Metody statystyczne w lokalizacji genów

Małgorzata Bogdan

Instytut Matematyki
Uniwersytet Wrocławski

Zakopane, 06/09/2016

Plan

- Lokalizacja genów

Plan

- Lokalizacja genów
- Wielokrotne testowanie

Plan

- Lokalizacja genów
- Wielokrotne testowanie
- Kryteria wyboru modelu

Plan

- Lokalizacja genów
- Wielokrotne testowanie
- Kryteria wyboru modelu
- Zmodyfikowane wersje Bayesowskiego Kryterium Informacyjnego

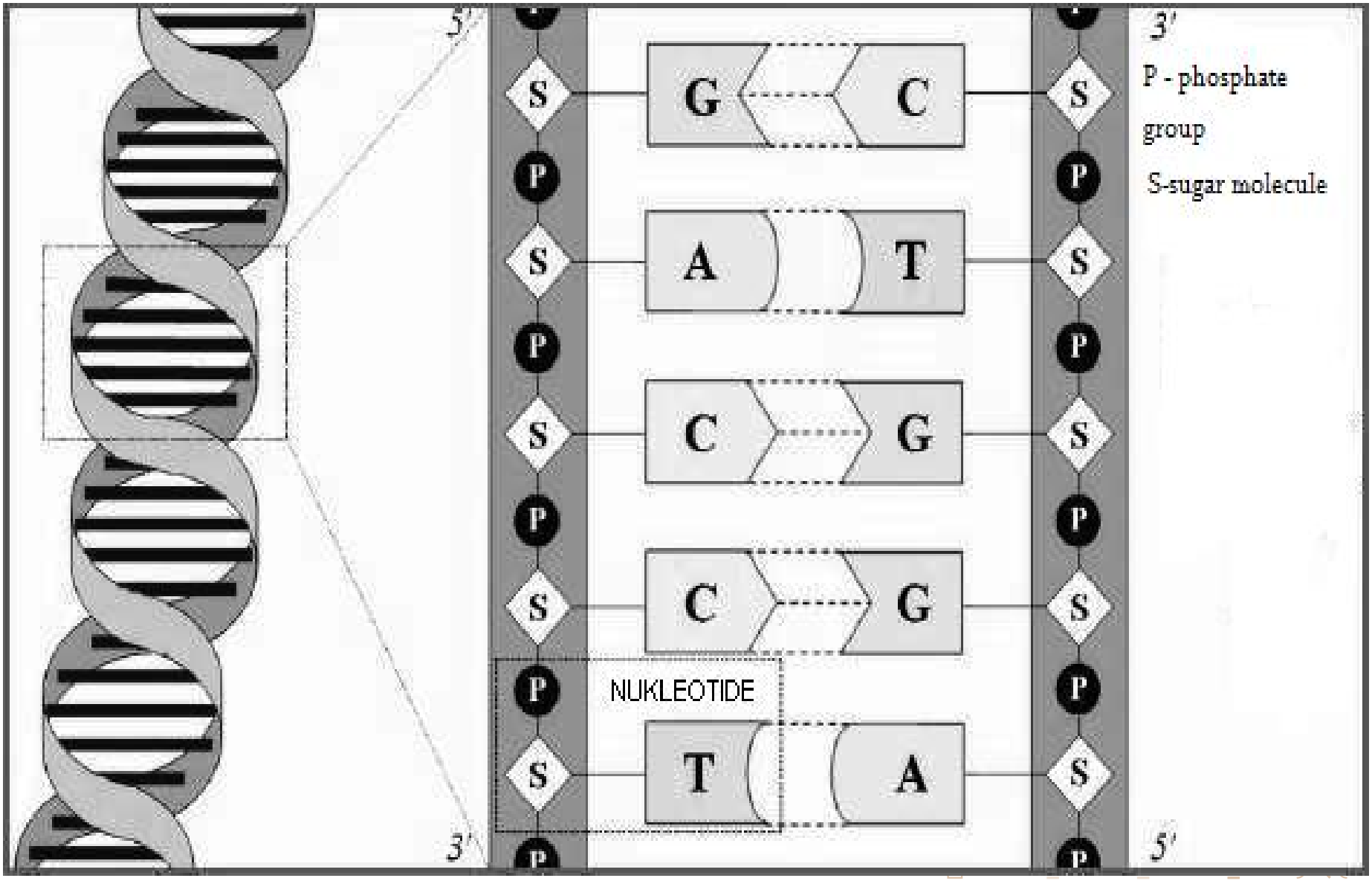
Plan

- Lokalizacja genów
- Wielokrotne testowanie
- Kryteria wyboru modelu
- Zmodyfikowane wersje Bayesowskiego Kryterium Informacyjnego
- SLOPE (Sorted L-one penalized estimation)

Plan

- Lokalizacja genów
- Wielokrotne testowanie
- Kryteria wyboru modelu
- Zmodyfikowane wersje Bayesowskiego Kryterium Informacyjnego
- SLOPE (Sorted L-one penalized estimation)
- Symulacje komputerowe

Struktura DNA



Zmienność genetyczna w populacjach ludzkich

- Około 99,9% informacji genetycznej jest dokładnie taka sama u wszystkich ludzi.
- ***Polimorfizm*** to różnica w strukturze DNA, która występuje u co najmniej 1% populacji.
- ***Polimorfizm Pojedynczego Nukelotydu (Single Nucleotide Polymorphism, SNP)*** - polimorfizm w pojedynczej bazie nukleotydowej:
  - Typowy SNP: pozycja w której
    - 85% populacji ma Cytozynę (C)
    - 15% ma Tyminę (T).
- Zwykle w danym lokusie występuję tylko dwie formy SNPa
- trzy genotypy : AA, Aa, aa.

Główny cel

GŁÓWNY CEL: identyfikacja mutacji które wpływają na badaną cechę.

Główny cel

GŁÓWNY CEL: identyfikacja mutacji które wpływają na badaną cechę.

Przykład - identyfikacja pacjentów korzystnie reagujących na terapię

Główny cel

GŁÓWNY CEL: identyfikacja mutacji które wpływają na badaną cechę.

Przykład - identyfikacja pacjentów korzystnie reagujących na terapię

Y - cecha ilościowa

Główny cel

GŁÓWNY CEL: identyfikacja mutacji które wpływają na badaną cechę.

Przykład - identyfikacja pacjentów korzystnie reagujących na terapię

Y - cecha ilościowa

Przykłady: zmiana ciśnienia krwi, poziomu cholesterolu

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cechy dla n pacjentów

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cechy dla n pacjentów

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - wskaźnik terapii

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cechy dla n pacjentów

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - wskaźnik terapii

$G_{n \times m}$ - macierz genotypów

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cechy dla n pacjentów

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - wskaźnik terapii

$G_{n \times m}$ - macierz genotypów

Zwykle $n \approx k \times 100$ lub $k \times 1000$, $m \approx k \times 100,000$

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cechy dla n pacjentów

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - wskaźnik terapii

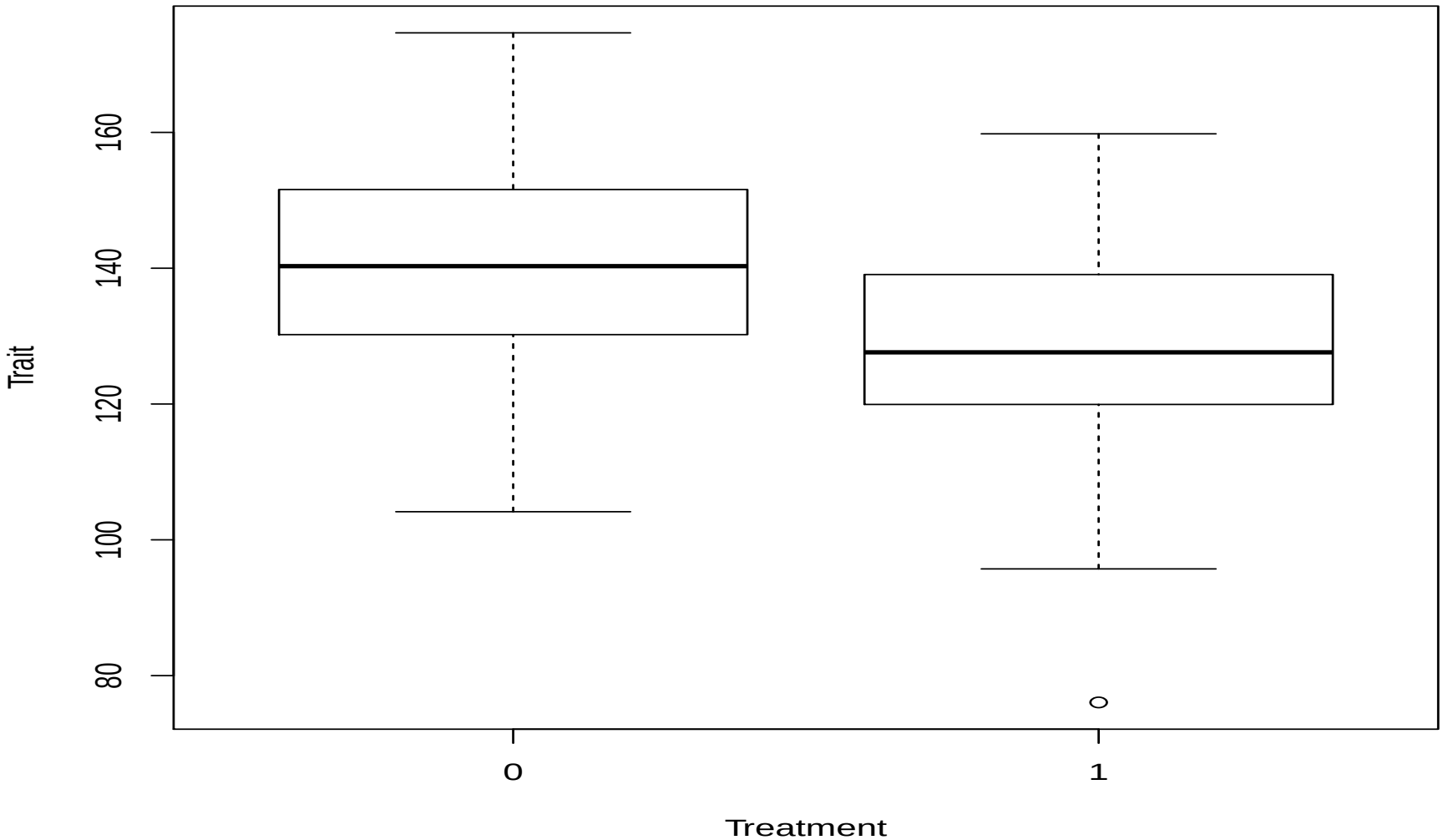
$G_{n \times m}$ - macierz genotypów

Zwykle $n \approx k \times 100$ lub $k \times 1000$, $m \approx k \times 100,000$

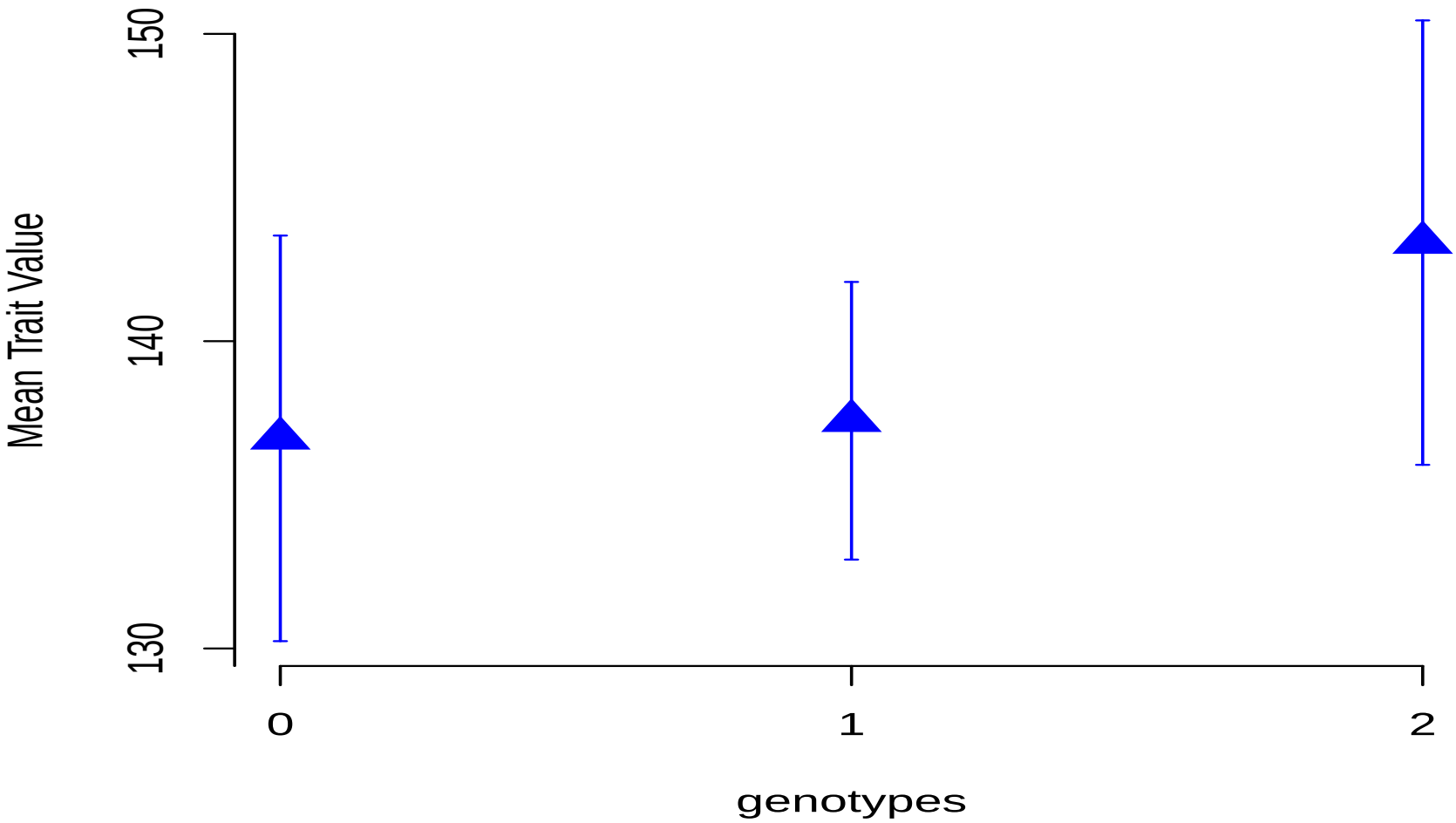
Typowe kodowanie oddziaływań addytywnych

$$Z_{ij} = \begin{cases} 0 & \text{gdy} \quad G_{ij} = AA \\ 1 & \text{gdy} \quad G_{ij} = Aa \\ 2 & \text{gdy} \quad G_{ij} = aa \end{cases}$$

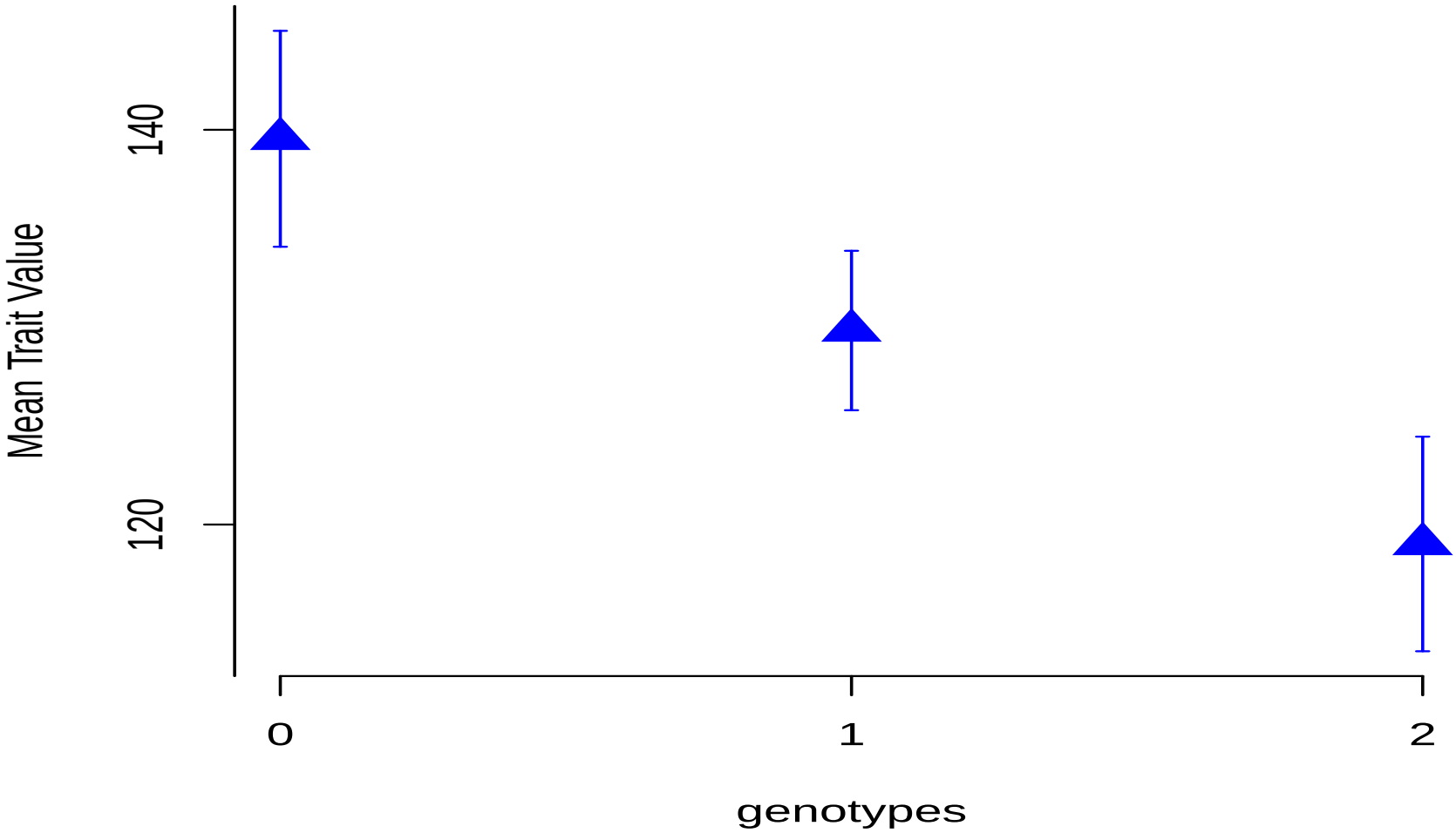
Fikcyjny przykład - obniżenie ciśnienia krwi



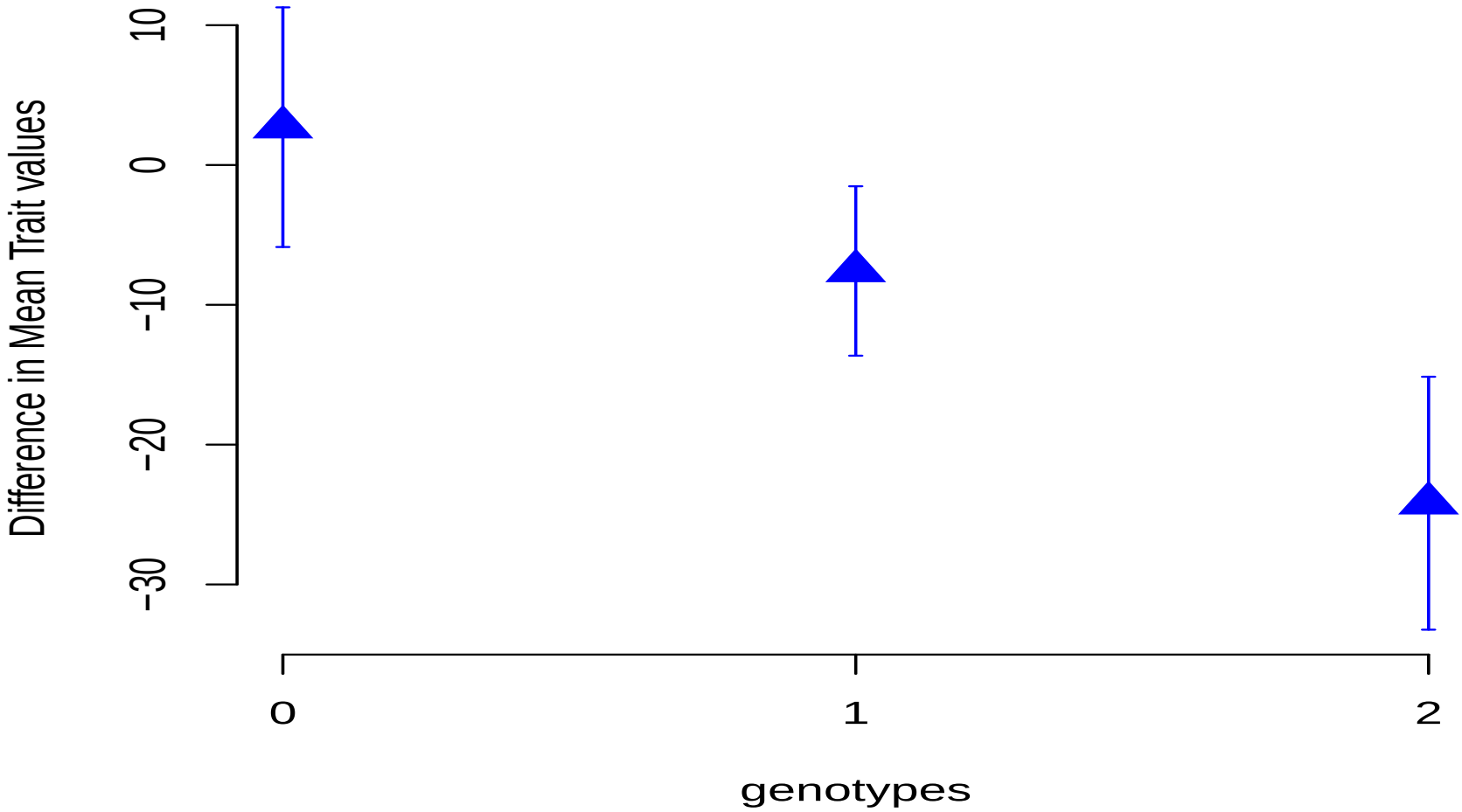
Grupa kontrolna



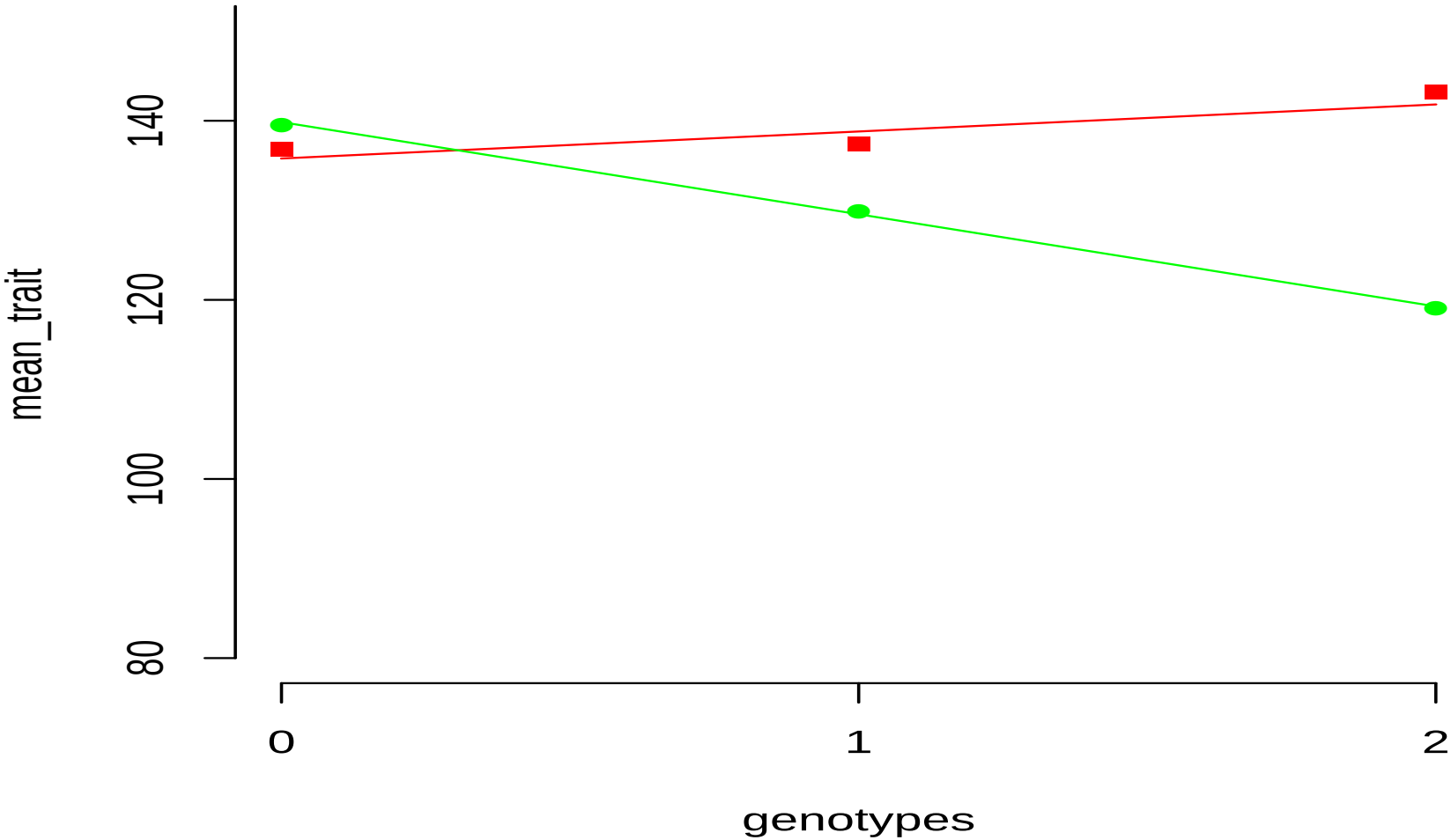
Grupa zabiegowa



Wpływ leku



Gene-Treatment Interaction



Model statystyczny

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

Model statystyczny

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

Model statystyczny

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

$$p = m + 2, \quad X = [1|T|Z|ZT] \quad , \beta = [\beta_0, \beta_1, \nu, \gamma]^T$$

Model statystyczny

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

$$p = m + 2, \quad X = [1|T|Z|ZT] \quad , \beta = [\beta_0, \beta_1, \nu, \gamma]^T$$

a) Identyfikacja istotnych genów - wybór modelu statystycznego

b) Predykcja: przewidywanie indywidualnej odpowiedzi na terapę

Prosta regresja liniowa

$$Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

Prosta regresja liniowa

$$Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

$\hat{\beta}_j$: estymator liczony metodą najmniejszych kwadratów β_j

$$\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	p_0
H_0 fałszywa	T	S	p_1
	W	R	p

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	p_0
H_0 fałszywa	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	p_0
H_0 fałszywa	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

Wielokrotne testowanie (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Odrzucamy H_{0j} gdy $z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$

Poziom istotności: $\alpha = P_{H_{0j}}(|z_j| > c)$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	p_0
H_0 fałszywa	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 2000 \rightarrow E(V) = 100$$

Procedury wielokrotnego testowania

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{p}$.

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{p}$.

$$FWER = P\left(\bigcup_{j=1}^{p_0}\{|z_j| > c\}\right) \leq \sum_{j=1}^{p_0} P(\{|z_j| > c\} = p_0\alpha/p < \alpha$$

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{p}$.

$$FWER = P\left(\bigcup_{j=1}^{p_0}\{|z_j| > c\}\right) \leq \sum_{j=1}^{p_0} P(\{|z_j| > c\} = p_0\alpha/p < \alpha$$

Odrzucamy H_{0j} gdy $|z_j| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2\log p}(1 + o_p)$

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{p}$.

$$FWER = P\left(\bigcup_{j=1}^{p_0}\{|z_j| > c\}\right) \leq \sum_{j=1}^{p_0} P(\{|z_j| > c\} = p_0\alpha/p < \alpha$$

Odrzucamy H_{0j} gdy $|z_j| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2\log p}(1 + o_p)$

Procedura Benjaminiego-Hochberga

(1) $|z|_{(1)} \geq |z|_{(2)} \geq \ldots \geq |z|_{(p)}$

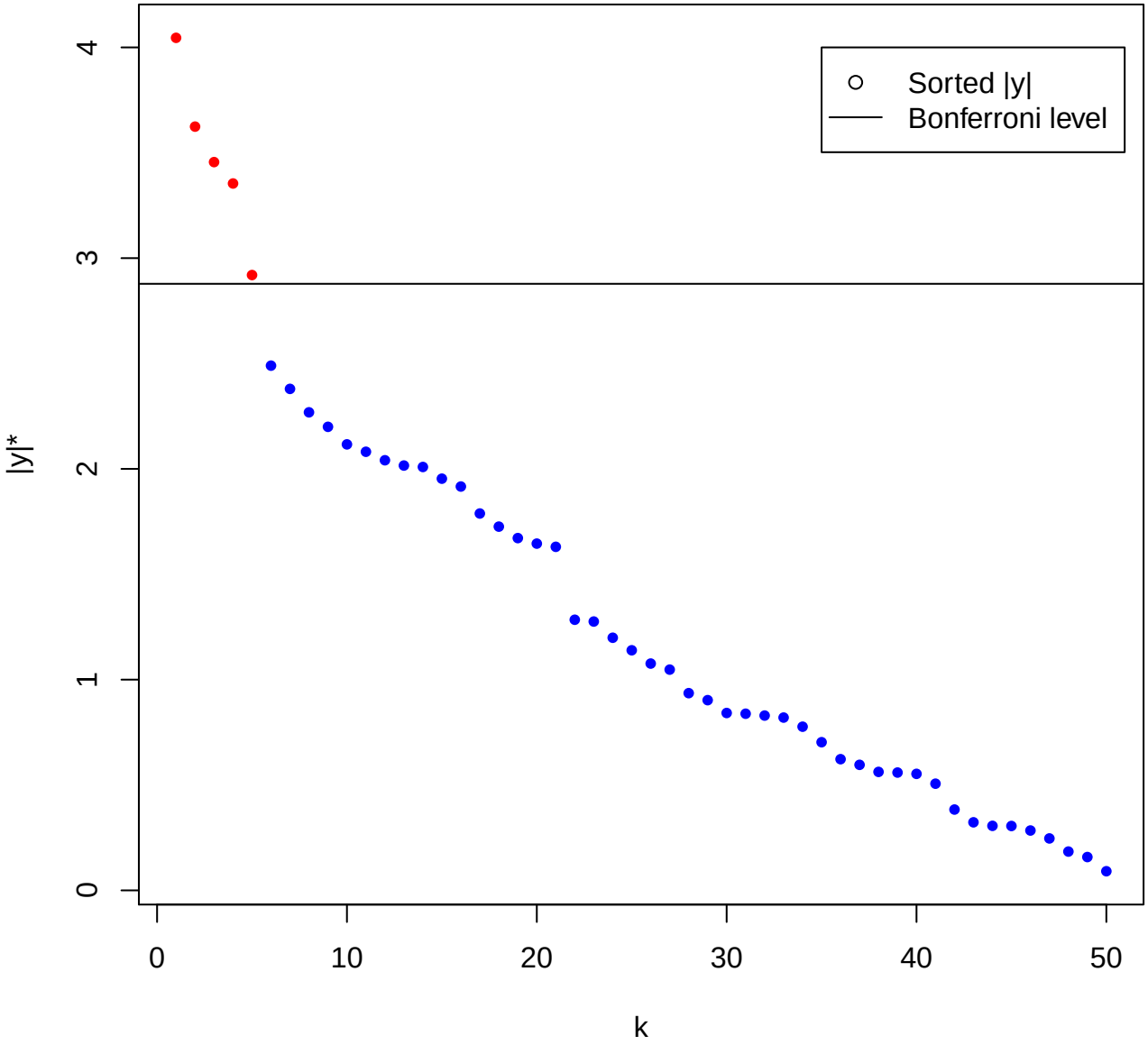
(2) Ustal największe j takie, że

$$|z|_{(j)} \geq \Phi^{-1}(1 - \alpha_j), \quad \alpha_j = \alpha\frac{j}{2p}, \tag{1}$$

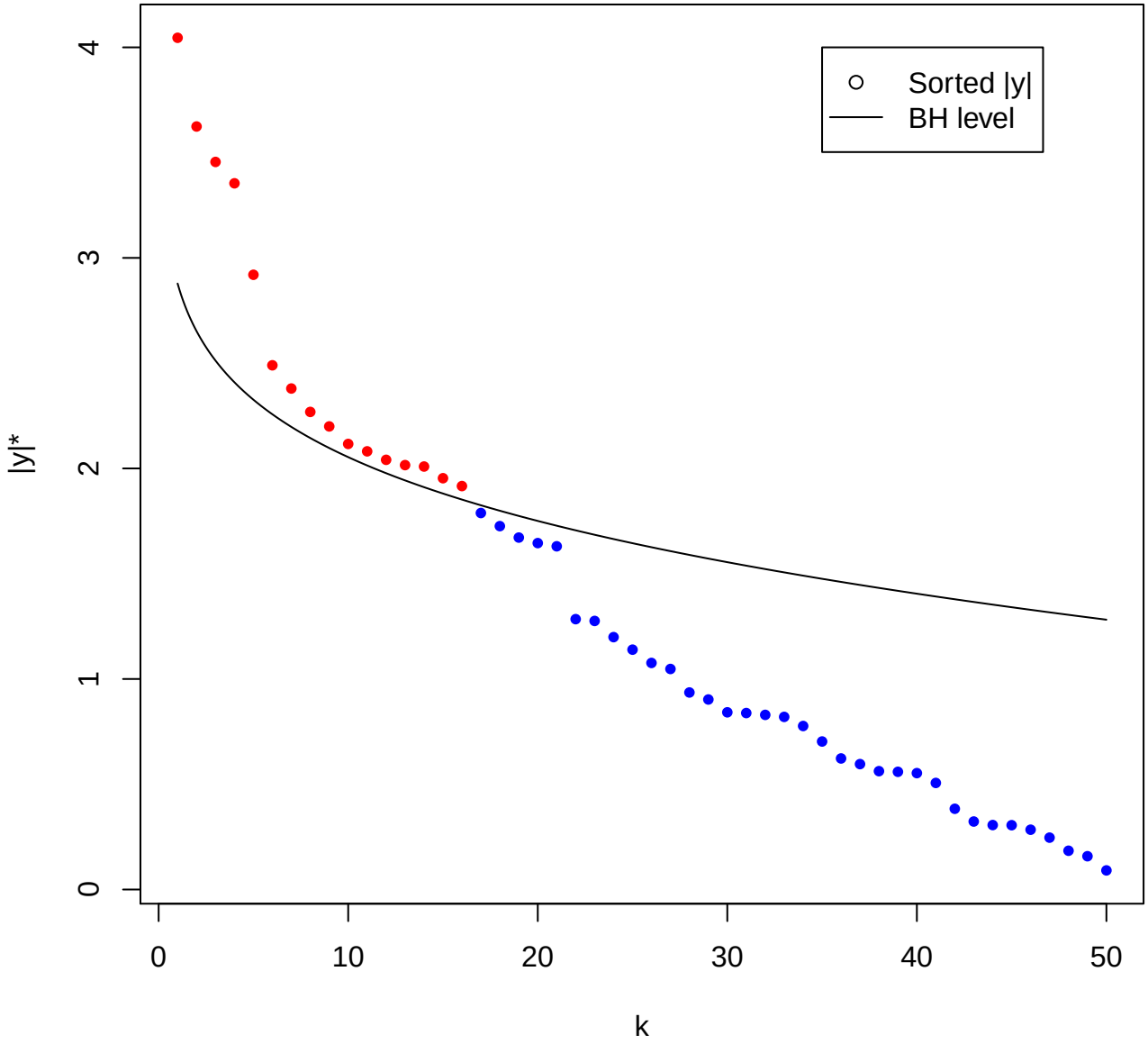
Nazwij ten indeks j_{SU}.

(3) Odrzuć $H_{(j)}$ wtedy i tylko wtedu gdy $j \leq j_{\mathsf{SU}}$

Korekta Bonferroniego



Korekta Benjaminiego-Hochberga



Kontrola FDR

(Benjamini,Hochberg, JRSSB 1995) Jeżeli $z_1, \ldots, z_p$ są niezależne to BH kontroluje FDR na poziomie

Kontrola FDR

(Benjamini,Hochberg, JRSSB 1995) Jeżeli $z_1, \ldots, z_p$ są niezależne to BH kontroluje FDR na poziomie

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}. \tag{2}$$

(Benjamini,Hochberg, JRSSB 1995) Jeżeli $z_1, \ldots, z_p$ są niezależne to BH kontroluje FDR na poziomie

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}. \tag{2}$$

(Benjamini, Yekutieli, Ann. Statist. 2001) Jeżeli statystyki testowe są "dodatnio skorelowane"to BH kontroluje FDR na poziomie $\alpha \frac{p_0}{p}$. Niezależnie od stopnia i rodzaju zależności między statystykami testowymi FDR jest kontrolowane jeżeli $|z|_{(j)}$ porównujemy z $\Phi^{-1}\left(1 - \frac{j\alpha}{2p\sum_{i=1}^{p}\frac{1}{i}}\right)$.

Asymptotyczna optymalność gdy testy są niezależne

$\tau = \frac{p-p_0}{p}$ - procent hipotez alternatywnych (rzadkość)
$\tau \to 0$ gdy $p \to \infty$

Asymptotyczna optymalność gdy testy są niezależne

$\tau = \frac{p-p_0}{p}$ - procent hipotez alternatywnych (rzadkość)
$\tau \to 0$ gdy $p \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006
- estymacja β z wykorzystaniem BH jest asymptotycznie optymalna (minimalizacja $||\hat{\beta} - \beta||$)

Asymptotyczna optymalność gdy testy są niezależne

$\tau = \frac{p-p_0}{p}$ - procent hipotez alternatywnych (rzadkość)
$\tau \to 0$ gdy $p \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - estymacja β z wykorzystaniem BH jest asymptotycznie optymalna (minimalizacja $||\hat{\beta} - \beta||$)

Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - minimalizacja średniego kosztu, γ_0 - koszt błędu I rodzaju (fałszywego odkrycia), γ_A - koszt błędu II rodzaju (pominięcie istotnego składnika)

Asymptotyczna optymalność gdy testy są niezależne

$\tau = \frac{p-p_0}{p}$ - procent hipotez alternatywnych (rzadkość)
$\tau \to 0$ gdy $p \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - estymacja β z wykorzystaniem BH jest asymptotycznie optymalna (minimalizacja $||\hat{\beta} - \beta||$)

Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - minimalizacja średniego kosztu, γ_0 - koszt błędu I rodzaju (fałszywego odkrycia), γ_A - koszt błędu II rodzaju (pominięcie istotnego składnika)

Korekta Bonferroniego jest asymptotycznie optymalna gdy $\tau \propto \frac{1}{p}$ ($p - p_0 = const$,ekstremalna rzadkość)

Asymptotyczna optymalność gdy testy są niezależne

$\tau = \frac{p-p_0}{p}$ - procent hipotez alternatywnych (rzadkość)
$\tau \to 0$ gdy $p \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - estymacja β z wykorzystaniem BH jest asymptotycznie optymalna (minimalizacja $||\hat{\beta} - \beta||$)

Bogdan, Chakrabarti, Frommlet, Ghosh, Ann.Statist. 2011 - minimalizacja średniego kosztu, γ_0 - koszt błędu I rodzaju (fałszywego odkrycia), γ_A - koszt błędu II rodzaju (pominięcie istotnego składnika)

Korekta Bonferroniego jest asymptotycznie optymalna gdy $\tau \propto \frac{1}{p}$ ($p - p_0 = const$,ekstremalna rzadkość)

BH jest asymptotycznie optymalna gdy $\tau \to 0$ i $\tau p \to C \in (0, \infty]$

Badania symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osobników pochodzenia europejskiego

Badania symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osobników pochodzenia europejskiego
- k = 40 przyczynowych niezależnych SNPów

Badania symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

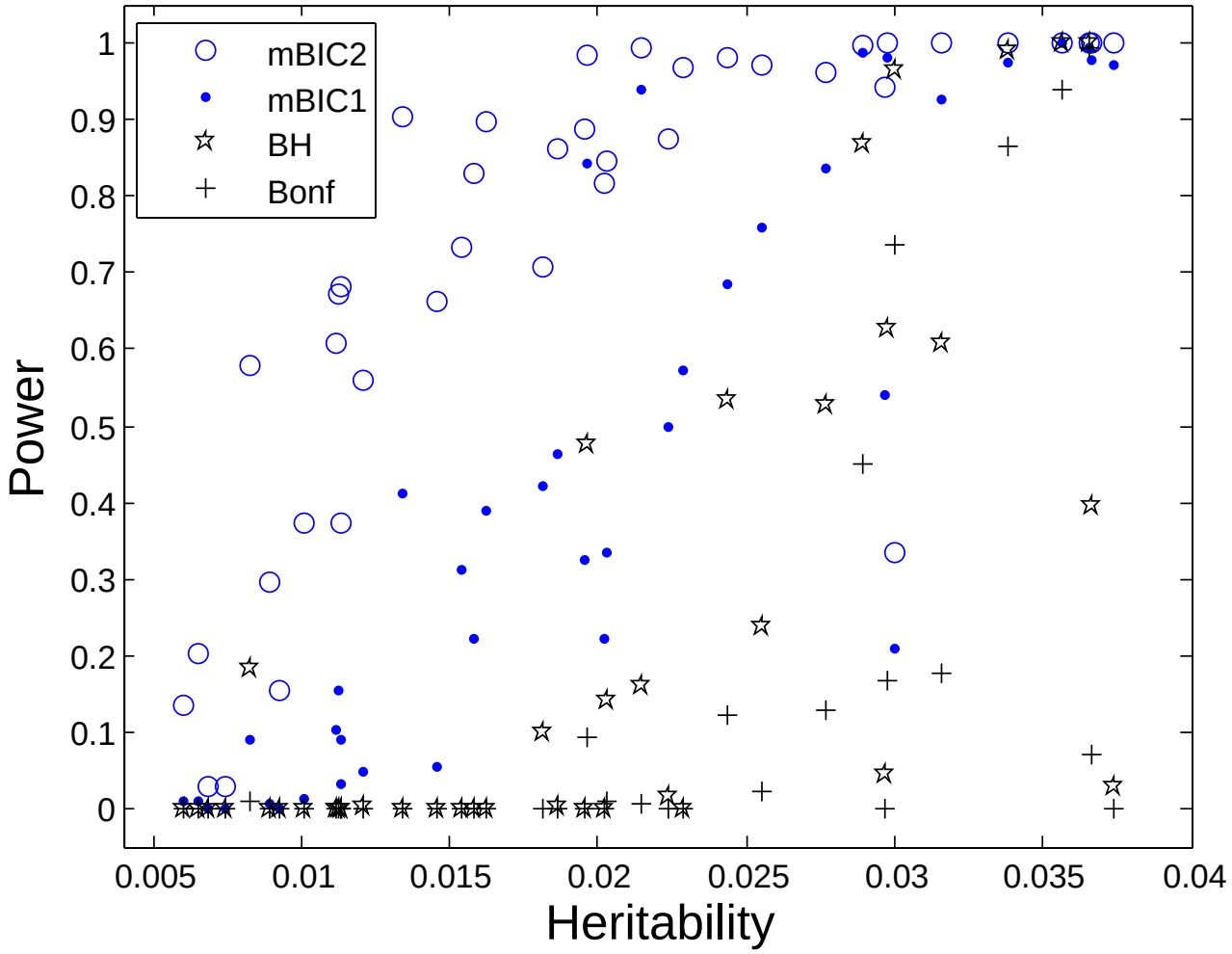
- 309790 SNPów dla 649 osobników pochodzenia europejskiego
- k = 40 przyczynowych niezależnych SNPów
- 1000 replikacji z modelu addytywnego M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$

Badania symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osobników pochodzenia europejskiego
- k = 40 przyczynowych niezależnych SNPów
- 1000 replikacji z modelu addytywnego M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$
- β_j równomiernie rozłożone na odcinku [0.27, 0.66]

Power


mBIC2
mBIC1
BH
Bonf
Power
Heritability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$\hat{C}ov(Y, X_1) = \beta_1 \hat{V}arX_1 + \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) + \hat{C}ov(X_1, \epsilon)$$

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$\hat{C}ov(Y, X_1) = \beta_1 \hat{V}arX_1 + \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) + \hat{C}ov(X_1, \epsilon)$$

Załóżmy, że dla $i > 1$, $\hat{C}ov(X_1, X_i) \sim N(0, \sigma_c^2)$

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$\hat{C}ov(Y, X_1) = \beta_1 \hat{V}arX_1 + \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) + \hat{C}ov(X_1, \epsilon)$$

Załóżmy, że dla $i > 1$, $\hat{C}ov(X_1, X_i) \sim N(0, \sigma_c^2)$

$$E \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) = 0$$

Problem z testami pojedynczymi

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$\hat{C}ov(Y, X_1) = \beta_1 \hat{V}arX_1 + \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) + \hat{C}ov(X_1, \epsilon)$$

Załóżmy, że dla $i > 1$, $\hat{C}ov(X_1, X_i) \sim N(0, \sigma_c^2)$

$$E \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) = 0$$

$$Var(\sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i)) \approx \sum_{i=2}^{k} \beta_i^2 \sigma_c^2$$

Informacyjne kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

$$Y = X_{n\times p}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n}), \quad p >> n$$

Informacyjne kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

$$Y = X_{n\times p}\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I_{n\times n}), \;\; p >> n$$

Zadanie wykonalne przy założeniu,że liczba niezerowych elementów $||\beta||_0 = k << n$ (założenie rzadkości)

Informacyjne kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

$$Y = X_{n\times p}\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I_{n\times n}), \;\; p >> n$$

Zadanie wykonalne przy założeniu,że liczba niezerowych elementów $||\beta||_0 = k << n$ (założenie rzadkości)

Kryteria wyboru modelu: minimalizujemy $||Y - X\beta||^2 + pen(k)$

Informacyjne kryteria wyboru modelu (2)

Bayesowskie Kryterium Informacyjne, BIC : $pen(k) = \sigma^2 k \log n$

Informacyjne kryteria wyboru modelu (2)

Bayesowskie Kryterium Informacyjne, BIC : $pen(k) = \sigma^2 k \log n$

BIC nie jest zgodne gdy $\frac{p}{\sqrt{n}} \to \infty$ (Bogdan et al. (2008, QREI)).

Informacyjne kryteria wyboru modelu (2)

Bayesowskie Kryterium Informacyjne, BIC : $pen(k) = \sigma^2 k \log n$

BIC nie jest zgodne gdy $\frac{p}{\sqrt{n}} \to \infty$ (Bogdan et al. (2008, QREI)).

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

Informacyjne kryteria wyboru modelu (2)

Bayesowskie Kryterium Informacyjne, BIC : $pen(k) = \sigma^2 k \log n$

BIC nie jest zgodne gdy $\frac{p}{\sqrt{n}} \to \infty$ (Bogdan et al. (2008, QREI)).

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

BHRIC (Abramovich et al. (2006), Foster and Stine (1999), Birge and Massart (2001))

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i)$ - "korekta BH"

Informacyjne kryteria wyboru modelu (2)

Bayesowskie Kryterium Informacyjne, BIC : $pen(k) = \sigma^2 k \log n$

BIC nie jest zgodne gdy $\frac{p}{\sqrt{n}} \to \infty$ (Bogdan et al. (2008, QREI)).

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

BHRIC (Abramovich et al. (2006), Foster and Stine (1999), Birge and Massart (2001))

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i)$ - "korekta BH"

Bogdan et al. (Genetics, 2004), Żak-Szatkowska and Bogdan (CSDA,2011)

mBIC1, mBIC2 - RIC i mRIC + kara $\sigma^2 k \log n$

Wypukła relaksacja - LASSO, (Tibshirani, JRSSB 1995)

- LASSO to rozwiązanie problemu optymalizacji wypukłej

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

gdzie $\lambda_L > 0$ jest parametrem wygładzającym

Wypukła relaksacja - LASSO, (Tibshirani, JRSSB 1995)

- LASSO to rozwiązanie problemu optymalizacji wypukłej

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

  gdzie $\lambda_L > 0$ jest parametrem wygładzającym
- Trudność - wybór λ_L

Wypukła relaksacja - LASSO, (Tibshirani, JRSSB 1995)

- LASSO to rozwiązanie problemu optymalizacji wypukłej

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

  gdzie $\lambda_L > 0$ jest parametrem wygładzającym
- Trudność - wybór λ_L
- $\lambda_L = \frac{\sqrt{2 \log p}}{n}$ - korekta Bonferroniego

Wypukła relaksacja - LASSO, (Tibshirani, JRSSB 1995)

- LASSO to rozwiązanie problemu optymalizacji wypukłej

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

  gdzie $\lambda_L > 0$ jest parametrem wygładzającym
- Trudność - wybór λ_L
- $\lambda_L = \frac{\sqrt{2 \log p}}{n}$ - korekta Bonferroniego
- walidacja krzyżowa - optymalizacja własności predykcyjnych

SLOPE (Sorted L-One Penalized Estimation, Bogdan, van den Berg, Sabatti, Su and Candés (AOAS, 2015))

- SLOPE to rozwiązanie problemu wypukłego

$$argmin_b \ \frac{1}{2}\|y - Xb\|_2^2 + \sigma \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \qquad \text{(SLOPE)}$$

gdzie $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ i $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

pozwala na kontrolę FDR na poziomie qp_0/p.

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.
- Interesujące własności predykcyjne

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.
- Interesujące własności predykcyjne
- Ogólnodostępne pakiety na CRAN

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.
- Interesujące własności predykcyjne
- Ogólnodostępne pakiety na CRAN
  i) *SLOPE* - Bogdan et al. 2015, autor E. Patterson

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.
- Interesujące własności predykcyjne
- Ogólnodostępne pakiety na CRAN
  i) *SLOPE* - Bogdan et al. 2015, autor E. Patterson
  ii) *grpSLOPE* - grupowe SLOPE (wybór grup predykatorów), [Brzyski et al. (arXiv, 2015)], autor - A. Gossman

Kontrola FDR

- Gdy $X_i^\top X_j = 0$ dla $i \neq j$, to ciąg

$$\lambda_i := \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

  pozwala na kontrolę FDR na poziomie qp_0/p.
- Heurystyczne procedury wyboru ciągu λ kontrolującego FDR gdy regresory są niezależnymi zmiennymi losowymi.
- Interesujące własności predykcyjne
- Ogólnodostępne pakiety na CRAN
  i) *SLOPE* - Bogdan et al. 2015, autor E. Patterson
  ii) *grpSLOPE* - grupowe SLOPE (wybór grup predykatorów), [Brzyski et al. (arXiv, 2015)], autor - A. Gossman
  iii) *geneSLOPE* - aplikacja do badań asocjacyjnych, [Brzyski et al. (przyjęte do druku w *Genetics*)], autor - P. Sobczyk

Symulacje w kontekście genetycznym

$n = p = 5000$, niezależne SNPy

Symulacje w kontekście genetycznym

$n = p = 5000$, niezależne SNPy

Scenariusz 1: $Y = X\beta + z$ - idealny model

Symulacje w kontekście genetycznym

$n = p = 5000$, niezależne SNPy

Scenariusz 1: $Y = X\beta + z$ - idealny model

Scenariusz 2: Nieliniowość - efekty dominancji:

$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases}, \tag{3}$$

$$y = [X, Z][\beta'_X, \beta'_Z]' + \epsilon$$

Szukamy tylko efektów addytywnych.

Symulacje w kontekście genetycznym

$n = p = 5000$, niezależne SNPy

Scenariusz 1: $Y = X\beta + z$ - idealny model

Scenariusz 2: Nieliniowość - efekty dominancji:

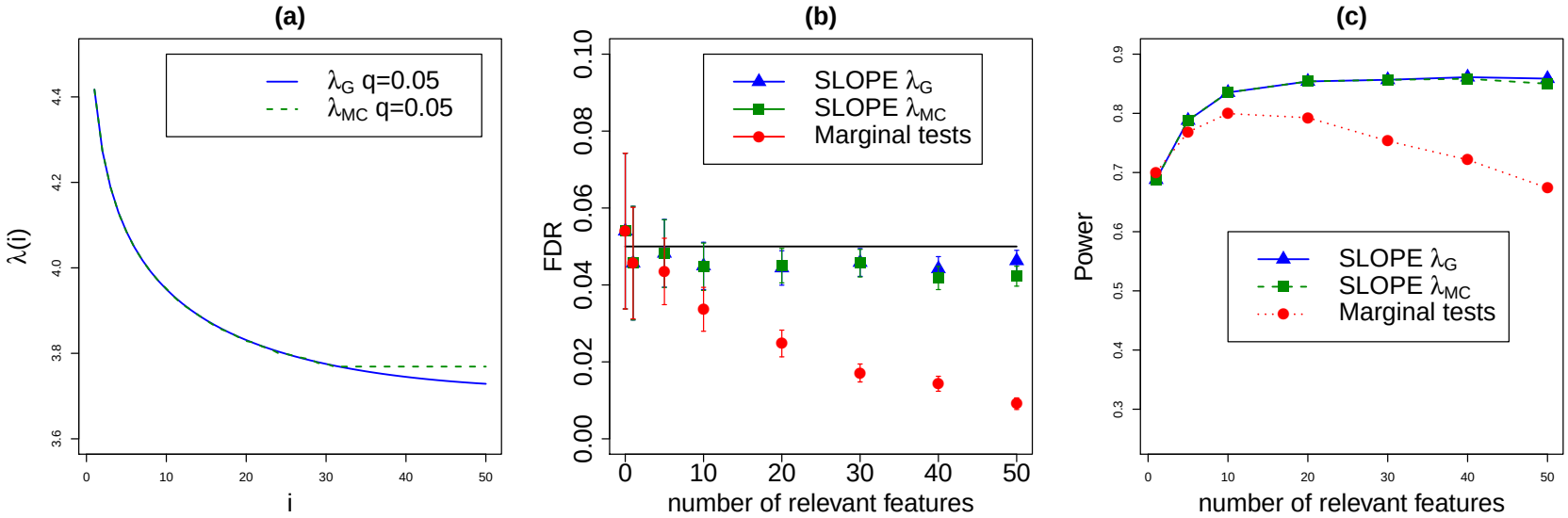
$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases}, \tag{3}$$

$$y = [X, Z][\beta'_X, \beta'_Z]' + \epsilon$$

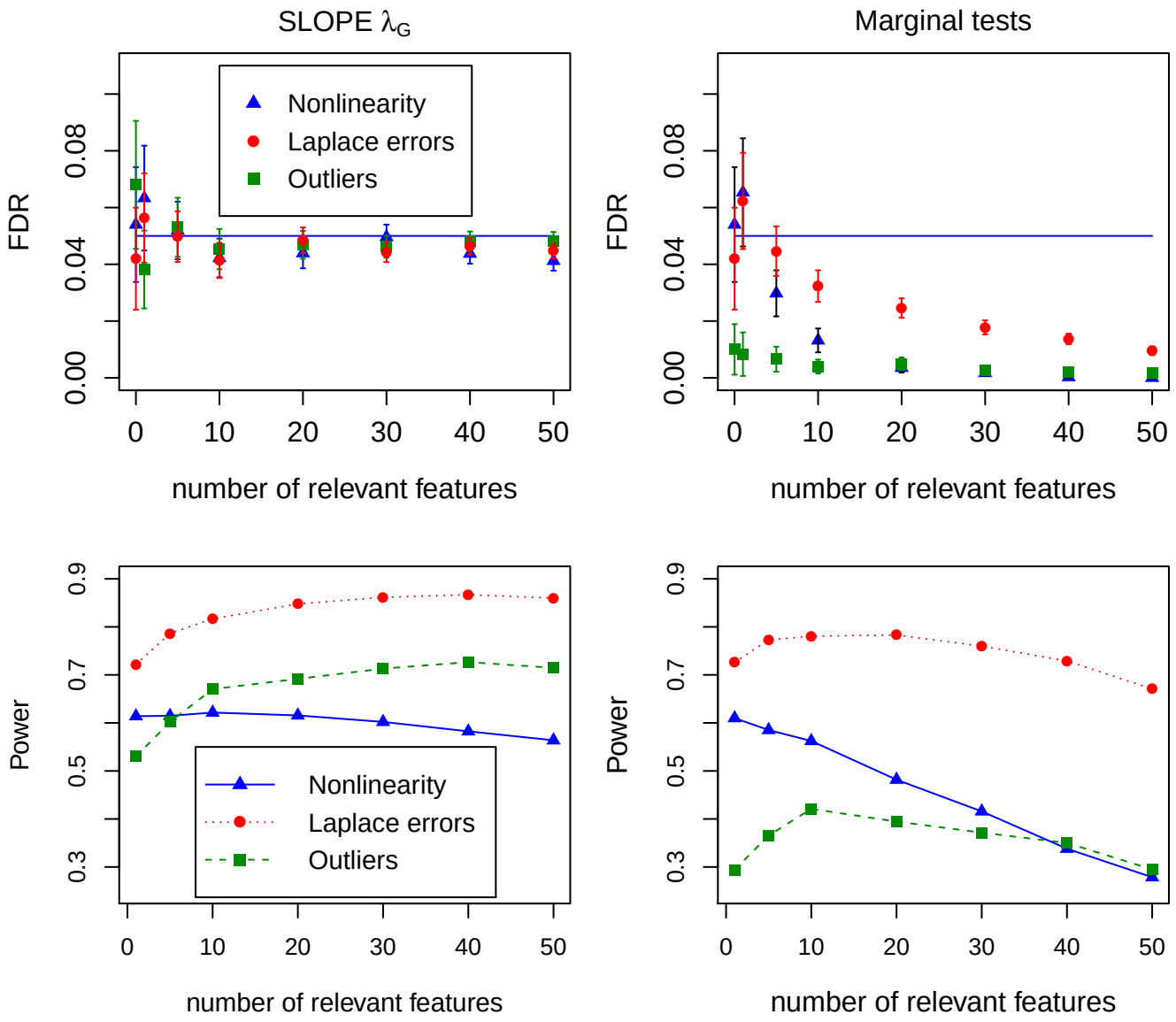
Szukamy tylko efektów addytywnych.

Scenariusze 3 i 4: Błędy z rozkładu Laplace'a lub z pewnym odsetkiem obserwacji odstających

Model idealny



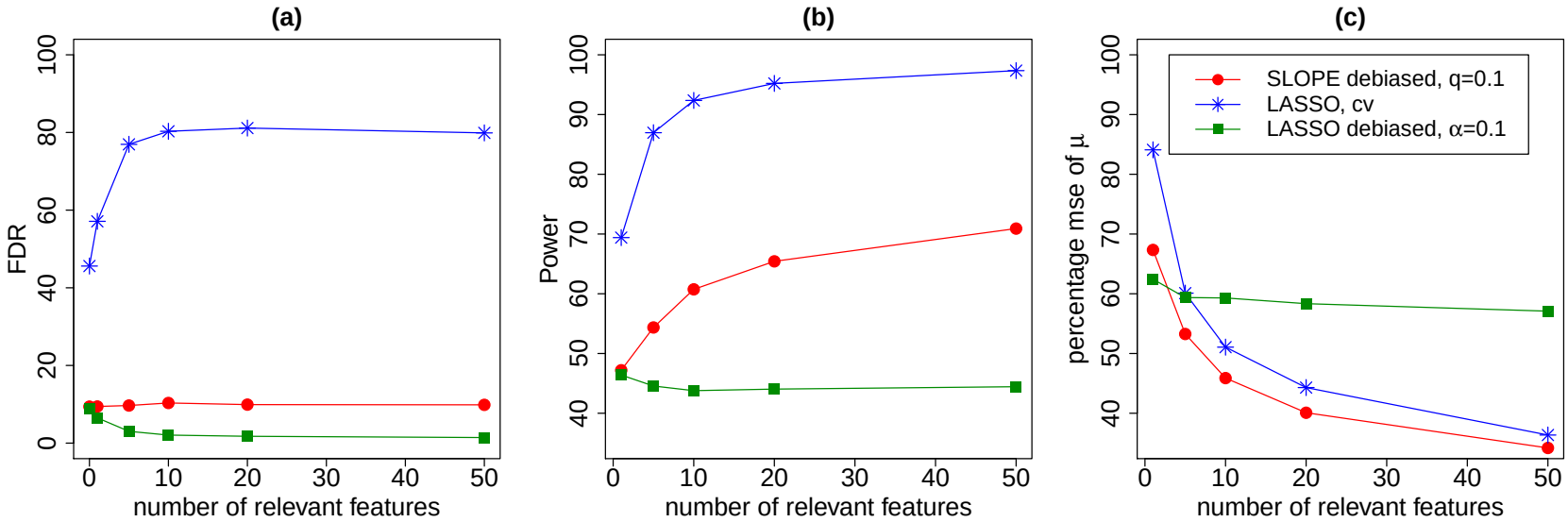
Odstępstwa od założeń



Własności predykcyjne SLOPE

1. Su and Candes (Ann. Statist., 2016) - asymptotyczna optymalność gdy zmienne są ortogonalne lub niezależne gausowskie
2. Nowak and Figuereido (AISTATS, 2016) - skupianie skorelowanych predykatorów i optymalna predykcja dla skorelowanych gaussowskich planów eksperymentu

Własności predykcyjne



Predykcja - ukryte zmienne

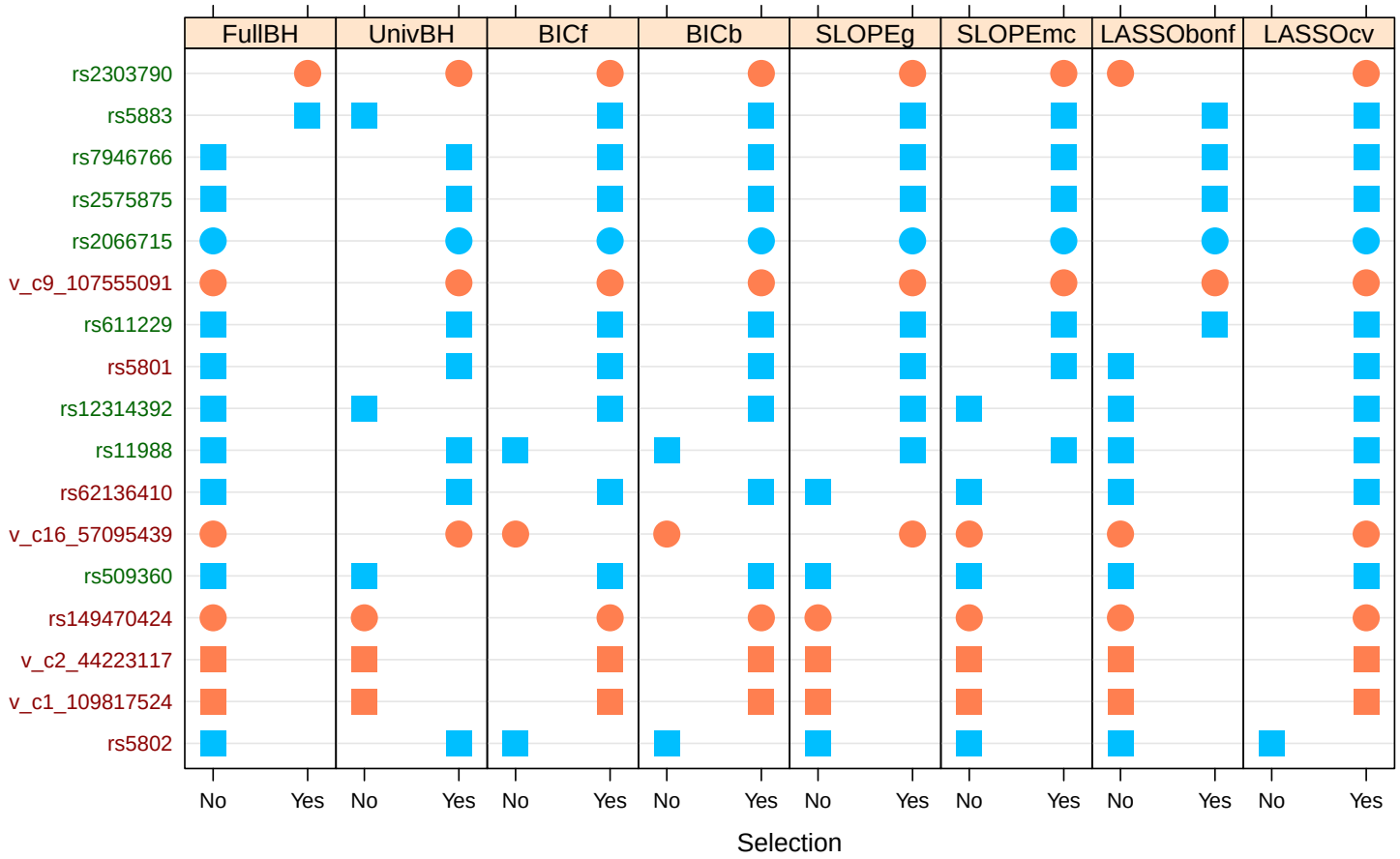
$$X_{n\times p} = F_{n\times 20} C_{20\times p} + \epsilon_1$$

$$Y = F\beta + \epsilon$$

Tablica : Procentowy błąd predykcji

p	n	adaptive lasso	SLOPE $q = 0.1$
1000	1000	0.209	0.257
2000	2000	0.251	0.242
2000	2000	0.346	0.240

Analiza danych rzeczywistych



Współpracownicy

1. Polska - M. Żak-Szatkowska, P. Biecek (UW), D. Brzyski (UJ), P. Sobczyk (PWr), P. Szulc (PWr)
2. Vienna University, Medical University of Vienna - A. Futschik, A. Baierl, F. Frommlet, F. Koenig
3. TU Muenchen - C. Czado, V. Erhart
4. Stanford - E.J. Candes, C. Sabatti, W. Su, E. Van den Berg, E. Patterson, C. Peterson
5. Tulane University - A. Gossman