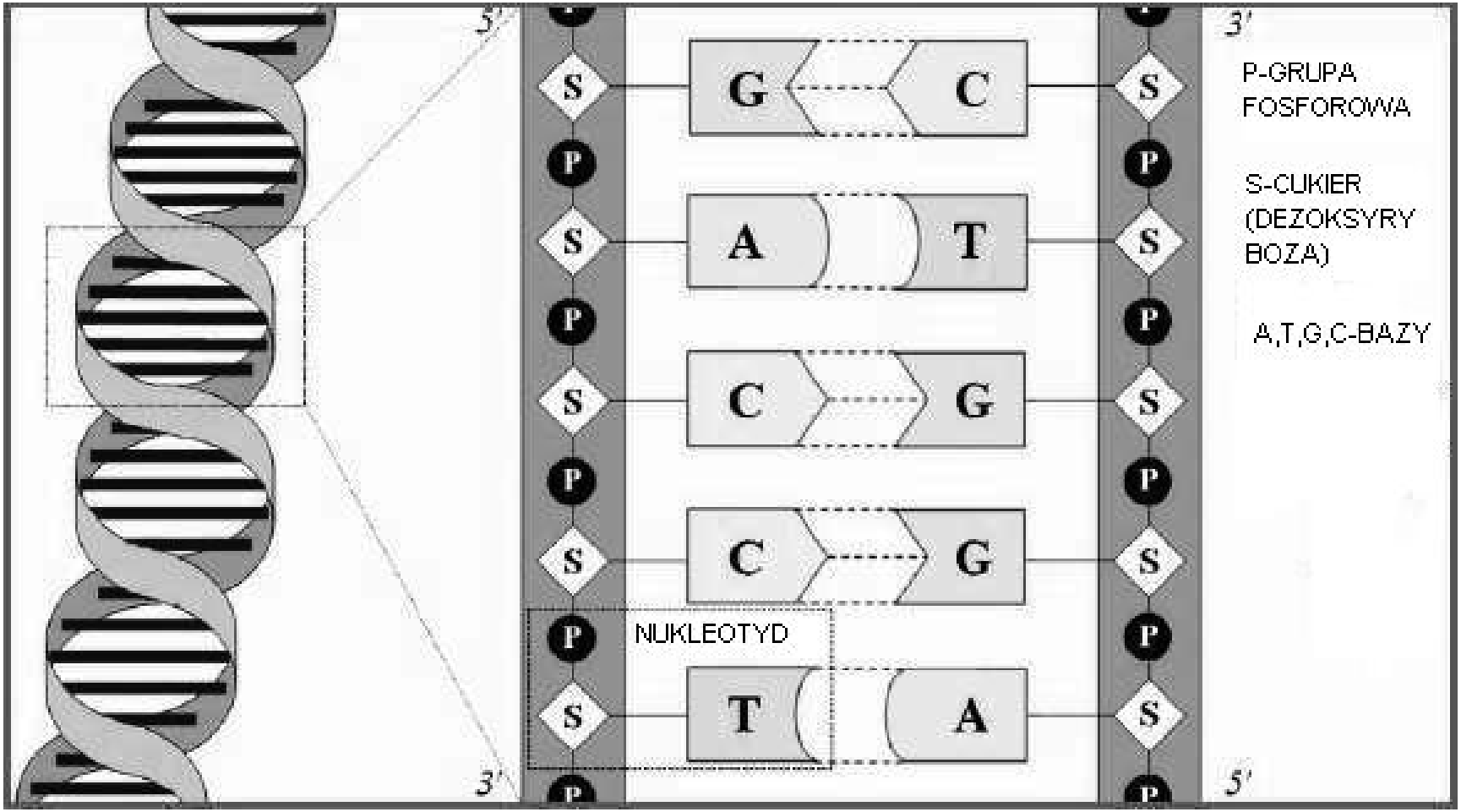
Metody oparte na logicznej regresji w zastosowaniu do wykrywania interakcji SNPów

Magdalena Malina

Instytut Matematyczny, Uniwersytet Wrocławski

Wisła, 9 grudnia 2009

DNA


5'
3'
S
P
G
C
A
T
C
G
C
G
NUKLEOTYD
T
A
3'
P-GRUPA FOSFOROWA
S-CUKIER (DEZOKSYRYBOZA)
A,T,G,C-BAZY
5'

Zmienność genetyczna

- ***Polimorfizm*** to zmiana w strukturze DNA, obecna u co najmniej 1% populacji
- Polimorfizm pojedynczego nukleotydu ***SNP(Single Nucleotide Polimorphism)***-zmiany dotycza różnicy na pojedynczej bazie
- **GŁÓWNY CEL:** znalezienie mutacji w ciągu DNA, mających wpływ na badaną cechę.
- **METODY:**
  - regresja logiczna (*Ruczinski, Kooperberg, LeBlanc* (2003)) - zmiennymi objaśniającymi są wyrażenia logiczne otrzymane ze zmiennych binarnych
  - regresja logiczna z algorytmem Monte Carlo (*Ruczinski, Kooperberg* (2005))
  - w pełni bayesowska wersja logicznej regresji (*Fritsch*(2006))
  - bayesowska wersja regresji logistycznej, w której zmiennymi objaśniającymi są iloczyny binarnych predykatorów

Regresja Logiczna z predykatorami genetycznymi

Predykatorami są SNPy o możliwych trzech wartościach: $0, 1$ lub 2.
Zmienne pomocnicze (*Kooperberg et.al*(2001)):

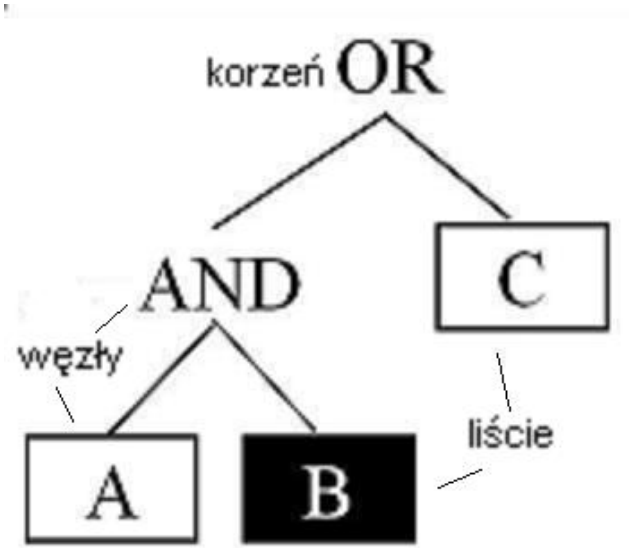
- X_d- efekt dominujący, $X_d = \begin{cases} 0, & SNP = 0 \\ 1, & SNP = 1, 2 \end{cases}$,
- X_r-efekt recesywny, $X_r = \begin{cases} 0, & SNP = 0, 1 \\ 1, & SNP = 2 \end{cases}$.

SNP	X_d	X_r	Genotyp
0	0	0	Homozygota (reference)
1	1	0	Heterozygota
2	1	1	Homozygota (variant)

Wyrażenia logiczne

Niech $X_1, X_2, \ldots, X_m$ będą predykatorami binarnymi.
Wówczas:

- ***wyrażeniem logicznym*** jest każda kombinacja zmiennych X_i, uzyskana przez zastosowanie operatorów logicznych $\wedge$ (AND), $\vee$ (OR) oraz ^{C} (NOT)
- Każde wyrażenie logiczne może być przedstawione za pomocą drzewa binarnego, np: $L = (A \wedge B^c) \vee C$ możemy przedstawić w postaci drzewa



Model regresji logicznej

- Dopasowujemy model regresji

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j,$$

gdzie L_j są wyrażeniami logicznymi otrzymanymi ze zmiennych binarnych $X_i, i = 1, 2, \ldots, m$.

- ***Rozmiar modelu*** to liczba liści w modelu; ustalany w oparciu o
  - zbiór testowy i treningowy
  - metodę kroswalidacji
- **Przykład:** $\mathbb{X} = (X_1, X_2, \ldots, X_m)$ - macierz zmiennych binarnych; Y - status chory/zdrowy
  - Model Regresji Logicznej: $\log(\frac{E[Y]}{1-E[Y]}) = X_1 \vee (X_2 \wedge X_3)$
  - Model Regresji Logistycznej z iloczynami : $\log(\frac{E[Y]}{1-E[Y]}) = X_1 + X_2 * X_3 - X_1 * X_2 * X_3$

W pełni bayesowska wersja Logicznej Regresji (FBLR)

Fritsch (2006) zaproponował wersję bayesowską logicznej regresji (FBLR)

- wyrażenia logiczne tylko ze spójnikiem AND w jednoznacznej reprezentacji ***np.:*** $X_4 \wedge X_1^C \wedge X_7$ jako $(1C, 4, 7)$

Rozkłady a priori:

- k - liczba predykatorów , $p(k) \sim \mathcal{U}(\{0, 1, \ldots, k_{max}\})$,
- β - wektor parametrów regresji, $p(\beta|v, k) \sim \mathcal{N}(0, v * \mathbb{I}_{k+1})$,
- $p(v) \sim \mathrm{InvGamma}(\tau = 0.001; \nu = 0.1)$, (Holmes, Denison(2003)).
- s_i - liczba zmiennych binarnych w wyrażeniu logicznym L_i, $p(s_i) \propto (a)^{s_i}$, $(a = 0.7)$
- *bin* - wektor indeksów zmiennych binarnych włączonych do modelu $p(bin|s_i) = \frac{1}{\binom{2m}{s_i}}$

$$p(\theta) = p(k) * p(v) * p(\beta|v, k) * \prod_{i=1}^{k} p(s_i) * p(bin|s_i),$$

Rozważany model

Rozważamy prosty model z jedną zmienną objaśniającą:

$$M_1 : y_i = \beta_0 + \beta_1 \cdot x_{ij} + \varepsilon_i,$$

$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m,$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ jest znane.

- $x_{ij} \sim b(1, p)$, p znane,
- $\beta = (\beta_0, \beta_1)$, $\hat{\beta}$ - estymator największej wiarogodności dla β,
- $f(\beta|M_1) = p(v)p(\beta|v)$ - gęstość *a priori* rozkładu β w modelu M_1,
  - $p(\beta|v) = \frac{1}{2\pi v} \exp(-\frac{1}{2v}(\beta_0^2 + \beta_1^2))$,
  - $p(v) = \frac{\nu^\tau \exp(-\frac{\nu}{v})}{v^{\tau+1}\Gamma(\tau)} \mathbb{I}_{(0,+\infty)}(v)$.
- $\pi(M_K) \propto a^K \cdot \frac{1}{m_K}$- prawdopodobieństwo *a priori* modelu M_K, $K = 0, 1$, m_K - liczba modeli rozmiaru K.
- prawdopodobieństwo *a posteriori* modelu M_K:

$$P(M_K|Y) \propto L(Y|M_K)\pi(M_K)$$

Rozważany model

Testujemy hipotezę $H0 : \beta_1 = 0$ przy alternatywie $H1 : \beta_1 \neq 0$.

- metoda FBLR odrzuca hipotezę $H0$, gdy $P(M_1|Y) > P(M_0|Y)$.
- funkcja wiarogodności w K-tym modelu:

$$L(Y|M_K) = \int L(Y|M_K, \beta) f(\beta|M_K) d\beta$$

czyli

$$L(Y|M_1) = \frac{\nu^\tau \tau 2^{\tau+1}}{(\sqrt{2\pi}\sigma)^n 2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(- \frac{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{ij}))^2}{2\sigma^2} \right) \cdot$$

$$\cdot (2\nu + \beta_0^2 + \beta_1^2)^{-(\tau+1)} d\beta_0 d\beta_1$$

Stosując aproksymację Laplace'a otrzymujemy oszacowanie

$$\log(L(Y|M_1)) = \log(L(Y|M_1, \hat{\beta})) - \log(n) + \log(R_1),$$

gdzie

$$R_1 = \frac{\nu^\tau \tau 2^{\tau+1} \sigma^2}{S(2\nu + \hat{\beta}_0^2 + \hat{\beta}_1^2)^{\tau+1}},$$

i podobnie

$$\log(L(Y|M_0)) = \log(L(Y|M_0, \hat{\beta})) + \log(R_0),$$

gdzie

$$R_0 = \frac{\nu^\tau \tau 2^{\tau+1} \sigma^2}{2\pi S(2\nu)^{\tau+1}}.$$

A więc odrzucamy hipotezę H_0 gdy

$$\log\left(\frac{L(Y|M_1, \hat{\beta})}{L(Y|M_0, \hat{\beta})}\right) > \log(n) + \log\left(\frac{m_1}{a}\right) - \log(R_1) + \log(R_0).$$

Oszacowanie prawdopodobieństwa błędu I rodzaju

Przy założeniu, że $\mathbb{X}^T\mathbb{X} = n\mathbb{I}_{(m+1)\times(m+1)}$, mamy

$$\log\left(\frac{L(Y|M_1,\hat{\beta})}{L(Y|M_0,\hat{\beta})}\right) = \frac{n\hat{\beta}_1^{\,2}}{2\sigma^2}.$$

Przy $c = 2(\log(n) + \log\left(\frac{m_1}{a}\right) - \log(R_1) + \log(R_0))$, prawdopodobieństwo błędu I rodzaju α_{n,m_1} wynosi

$$P\left(\frac{n\hat{\beta}_1^{\,2}}{\sigma^2} > c\right) = 2P(\frac{\sqrt{n}\hat{\beta}_1}{\sigma} > \sqrt{c}), \quad \text{gdzie} \quad \frac{\sqrt{n}\hat{\beta}_1}{\sigma} \sim \mathcal{N}(0,1).$$

Stosując oszacowanie : $P(\frac{\sqrt{n}\hat{\beta}_1}{\sigma} > \sqrt{c}) \approx \frac{1}{\sqrt{2\pi c}}\exp\left(-\frac{c}{2}\right)$ dostajemy

$$\alpha_{n,m_1} = \frac{a}{n\cdot m_1\sqrt{\pi(\log(n) + \log(\frac{m_1}{a}) - \log(R_1) + \log(R_0)}}\frac{R_1}{R_0}.$$

Korekta na wielokrotne testowanie

W modelu z interakcjami rzędu 2

$$\alpha_{int,n,m_2} = \frac{a^2}{n \cdot m_2 \sqrt{\pi(\log(n) + \log(\frac{m_2}{a^2}) - \log(R_1) + \log(R_0)}} \frac{R_1}{R_0}.$$

Przy pominięciu składników resztowych R_0, R_1:

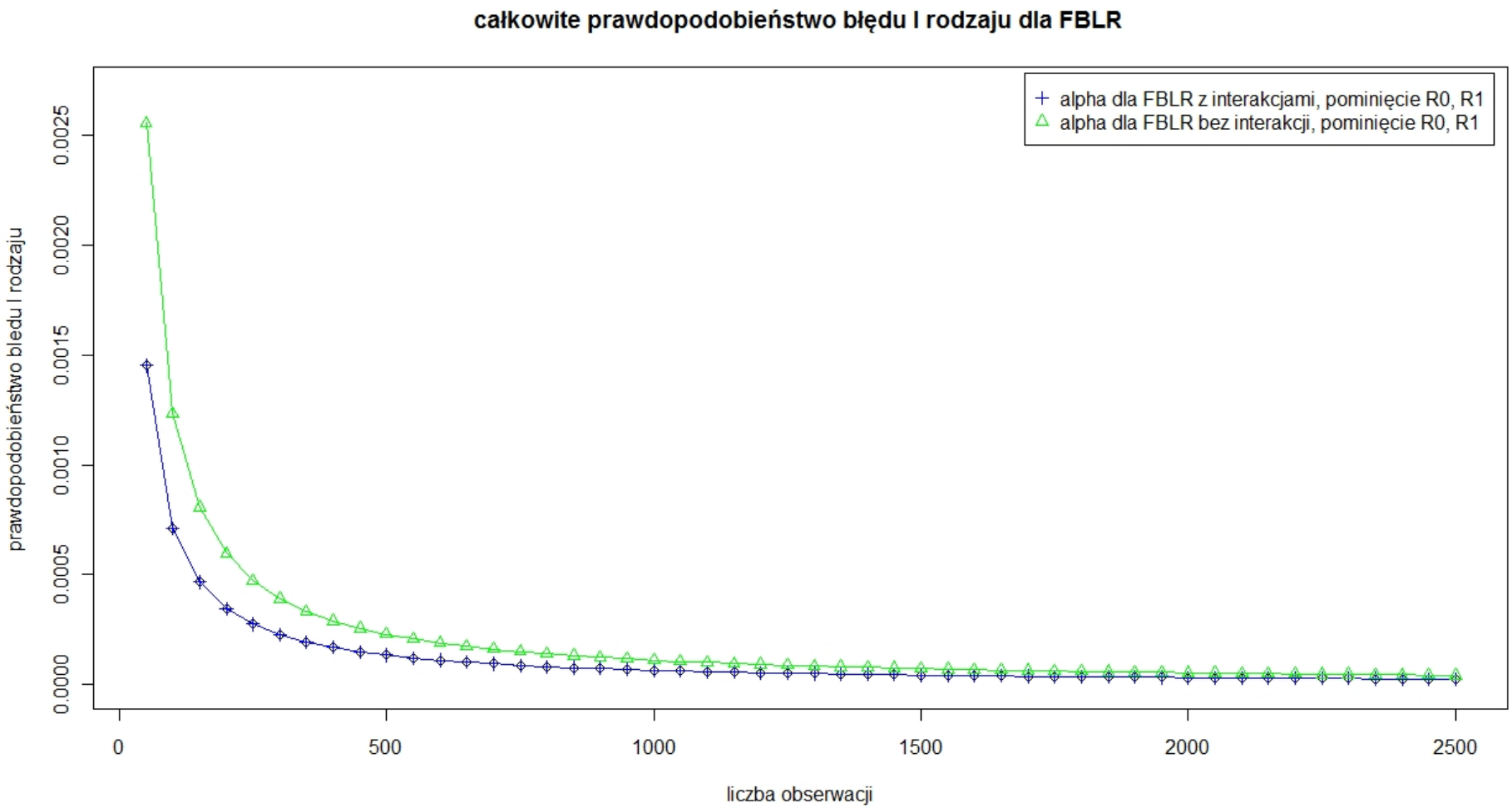
$$\alpha_{n,m_1} = \frac{a}{n \cdot m_1 \sqrt{\pi(\log(n) + \log(\frac{m_1}{a})}},$$

$$\alpha_{int,n,m_2} = \frac{a^2}{n \cdot m_2 \sqrt{\pi(\log(n) + \log(\frac{m_2}{a^2})}}.$$

- automatyczna korekta na wielokrotne testowanie.

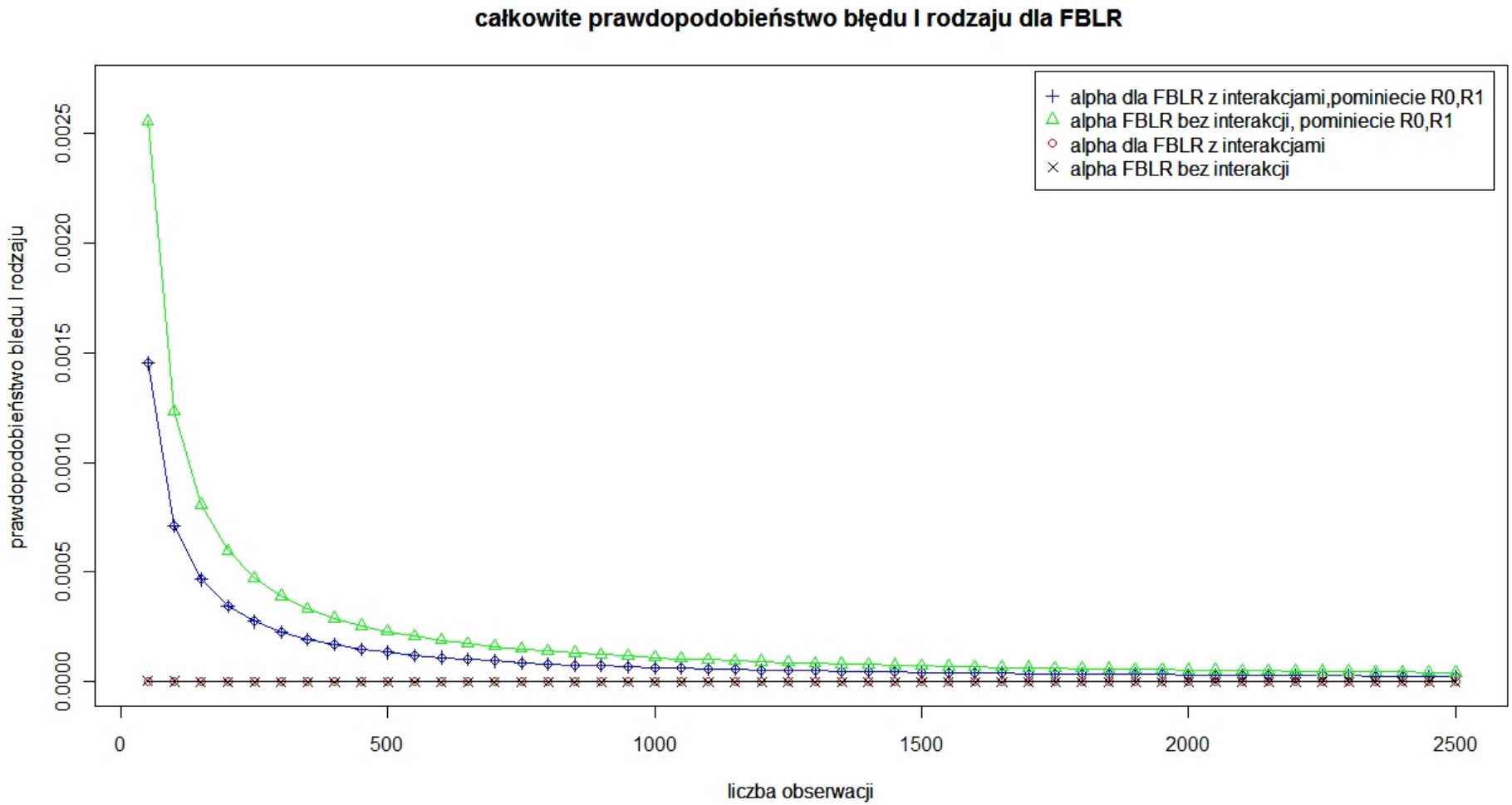
Całkowity błąd I rodzaju - oszacowanie Laplace'a

Liczba obserwacji n rośnie od 50 do 2500:



Badanie wpływu czynników R_0 i R_1

Badamy wartości α_n i $\alpha_{int,n}$ z włączeniem czynników R_0, R_1 i z ich pominięciem.
Liczba obserwacji n rośnie od 50 do 2500:



Uwzględnienie rozkładu składników resztowych

Gdy R_0, R_1 uwzględniamy jako mające wpływ na rozkład

- Prawdopodobieństwo błędu I rodzaju dla FBLR:

$$\alpha_{n,m_1} = P\left(Z > \log(n) + \log\left(\frac{m_1}{a}\right) - \log\left(\frac{(2\nu)^{(\tau+1)}}{2\pi}\right)\right),$$

gdzie

$$Z = \frac{n\hat{\beta}_1^{\,2}}{\sigma^2} - (\tau+1)\log(2\nu + \hat{\beta}_0^{\,2} + \hat{\beta}_1^{\,2}).$$

- Konsekwencją asymptotycznej normalności estymatorów największej wiarogodności $\hat{\beta}$ jest, że

$$n\hat{\beta}_0^{\,2} \cdot \left(\frac{n\nu + (\tau+1)\sigma^2}{\sigma^2\nu} - \frac{n^2p^2\nu}{\sigma^2(n\nu p + (\tau+1)\sigma^2)}\right) \sim \chi^2(1)$$

oraz

$$n\hat{\beta}_1^{\,2} \cdot \left(\frac{n\nu p + (\tau+1)\sigma^2}{\sigma^2\nu} - \frac{n^2p^2\nu}{\sigma^2(n\nu + (\tau+1)\sigma^2)}\right) \sim \chi^2(1)$$

Uwzględnienie rozkładu składników resztowych

Stąd i z dwustronnego oszacowania logarytmu

$$\frac{\hat{\beta}_1^{\,2}\left(\frac{p^2+1}{2\nu}\right)}{1+\hat{\beta}_1^{\,2}\left(\frac{p^2+1}{2\nu}\right)} \leqslant \log\left(1+\hat{\beta}_1^{\,2}\left(\frac{p^2+1}{2\nu}\right)\right) \leqslant \hat{\beta}_1^{\,2}\left(\frac{p^2+1}{2\nu}\right)$$

otrzymujemy ($\forall t \in \mathbb{R}$) następujące oszacowanie rozkładu zmiennej Z

$$F_{\chi^2(1)}\left(\frac{c_1(t+(\tau+1)\log(2\nu))}{\frac{1}{2\sigma^2}-(\tau+1)\frac{p^2+1}{2\nu n}}\right) \leqslant F_Z(t) \leqslant F_{\chi^2(1)}\left(\frac{c_1(t+(\tau+1)\log(2\nu))}{\frac{1}{2\sigma^2}-(\tau+1)\frac{p^2+1}{3\nu n}}\right),$$

gdzie $F_{\chi^2(1)}(\cdot)$ oznacza dystrybuantę centralnego rozkładu $\chi^2(1)$ z jednym stopniem swobody i

$$c_1 = \left(\frac{n\nu p+(\tau+1)\sigma^2}{\sigma^2\nu} - \frac{n^2p^2\nu}{\sigma^2(n\nu+(\tau+1)\sigma^2)}\right) \quad .$$

Oszacowanie prawdopodobieństwa błędu I rodzaju

Twierdzenie: Dla testowania hipotezy $H_0 : \beta_1 = 0$ przy alternatywie $H_1 : \beta_1 \neq 0$ w metodzie FBLR dla pojedynczych zmiennych prawdopodobieństwo błędu pierwszego rodzaju dla pojedynczego testu α_{n,m_1} spełnia warunek

$$1-F_{\chi^2(1)}\left(\frac{c_1 \log\left(\frac{2\pi m_1}{a}\right)}{\frac{1}{2\sigma^2}-(\tau+1)\frac{p^2+1}{3\nu n}}\right) \leqslant \alpha_{n,m_1} \leqslant 1-F_{\chi^2(1)}\left(\frac{c_1 \log\left(\frac{2\pi m_1}{a}\right)}{\frac{1}{2\sigma^2}-(\tau+1)\frac{p^2+1}{2\nu n}}\right)$$

gdzie $F_{\chi^2(1)}(\cdot)$ oznacza dystrybuantę centralnego rozkładu $\chi^2_{(1)}$ z jednym stopniem swobody i

$$c_1 = \left(\frac{n\nu p+(\tau+1)\sigma^2}{\sigma^2\nu} - \frac{n^2p^2\nu}{\sigma^2(n\nu+(\tau+1)\sigma^2)}\right) \quad .$$

Literatura

[1]Ruczinski I., Kooperberg C., LeBlanc M., *Logic regression*, J. Comput. Graphical Statist. 12 (3),(2003),474-511,
[2]Kooperberg C., Ruczinski I., *Identifying Interacting SNPs Using Monte Carlo Logic Regression*, Genetic Epidemiology 28, 157-170 (2005)
[3]Fritsch A., Ickstadt K., *Comparing Logic Regression Based Methods for Identifying SNP Interactions* , Springer Berlin / Heidelberg, Lecture Notes in Computer Science, Volume 4414/2007, pp 90-103
[4]Fritsch A., *A Full Bayesian Version of Logic regression for SNP Data* , Diploma Thesis, (2006)
[5]Scott J.G. and Berger J.O., *Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.*, Duke University Department of Statistical Science Technical Report (2008).
[6]Holmes,C.C and Denison D.G.T, *Classification with Bayesian MARS*, Mach. Learn.**50**(2003), 159-173
[7]Green, P.J. (1995). *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*. Biometrika 82, 711-732.

Symulacje

- **DANE:** Uproszczona wersja danych SNP-owych: po 1000 osób, z 50, 200 i 300 SNP-ami;
- **METODY:**
  - FBLR: $logit(P(Y=1)) = \beta_0 + \sum_{j=1}^{k} \beta_j L_j$,
    - w L_i możliwe dopełnienia
  - Bayesowska wersja regresji logistycznej dla interakcji:
    $logit(P(Y=1)) = \gamma_0 + \sum_{j=1}^{t} \gamma_j I_j$,
    - I_j proste iloczyny zmiennych, bez dopełnień
- Dla każdego modelu tworzymy 20 zbiorów danych
- Dla każdego zbioru danych obliczamy liczbę właściwie i niewłaściwie klasyfikowanych interakcji SNPów
- Wynik uśredniony przez 20 zbiorów danych
- Pojedynczy SNP reprezentowany przez dwie zmienne X_i z rozkładu $b(1, 0.25)$

- Liczba interakcji SNPów w prawdziwym modelu=1

Model 1. $Z = X_{1d} \wedge X_{2d} + X_{11d} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$

	#SNP	Poprawne	Niepoprawne
FBLR	50	0.3	0.0
	200	0.05	0.0
	300	0.0	0.0
Bayesowska	50	0.65	0.0
Regresja	200	0.2	0.0
Logistyczna	300	0.05	0.05

Model 2. $Z = X^C_{7d} \wedge X^C_{9d} \wedge X^C_{11d} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$

	#SNP	Poprawne rzędu 2 lub 3	Poprawne rzędu 1	Niepoprawne
FBLR	50	1.0	0.0	0.0
	200	0.85	0.3	0.15
	300	0.7	0.3	0.1
Bayesowska	50	0.15	2.75	0.2
Regresja	200	0.0	2.65	0.15
Logistyczna	300	0.0	2.85	0.1

Wyniki symulacji (Fritsch, Ickstadt(2007))

- 10 zbiorów danych dla każdego modelu
- Dla każdego zbioru danych obliczamy liczbę właściwie i niewłaściwie klasyfikowanych interakcji SNPów
- Wynik uśredniony przez 10 zbiorów danych
- Pojedynczy SNP reprezentowany przez dwie zmienne X_i z rozkładu $b(1, 0.3)$
- $\eta = \log(\frac{P(Y=1)}{P(Y=0)})$

1. $\eta = -0.70 + 1.0 * (X_{1d} \wedge X_{2d}^C) + 1.0 * (X_{3d} \wedge X_{4d}) + 1.0 * (X_{5d} \wedge X_{6d})$

- Liczba interakcji SNPów w prawdziwym modelu= 3

MODEL 1	poprawne	niepoprawne
Regresja Logiczna	1.3	2.9
MCLR $t = 2, a = \frac{1}{2}$	1.8	1.4
MCLR $t = 3, a = \frac{1}{\sqrt{2}}$	2.6	1.3
FBLR	2.7	0.0

Wyniki symulacji(Fritsch, Ickstadt(2007))

2. $\eta = -0.45 + 0.6 * (X_{1d} \wedge X_{2d}^C) + 0.6 * (X_{3d} \wedge X_{4d}) + 0.6 * (X_{5d} \wedge X_{6d})$

- Liczba interakcji SNPów w prawdziwym modelu= 3

MODEL 2	poprawne	niepoprawne
Regresja Logiczna	0.0	1.3
MCLR $t = 2, a = \frac{1}{2}$	0.5	0.3
MCLR $t = 3, a = \frac{1}{\sqrt{2}}$	0.9	0.7
FBLR	1.1	0.1

3. $\eta = -0.40 + 1.0 * (X_{1d} \wedge (X_{2d}^C \vee X_{3d}))$

- Liczba interakcji SNPów w prawdziwym modelu= 3

MODEL 3	poprawne	niepoprawne
Regresja Logiczna	0.1	2.1
MCLR $t = 2, a = \frac{1}{2}$	1.5	0.3
MCLR $t = 3, a = \frac{1}{\sqrt{2}}$	1.4	0.4
FBLR	0.9	0.2