkową niezależnością. Przykładowo dla danych dotyczących wron i gawronów mielibyśmy

$$\log p(i) = u + u^1(j) + u^2(k) + u^3(l) + u^{1\,2}(j,k) + u^{2\,3}(k,l) + u^{1\,3}(j,l) + u^{1\,2\,3}(j,k,l).$$

Parametry u zwane są *interaction terms*. Jeśli założymy, że nie ma efektu wspólnego dla liczby jak i gatunku, który by wpływał na liczbę takich obserwacji, to $u^{1\,2} \equiv 0$ i model upraszcza się.

Definicja 6.1. *Model jest **hierarchiczny**, jeśli prawdziwa jest implikacja*

$$u^A \equiv 0 \implies \left(\forall_{B\subset\Delta} : A \subset B \implies u^B \equiv 0\right).$$

Jeśli model jest hierarchiczny, to możemy go scharakteryzować poprzez podanie 'najwyższych' niezerowych współczynników u. Taki zbiór nazywamy zbiorem generatorów lub generującym.

Przykład

Jeśli $\Delta = \{1, 2, 3\}$, i model jest hierarchiczny, to model o generatorach $\{a, b\}$ $\{b, c\}$ jest postaci

$$\log p(i) = u + u^1(j) + u^2(k) + u^3(l) + u^{1\,2}(j, k) + u^{2\,3}(k, l).$$

Uwaga 6.1 (Faktoryzacja funkcji prawdopodobieństwa)**.** *Niech* $\mathbf{X} = (X_1, \ldots, X_N)$*. Wówczas*

$$X_A \perp X_B | X_C \iff p(x) = p_1\left(x_{A\cup C}\right) p_2\left(x_{B\cup C}\right),$$

gdzie p_1 *– pewna funkcja* $|A \cup C|$ *zmiennych,* p_2 *– pewna funkcja* $|B \cup C|$ *zmiennych oraz* $x_D = (x_j)_{j\in D}$ *dla dowolnego podzbioru* D *zbioru* $\{1, \ldots, N\}$*.*

(Jeśli ktoś tego nie wie, to powinien sobie udowodnić!)

Definicja 6.2. *Grafem zależności hierarchicznego modelu log–liniowego nazywamy nieskierowany graf* $G = (V, E)$*, gdzie* $V = \Delta$ *i* $\{u, v\} \in E$ *jeśli* $u_{uv} \neq 0$*. Czyli krawędź występuje wszędzie tam, gdzie dozwolona jest interakcja czynników.*

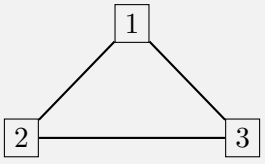
Definicja 6.3. *Hierarchiczny model log–liniowy jest modelem graficznym, jeśli istnieje graf* $G = (V, E)$ *taki, że* $V = \Delta$ *i zbiór klik* G *jest tożsamy ze zbiorem generatorów modelu.*

model niegraficzny

Najprostszym przykładem modelu, który nie jest graficzny jest

$$\log p(i) = u + u^1(j) + u^2(k) + u^3(l) + u^{1\,2}(j, k) + u^{2\,3}(k, l) + u^{1\,3}(j, l).$$

Łatwo zauważyć, że graf zależności tego modelu jest taki sam jak graf zależności modelu wysyconego (ze wszystkimi interakcjami, a więc dodatkowo z interakcją wszystkich trzech czynników) i wygląda następująco:



Zaletą tych modeli log–liniowych, które są graficznymi jest to, że mogą być całkowicie interpretowane (czy reprezentować) w terminach warunkowych niezależności, które z kolei możemy przedstawiać za pomocą grafu (zależności)

6.1.3 Estymacja i dopasowanie modelu

Definicja 6.4. *Model $\mathcal{M}$ jest dekomponowalny jeśli jego graf (zależności) jest chordalny (po angielsku jest też słowo triangulated, ale po polsku brak odpowiednika, bo 'trangulizowalny' brzmi dziwnie).*

Definicja 6.5. *Graf jest chordalny, jeśli każdy cykl długości co najmniej 4 ma przekątną.*

Dla modeli dekomponowalnych można podać analityczne formuły na estymatory największej wiarogodności. (Dlatego te modele są szczególnie ważne!)

Dane jest model graficzny: Δ, N, $n(i)$ dla $i \in \mathcal{I}$ oraz graf zależności $G = (V, E)$. Niech $\mathcal{C} = (C_1, \ldots, C_k)$ będzie sekwencją klik uporządkowaną zgodnie z **porządkiem topologicznym** wierzchołków. Niech $\mathcal{S} = (S_1, \ldots, S_k)$ będzie sekwencją odpowiadających separatorów: $S_k = C_k \cap C_{k-1}$, więc w szczególności $S_1 = \emptyset$. Wówczas estymator największej wiarogodności wartości oczekiwanej liczności komórki i jest równy

$$\hat{m}(i) = \frac{\prod_{j=1}^{k} n(i_{C_j})}{\prod_{j=1}^{k} n(i_{S_j})}, \quad i = (i_1, \ldots, i_d) \in \mathcal{I}.$$

Dalej $\hat{p}(i) = \hat{m}(i)/N$.

W przypadku modeli, które nie są dekomponowalne stosowane są metody iteracyjne, np. IPS – iterative proportional scaling.

6.1.4 Testowanie hipotez

Wcześniej (6.1.1) uzasadniliśmy, że $L(p) \propto \prod_{i \in \mathcal{I}} p(i)^{n(i)}$. Zatem maksymalna wartość log–wiarogodności, z dokładnością do stałej **addytywnej**, wynosi

$$l = \sum_{i \in \mathcal{I}} n(i) \log \hat{p}(i),$$

gdzie $\hat{p}(i)$ jest MLE.

Z definicji dewiancji ($D = 2(l(\hat{p}) - l(p_{\mathcal{M}})$, gdzie $p_{\mathcal{M}}$ jest estymatorem uzyskanym w modelu)

$$D = 2 \sum_{i \in \mathcal{I}} n(i) \log \frac{n(i)}{\hat{m}(i)}.$$

Jeśli model $\mathcal{M}$ jest prawdziwy, to D ma asymptotycznie rozkład $\chi^2(k)$, gdzie k jest różnicą w liczbie parametrów modelu wysyconego i $\mathcal{M}$.

Alternatywną statystyką testową hipotezy o dobrym dopasowaniu $\mathcal{M}$ jest **test dopasowania Pearsona**. Statystyka asymptotycznie ma rozkład χ^2:

$$\chi^2 = \sum_{i \in \mathcal{I}} \frac{(n(i) - \hat{m}(i))^2}{\hat{m}(i)} \overset{H_0}{\sim} \chi^2(k).$$

Więcej na temat dopasowywania i porównywania modeli znajduje się w części praktycznej (wraz z przykładami).

6.2 Podstawowe pojęcia

Graf skierowany to para $G = (V, E)$, gdzie V jest skończonym zbiorem wierzchołków, a $E \subseteq V \times V$ jest zbiorem krawędzi. Z każdym wierzchołkiem $i \in V$ kojarzymy zmienną losową X_i. Zakładamy, że $V = \{1, 2, \ldots, n\}$ i wszystkie zmienne losowe skojarzone z wierzchołkami ze zbioru V są osadzone na tej samej przestrzeni probabilistycznej. Będziemy zakładać, że wszystkie X_i są dyskretne (tak jak było w modelach log–liniowych) lub że wszystkie są ciągłe. Symbolem X_A, gdzie $A = \{a_1, \ldots, a_k\} \subseteq \{1, \ldots, n\}$ oznaczamy wektor losowy $(X_{a_1}, \ldots, X_{a_k})$. Będziemy również korzystać z pojęcia podgrafu indukowanego. Dla danego podzbioru wierzchołków $W \subset V$ podgrafem indukowanym przez W nazywamy graf

$$G(W) = (W, E(W)), \quad \text{gdzie } E(A) = \{(x, y) \in E : x, y \in A\}.$$

Notacja

- $i \to j \iff (i, j) \in E$,
- $i \sim j \iff i \to j \vee j \to i$,
- $N(v) = \{w \in V : w \sim v\}$ – zbiór sąsiadów wierzchołka v,
- $\mathfrak{P}(v) = \{w \in V : w \to v\}$ – zbiór rodziców wierzchołka v,
- uv-ścieżka, to zbiór krawędzi łączących wierzchołki u i v (bez powtórzeń),
- podgraf indukowany przez zbiór $A \subset V$, to

$$G_A = (A, E_A), \quad E_A = E \cap A \times A,$$

- **kliką** nazywamy maksymalny (w sensie inkluzji) podgraf pełny, czyli zbiór $W \subseteq V$ jest kliką, jeśli $G_W = K_{|}W|$ oraz $G_{W \cup \{w\}} \neq K_{|W|+1}$ dla dowolnego $w \in V \setminus W$.

Definicja 6.6. *Wierzchołek v jest **simplicjalny**, (simplicial) jeśli jego sąsiedztwo tworzy graf pełny, tzn. $G_{Nv} = K_{|N(v)|}$.*

Dekomponowalność

Niech $(A, B; S)$ będzie trójką podzbiorów zbioru V. $(A, B; S)$ tworzy **dekompozycję** grafu G, jeśli

(i) zbiory A, B i S są rozłączne oraz $A \cup B \cup S = V$;

(ii) G_S jest grafem pełnym;

(iii) S separuje A i B.

Definicja 6.7. *Graf jest **dekomponowalny**, jeśli jest grafem pełnym lub istnieje dekompozycja G na grafy dekomponowalne (tzn. istnieje trójka $(A, B; S)$ tworząca dekompozycję G taka, że grafy G_A, G_B i G_S są dekomponowalne).*

Definicja dekomponowalności ma rekursywny charakter i stąd dekomponowalność jest trudna do sprawdzenia. Jednak jest wiele warunków równoważnych, których sprawdzenie jest często dużo prostsze.

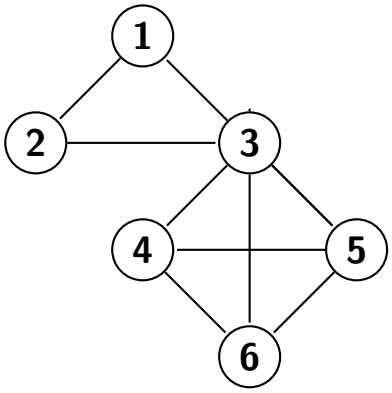
Twierdzenie 6.1. *Graf jest dekomponowalny wtedy i tylko wtedy, gdy jest chordalny.*

Doskonałe uporządkowanie wierzchołków

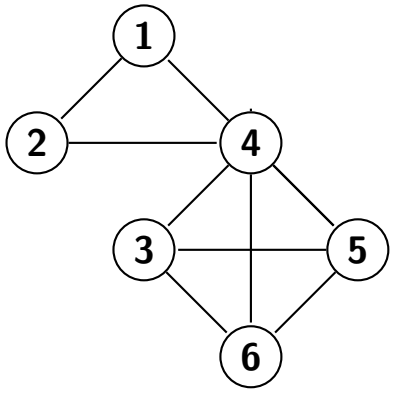
Niech $V = \{1, 2, \ldots, n\}$. Mówimy, że wierzchołki są doskonale uporządkowane (*perfect ordered*), jeśli $\forall_{i=2,\ldots,n}\ S(i) = N(i) \cap \{1; \ldots; i-1\}$ jest grafem pełnym.

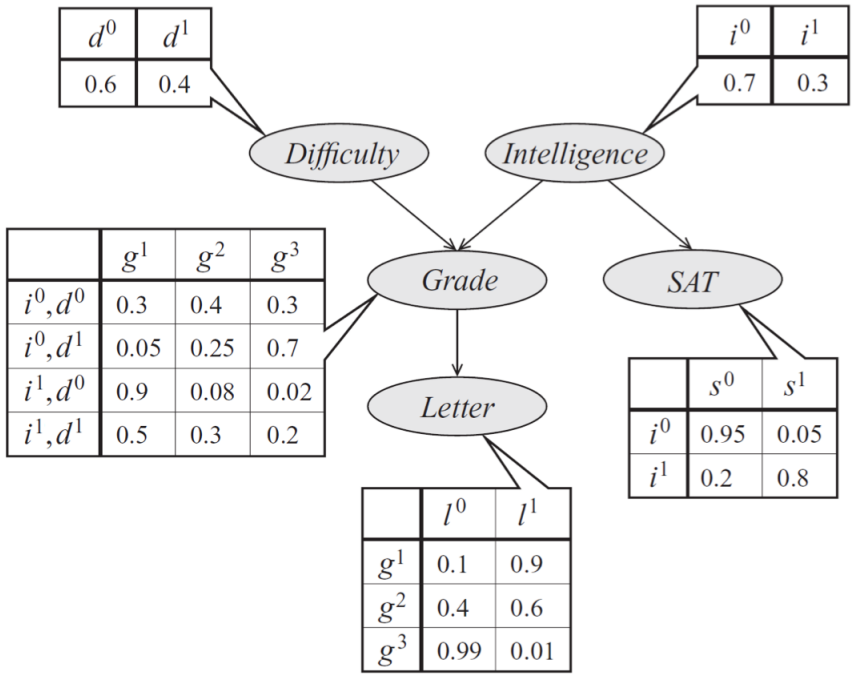
Przykład

To jest porządek doskonały:



A to nie:





Rysunek 6.1: Przykład do omówienia pochodzący ze strony https://ermongroup.github.io/cs228-notes/

6.3 Sieci Bayesowskie

Pojęcie sieci Bayesowskich jest często używane przez praktyków. Jest to "potoczne" określenie na **modele graficzne** oparte na DAGu, tj. grafie skierowanym, w którym nie ma skierowanych cykli. Idea opiera się na zasadzie łańcuchowej:

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)\ldots p(x_n|x_{n-1}, \ldots, x_2, x_1).$$

Pozostajemy przy oznaczeniach z poprzedniego rozdziału: Niech $X = X_V = (X_v)_{v\in V}$ będzie dyskretnym wektorem losowym. Poziomy przyjmowane przez X ozn. $i = i_V = (i_v;\ v \in V)$ i zbiór możliwych wartości $\mathcal{I}$.

Sieć Bayesowska

Sieć Bayesowska to para (G, p), gdzie $G = (V, E)$ jest grafem skierowanym acyklicznym (DAG) i p rozkładem prawdopodobieństwa. O p zakładamy, że **faktoryzuje się względem** G, tzn.

$$p(i_V) = \prod_{v\in V} p(i_v|i_{pa(v)}),$$

gdzie $pa(v)$ jest zbiorem rodziców wierzchołka v

Spróbuj sam!

Przekonaj się, że grafowi przedstawionemu na Rys. 6.1, odpowiada faktoryzacja:

$$p(l, g, i, d, s) = p(l \mid g)\, p(g \mid i, d)\, p(i)\, p(d)\, p(s \mid i).$$

Innym sposobem interpretacji wykresów kierunkowych jest opowieść o tym, jak dane zostały wygenerowane. Jaka historia może kryć się za tym grafem?
W powyższym przykładzie, w celu określenia jakości listu referencyjnego, możemy najpierw wylosować poziomu inteligencji i trudności egzaminu; następnie próbkowana jest ocena studenta z uwzględnieniem tych parametrów; w końcu, na podstawie tej oceny generowany jest list polecający.

Nietrudno zauważyć, że prawdopodobieństwo reprezentowane przez sieć bayesowską będzie poprawne: oczywiście, będzie nieujemne i można pokazać, że suma po wszystkich przypisaniach zmiennych będzie równa 1. I odwrotnie, możemy również pokazać za pomocą kontrprzykładu, że kiedy G zawiera cykle, to związane z nim prawdopodobieństwa nie może się sumować do jednego.

6.4 Modele graficzne nieskierowane

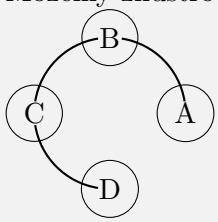
Sieci bayesowskie są klasą modeli, które mogą zgrabnie reprezentować wiele interesujących rozkładów prawdopodobieństwa. Jednak niektóre rozkłady mogą mieć założenia dotyczące niezależności, które nie mogą być reprezentowane przez strukturę sieci bayesowskiej.

W takich przypadkach, o ile nie chcemy wprowadzić fałszywych zależności pomiędzy zmiennymi naszego modelu, musimy wrócić do mniej zwartej reprezentacji (która może być postrzegana jako graf z dodatkowymi, niepotrzebnymi krawędziami). Prowadzi to do dodatkowych, niepotrzebnych parametrów w modelu, a także utrudnia naukę tych parametrów i wykonywanie przewidywań. (Więcej o tym powiemy w części praktycznej).

Istnieje jednak inna technika kompaktowego przedstawiania i wizualizacji rozkładu prawdopodobieństwa, która opiera się na języku grafów nieskierowanych. Ta klasa modeli znana jest pod nazwą Markowskich Pól Losowych (*Markov Random Fields* lub MRFs).

Przykład (wart implementacji!)

Osoby A, B, C, D głosują w wyborach. Wiemy, że (A, B), (B, C) i (C, D) są przyjaciółmi, a pozostali nie znają się. Przyjaciele wpływają na swoje wzajemne preferencje wyborcze. Możemy zilustrować to za pomocą grafu:



Definicja 6.8. *Zbiory A, $B \subseteq V$ są **separowane** przez zbiór $C \subseteq V$, jeśli dla każdego $a \in A$ i dla każdego $b \in B$ przecięcie ab-ścieżki i zbioru C nie jest puste. Oznaczamy $A \perp_C B$.*

Innymi słowy A i B są separowane przez zbiór C, jeśli każda ścieżka z A do B przechodzi przez zbiór C. Definicja nie wymaga, aby rozważane zbiory były rozłączne, jednak jeśli dla zbiorów A i B istnieje separator (czyli zbiór, który jest separuje), to A i B są rozłączne. Co więcej, jeśli C jest separatorem A i B i jeśli $C \subseteq D \subseteq V$, to D jest separatorem A i B. Z reguły przedmiotem zainteresowania będą minimalne separatory (w sensie inkluzji).

Własności Markowa

Niech dany będzie graf $G = (V, E)$. Rozkład p spełnia względem G:

(a) **pairwise własność Markowa**, jeśli

$$\forall u, v \in V : \neg(u \sim v) \implies X_u \perp X_v | X_{V \setminus \{u,v\}},$$

(b) **globalną własność Markowa**, jeśli

$$\forall A, B, C \subseteq V : (C \cap (A \cup B) = \emptyset \wedge A \perp_C B \implies X_A \perp X_B | X_C),$$

(c) **lokalną własność Markowa**, jeśli

$$\forall u \in V : X_u \perp X_{V \setminus N(u)} | X_{N(u)},$$

Definicja 6.9. *Graf G jest perfect–map rozkłądu P, jeśli*

$$A \perp B | C \iff A \perp_S B | C.$$

Twierdzenie 6.2. *Globalna własność Markowa implikuje lokalną własność Markowa. Co więcej, jeśli $p(i) > 0$ dla każdego $i \in \mathcal{I}$, to własności te są równoważne.*

6.5 Budowa sieci Bayesowskiej

Na podstawie R. Nagarayan *Bayesian Networks in R*, Springer, 2013.

W przypadku grafów acyklicznych skierowanych, czyli tzw. DAG (*directed acyclic graphs*), relacje niezależności warunkowych są reprezentowane przez pojęcie tzw. d-separacji (od angielskiego *directed separation*).

Definicja 6.10. *Jeśli A, B, C są trzema rozłącznymi podzbiorami zbioru wierzchołków w DAGu G, to C **d-separuje** A i B, ozn. $A \perp_d B|C$, jeśli każda ścieżka*[1] *z wierzchołka z A do wierzchołka B ma co najmniej jeden wierzchołek v taki, że*

(a) v jest w C i nie jest kolidujący;

LUB

(b) v jest kolidujący i ani v ani żaden z jego potomków nie jest w C.

Collider

Wierzchołek u jest kolidujący na ścieżce P, jeśli istnieją $v \to u, u \leftarrow w \in P$.

Jeśli rozkład P spełnia własność Markowa względem grafu G, to oznacza, że z d-separacji wynikają warunkowe niezależności. Powiemy, że rozkład ten jest **wierny**, jeśli innych warunkowych niezależności nie ma. Mając dany DAG G i rozkład P, który spełnia MP dla G, otrzymujemy, że P faktoryzuje się na iloczyn warunkowych rozkładów w następujący sposób:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\mathfrak{P}(i)),$$

gdzie $\mathfrak{P}(i)$ jest zbiorem rodziców wierzchołka $i \in V(G)$.

[1] przez ścieżkę rozumiemy tutaj sekwencje krawędzi z A do B niezależnie od ich skierowania

3 wierzchołki

Rozważmy trzy możliwe konfiguracje skierowań dla drzewa o 3 wierzchołkach:

(a) $A \to C \to B$,

(b) $A \leftarrow C \to B$,

(c) $A \to C \leftarrow B$.

W przypadku (a), mamy: $\mathfrak{P}(A) = \emptyset$, $\mathfrak{P}(B) = \{C\}$ i $\mathfrak{P}(C) = \{A\}$. Zatem

$$P(A, B, C) = P(B|C) \cdot P(C|A) \cdot P(A).$$

Dla grafu (b): $\mathfrak{P}(A) = \{C\} = \mathfrak{P}(B)$, $\mathfrak{P}(C) = \emptyset$. Stąd

$$P(A, B, C) = P(B|C) \cdot P(A|C) \cdot P(C).$$

W ostatnim przypadku, tj. przypadku gdy C jest wierzchołkiem kolidującym, otrzymujemy: $\mathfrak{P}(A) = \emptyset = \mathfrak{P}(B)$, $\mathfrak{P}(C) = \{A, B\}$ i

$$P(A, B, C) = P(B) \cdot P(A) \cdot P(C|A, B).$$

Zauważmy, że w tym przypadku A i B są niezależne, ale nie są warunkowo niezależne pod warunkiem C.

6.5.1 Klasy równoważności

W powyższym przykładzie można sprawdzić, że przypadki (a) i (b) są tak naprawdę równoważne. Jedno wyrażenie na $P(A, B, C)$ można przekształcić do drugiego stosują twierdzenie Bayesa. (Zrób to jako ćwiczenie!) DAGi, które reprezentują równoważne struktury niezależności tworzą jedną **Markowską klasę równoważności**.

Definicja 6.11. *Graf $G_1 = (V, E_1)$ i graf $G_2 = (V, E_2)$ mają te same v-struktury, jeśli $\{(u, v, w) : u \to v \leftarrow w \text{ w } E_1\} = \{(u, v, w) : u \to v \leftarrow w \text{ w } E_2\}$.*

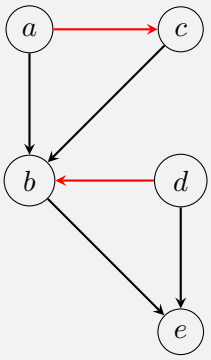
Twierdzenie 6.3 (Chickering, 1995)**.** *Grafy o tych samych v-strukturach są w tej samej Markowskiej klasie równoważności.*

6.5.2 Moralizacja

Niech Z będzie zbiorem v-struktur grafu G. Graf G jest moralny, jeśli $(a, b, c) \in Z \implies a \sim c$ dla dowolnych $a, b, c \in V$. Operacja **moralizacji** grafu polega na dodaniu krawędzi do grafu tak, aby był moralny. Znów, w praktyce interesuje nas minimalny (w sensie inkluzji) zbiór krawędzi, który trzeba dodać. Dodawane krawędzie mogą być skierowane w dowolny sposób.

Przykład

Dany graf (tylko czarne krawędzie) jest niemoralny. Moralizacja polega na dodaniu krawędzi łączących a i c oraz b i d. Na rysunku czerwone krawędzie.



6.6 Modelowanie statycznej sieci Bayesowskiej

Budowanie statycznej sieci bayesowskiej odbywa się w dwóch krokach: wybór modelu (grafu) i estymacji parametrów rozkładu. Pierwszy krok nazywamy *uczeniem struktury* (structure learning) i polega on na zidentyfikowaniu grafu sieci Bayesowskiej. Najważniejsze przykłady używanych w celu zbudowania sieci algorytmów omawiamy na praktycznej części kursu.

6.7 Drzewa łączące i propagowanie informacji

Algorytm propagowania informacji (*message passing*) po drzewie łączącym (*junction tree*) pozwala szybko (i wykonalnie) obliczać prawdopodobieństwa w dużych sieciach bayesowskich: rozkłady łączne podzbiorów wierzchołków, rozkłady brzegowe, warunkowe jak i rozkłady pod warunkiem interwencji, o których będzie mowa później.

Definicja 6.12. *Drzewo T jest grafem łączącym, jeśli jego zbiór wierzchołków $\mathcal{V}$ jest zawarty w zbiorze podzbiorów pewnego V oraz każdy wierzchołek C na ścieżce łączącej C_1, $C_2 \in \mathcal{V}$ spełnia: $C_1 \cap C_2 \subseteq C$.*

Uwaga: Jeśli graf $G = (V, E)$ jest dekomponowalny, to istnieje drzewo łączące $T = (\mathcal{V}, \mathcal{E})$, takie, że zbiór $\mathcal{V}$ jest zbiorem klik w G. Nie ma tu jednoznaczności.

6.7.1 Potencjały

Definicja 6.13. *Niech $\mathcal{C}$ będzie skończoną rodziną podzbiorów zbioru V. Mówimy, że $\mathcal{C}$ spełnia running intersection property (RIP), jeśli $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ dla pewnego $k \in \mathbb{N}$ i dla każdego $j = 2, 3, \ldots, k$ istnieje $\sigma(j) < j$ takie, że*

$$C_j \cap \bigcup_{i=1}^{j-1} C_i = C_j \cap C_{\sigma(j)}.$$

Innymi słowy przecięcie każdego zbioru ze wszystkimi poprzednimi zawiera się w przecięciu z jednym z nich.

Przykład

Zbiory $\{1,2,3\}$, $\{3,4\}$, $\{2,3,5\}$, $\{3,5,6\}$ spełniają RIP w tej kolejności. Formalnie: Jeśli $C_1 = \{1,2,3\}$, $C_2 = \{3,4\}$, $C_3 = \{2,3,5\}$ i $C_4 = \{3,5,6\}$, to rodzina $\mathcal{C} = \{C_1, C_2, C_3, C_4\}$ spełnia RIP.

Nie istnieje kolejność taka, żeby zbiory $C_1 = \{1,2\}$, $C_2 = \{2,3\}$, $C_3 = \{3,4\}$, $C_4 = \{4,1\}$ spełniały RIP.

(Można się o tym przekonać sprawdzając wszystkie możliwe porządki.)

Niech T będzie drzewem łączącym takim, że $\mathcal{V} = \{C_1, \ldots, C_k\}$. Dla $C \in \mathcal{V}$ definiujemy potencjał $\psi_C : \mathbb{R}^{|C|} \to \mathbb{R}_+$.

Definicja 6.14. *Niech $C, D \in \mathcal{V}$. Potencjały ψ_D i ψ_C są zgodne, jeśli*

$$\sum_{x_{C \setminus D}} \psi_C(x_C) = \sum_{x_{D \setminus C}} \psi_D(x_D) = g(x_{C \cap D}).$$

Twierdzenie 6.4. *Niech $C_1, \ldots, C_k$ będą wierzchołkami drzewa łączącego ponumerowanymi zgodnie z RIP. Niech $S_2, \ldots, S_k$ będą odpowiadającymi separatorami, tzn. $S_k = C_k \cap C_{k-1}$. Niech $p(x_V) = \prod_{i=1}^{k} \frac{\phi_{C_i}(x_{C_i})}{\phi_{S_i}(x_{S_i})}$, gdzie $S_1 = \emptyset$ i $\phi_\emptyset = 1$). Wówczas $\phi_{C_i}(x_{C_i}) = p(x_{C_i})$ dla każdego $i = 1, \ldots, k$ wtedy i tylko wtedy, gdy każda para potencjałów (o nie pustym przecięciu) jest zgodna.*

Dowód. Dowód implikacji "tylko wtedy" jest natychmiastowy.

Implikację w drugą stronę (czyli w prawo) dowiedziemy poprzez indukcję po k. Zakładamy zatem, że każda para potencjałów jest zgodna. Dla $k = 1$ nie ma czego dowodzić. Dla k większych od jedynki, niech $R_k = C_k \setminus S_k$. Wtedy

$$p\left(x_{V \setminus R_k}\right) = \sum_{x_{R_k}} p(x_V) = \prod_{i=1}^{k-1} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})} \times \frac{1}{\psi_{S_k}(x_{S_k})} \sum_{x_{R_k}} \psi_{C_k}(x_{C_k}).$$

Ostatnia równość wynika z tego, że $R_k = C_k \setminus \sum_{j<k} C_j$. Dalej, ponieważ zakładamy zgodność potencjałów, to

$$\sum_{x_{R_k}} \psi_{C_k}(x_{C_k}) = \psi_{S_k}(x_{S_k}),$$

więc

$$p\left(x_{V\setminus R_k}\right) = \prod_{i=1}^{k-1} \frac{\psi_{C_i}(x_{C_i})}{\psi_{S_i}(x_{S_i})}. \tag{6.2}$$

Z założenia indukcyjnego $\psi_{C_i}(x_{C_i}) = p(x_{C_i})$ dla $i \leq k-1$. Z RIPa wynika, że $S_k = C_j \cap C_k$ dla pewnego $j < k$. To wraz ze zgodnością daje:

$$\psi_{S_k}(x_{S_k}) = \sum_{C_j \setminus S_k} \psi_{C_j}(x_{C_j}) = \sum_{C_j \setminus S_k} p(x_{C_j}) = p(x_{S_k}).$$

Wstawiamy teraz (6.2) do wyrażenia z treści twierdzenia:

$$p(x_V) = p(x_{V\setminus R_k}) \frac{\psi_{C_k}(x_{C_k})}{\psi_{S_k}(x_{S_k})} = p(x_{V\setminus R_k}) \frac{\psi_{C_k}(x_{C_k})}{p(x_{S_k})}.$$

Stąd, wprost z definicji prawdopodobieństwa warunkowego,

$$\frac{\psi_{C_k}(x_{C_k})}{p(x_{S_k})} = p(x_{R_k} | x_{V\setminus R_k}).$$

Ponieważ wyrażenie z lewej strony zależy tylko od C_k, to $p(x_{R_k}|x_{V\setminus R_k}) = p(x_{R_k}|x_{S_k})$ (bo $S_k = C_k \cap (V \setminus R_k)$). (Proszę się upewnić, że to widać!) Ostatecznie więc (z dwóch ostatnich równości)

$$\psi_{C_k}(x_{C_k}) = p(x_{S_k})\, p(x_{R_k}|x_{S_k}) = p(x_{R_k \cup S_k}) = p(x_{C_k}).$$

Uaktualnienie (*update*) potencjału

Przypuśćmy, że kliki C i D sąsiadują w drzewie łączącym oraz $S = D \cap C$. *Update* z C do D polega na aktualizacji wartości ψ_S i ψ_D w następujący sposób:

$$\psi'_S(x_S) = \sum_{x_{C\setminus S}} \psi_C(x_C), \quad \psi'_D(x_D) = \frac{\psi'_S(x_S)}{\psi_S(x_S)} \psi_D(x_D).$$

Zauważmy, że

1. po update, potencjały ψ_S i ψ_C są zgodne (wykonać rachunek),
2. jeśli ψ_D i ψ_S było zgodne przed update, to dalej są zgodne (wykonać rachunek),
3. Iloraz iloczynu potencjałów klik i iloczynu potencjałów separatorów pozostaje niezmienny.

Algorytm propagacji informacji na drzewie łączącym - API

Niech $1 < 2 < \ldots < k$ będzie topologicznym porządkiem wierzchołków w drzewie łączącym T.

```
For t in k, \ldots, 2 do:
  update z \psi_t do \psi_k, k\in pa(t);
End for.

For t in 2, ..., k do:
  update z \psi_k do \psi_t, k \in pa(t);
End for.

Zwróć potencjały.
```

Twierdzenie 6.5. *Niech T będzie drzewem łączącym. Po API wszystkie pary potencjałów będą zgodne.*

Dowód. Wynika z własności (1-3). ▼

Potencjały wygodnie jest inicjalizować następująco. Dla separatorów ustalamy $\psi_S = 1$. Ponadto każdy wierzchołek przypisujemy do dokładnie jednej z klik zawierających ten wierzchołek i wszystkich jego rodziców. Niech $v(C)$ oznacza zbiór wierzchołków przypisanych do kliki C. Wówczas możemy przyjąć: $\psi_c(x_C) = \prod_{v \in v(C)} p(x_v | x_{pa(v)})$.

6.7.2 Triangulizacja

W części praktycznej zapoznamy się z algorytmiczną metodą trangulizacji o nazwie *Tarjan method*.

6.8 Wnioskowanie przyczynowe (*causal inference*)

Korelacja czy zależność zmiennych nie musi oznaczać związku przyczynowo skutkowego między nimi (*correlation does not imply causation*, jak głosi słynny frazes). Żeby lepiej to sobie uświadomić rozważmy pewną osobę, która może mieć lub nie mieć raka płuc oraz może palić lub nie palić (tak jako kot Shrodingera). Mamy więc tu prosty układ dwóch zmiennych. Niech X będzie zmienną binarną oraz niech $\{X = 1\} = \{\omega \in \Omega : \text{ osoba pali}\}$. Dalej niech Y będzie zmienną binarną taką, że $\{Y = 1\} = \{\omega \in \Omega : \text{ osoba ma raka płuc}\}$. Założenie, że zmienne X i Y nie są niezależne wydaje się sensowne. Jednak sam brak niezależności nie mówi nam nic o kierunku wpływu! Czy palenie wpływa na prawdopodobieństwo wystąpienia raka czy odwrotnie?

Dlatego wprowadzono dodatkowy nośnik wpływu, który oznacza coś więcej niż tylko zależność. Mówi bowiem również o kierunku i charakterze tej zależności. Oczywiście, to nie jest doskonały nośnik, ale mówi więcej niż rozkład warunkowy.

Operator 'do'

$$p\left(x_{V\setminus\{w\}} \,|\, do(x_w^*)\right) = \frac{p(x_v)}{p\left(x_w^\star \,|\, x_{pa(w)}\right)} I_{\{x_w = x_w^\star\}}$$

Popatrzmy na model graficzny zadany przez graf: $Z \to X \to Y$ oraz $Z \to Y$. Policzmy $p\left(z, y \,|\, do(x^\star)\right)$.

Ponieważ

$$p\left(x, y, z\right) = p\left(z\right)\, p\left(x \,|\, z\right)\, p\left(y \,|\, z, x\right),$$

to

$$p\left(z, y \,|\, do(x^\star)\right) = p\left(z\right)\, p\left(y \,|\, z, x^\star\right).$$

Podczas gdy rozkład warunkowy wygląda następująco

$$p\left(z, y \,|\, x^\star\right) = p\left(z \,|\, x^\star\right)\, p\left(y \,|\, z, x^\star\right).$$

Wykorzystanie i interpretacje tego operatora omówimy i przećwiczymy w części praktycznej.

6.9 Literatura

Uogólnione modele Liniowe

- Agresti, *Foundations of linear and generalized linear models*, Wiley 2015;
- Dobson, Barnett, *Introduction to Generalized Linear Models*, Chapman & Hall / CRC Press 2008;

- McCullagh, Nelder, *Generalized Linear Models*, CRC Press Book 1989.

Modele graficzne

- Højsgaard, Edwards, Lauritzen, *Graphical Models with R*, Springer 2012;
- Evans, *Graphical models: lecture notes*, 2018 (na stronie internetowej autora).