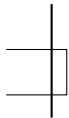


Pakiety statystyczne

Wykad 9

Tomasz Suchocki

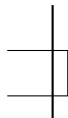
Uniwersytet Przyrodniczy we Wrocławiu
Katedra Genetyki



Plan wykładu

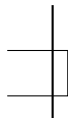
Model mieszany

1. Podstawy teoretyczne
2. Przykłady w R
3. Przykłady zastosowania



Model mieszany

- ▶ zmienna zależna jest typu ciągłego
- ▶ komplementarnie do efektów stałych możemy określić efekty losowe, które w modelu uwzględniamy jako realizacje zmiennej losowej o pewnych parametrach



Dlaczego używamy efektów losowych?

- ▶ Efektów tych może być tak dużo, że zamiast traktować je jako parametry modelu, przyjmujemy, że są to zmienne losowe z zadanego rozkładu prawdopodobieństwa, którego parametry będą parametrami modelu.
- ▶ Możemy nie być zainteresowani bezpośrednią oceną wartości efektów, lecz raczej zmiennością tych efektów w populacji.



Ogólna postać modelu:

$$y = X\beta + Zu + \epsilon,$$

Efekty stałe: $X\beta$

- ▶ X - macierz wystąpień dla efektów stałych ($n \times p$)
- ▶ β - parametry do estymacji ($p \times 1$)

Efekty losowe: Zu

- ▶ Z - macierz wystąpień dla efektów losowych ($n \times q$)
- ▶ u - wektor zmiennych losowych odpowiadających efektom losowym ($q \times 1$)



Residua:

- ▶ ϵ - błędy losowe ($n \times 1$)

Zakłada się, że wektory u i ϵ są niezależne i pochodzą z rozkładów normalnych:

- ▶ $u \sim \mathcal{N}(0, G)$, z macierzą $G = A\sigma_u^2$ o wymiarach ($q \times q$)
- ▶ $\epsilon \sim \mathcal{N}(0, R)$, z macierzą $R = I\sigma_\epsilon^2$ o wymiarach ($n \times n$)

Przy powyższych założeniach wariancja zmiennej y ma postać:

$$\text{Var}(y) = V = ZGZ^T + R$$



Estymatory efektów stałych β uzyskane przy pomocy metody największej wiarygodności:

- ▶ $y \sim \mathcal{N}(X\beta, V)$ z macierzą $V = ZGZ^T + R$
- ▶ Funkcja wiarygodności powyższego modelu mieszanego ma postać

$$-2 \log L = n \log(2\pi) + \log |V| + (y - X\beta)^T V^{-1} (y - X\beta)$$

- ▶ $\hat{\beta} = (X^T V^{-1} X)^{-1} \cdot X^T V^{-1} y$



Jak otrzymać funkcję największej wiarygodności? Wiemy, że jeżeli $z \sim \mathcal{N}(A, B)$ to gęstość prawdopodobieństwa zmiennej y ma postać

$$\frac{1}{(2\pi)^{\frac{n}{2}} |B|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (z - A)^T B^{-1} (z - A) \right)$$

Korzystając z tego faktu mamy i z tego, że $y \sim \mathcal{N}(X\beta, V)$ otrzymujemy

$$L = \frac{1}{(2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (y - X\beta)^T V^{-1} (y - X\beta) \right)$$



Logarytmując otrzymaną funkcję wiarygodności otrzymujemy

$$\log L = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |V| - \frac{1}{2} (y - X\beta)^T V^{-1} (y - X\beta)$$

Mnożąc otrzymany logarytm funkcji wiarygodności przez liczbę (-2) otrzymujemy ostateczną postać funkcji wiarygodności.



Jak otrzymać wzór na estymatory parametrów β metodą największej wiarygodności?

$$\begin{aligned} -2 \log L = & n \log(2\pi) + \log |V| + y^T V^{-1} y - y^T V^{-1} X \beta \\ & - \beta^T X^T V^{-1} y + \beta^T X^T V^{-1} X \beta \end{aligned}$$

Różniczkując powyższą funkcję względem parametru β i otrzymujemy

$$\begin{aligned} \frac{\partial(-2 \log L)}{\partial \beta} = & -y^T V^{-1} X - X^T V^{-1} y + 2X^T V^{-1} X \beta \\ = & -2X^T V^{-1} y + 2X^T V^{-1} X \beta \end{aligned}$$



Otrzymaną różniczkę funkcji wiarygodności przyrównujemy do 0 i otrzymujemy

$$\frac{\partial(-2 \log L)}{\partial \beta} = -2X^T V^{-1}y + 2X^T V^{-1}X\beta = 0$$

$$\hat{\beta} = \left(X^T V^{-1}X\right)^{-1} \cdot X^T V^{-1}y$$



FAKT:

$$\text{Var}(Ay) = A \cdot \text{Var}(y) \cdot A^T,$$

dla zmiennej losowej y i znanej macierzy A



Macierz wariancji-kowariancji dla efektów stałych:

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}\left(\left(X^T V^{-1} X\right)^{-1} \cdot X^T V^{-1} y\right) \\ &= \left(X^T V^{-1} X\right)^{-1} \left(X^T V^{-1} \text{Var}(y) V^{-1} X\right) \left(X^T V^{-1} X\right)^{-1} \end{aligned}$$

Podstawiając za $\text{Var}(y) = V$ otrzymujemy

$$\text{Var}(\hat{\beta}) = \left(X^T V^{-1} X\right)^{-1} \left(X^T V^{-1} V V^{-1} X\right) \left(X^T V^{-1} X\right)^{-1}$$

i ostatecznie mamy

$$\text{Var}(\hat{\beta}) = \left(X^T V^{-1} X\right)^{-1}$$



Testowanie istotności efektów stałych (Test Walda):

Dla każdego efektu β_j , $j = 1, \dots, p$ mamy $SE(\hat{\beta}_j) = \text{diag}(\text{Var}(\hat{\beta}))_j$

- ▶ Hipotezy $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
- ▶ Statystyka testowa $W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$
- ▶ Statystyka W przy prawdziwości H_0 ma rozkład $\mathcal{N}(0, 1)$



Testowanie istotności efektów stałych (Test Walda):

- ▶ Hipotezy $H_0 : L\beta = 0$ vs. $H_1 : L\beta \neq 0$, dla danej macierzy kontrastu L
- ▶ Statystyka testowa $W = \hat{\beta}^T L^T [L(X^T V^{-1} X)^{-1} L^T]^{-1} L \hat{\beta}$
- ▶ Statystyka W przy prawdziwości H_0 ma rozkład χ^2 z liczbą stopni swobody równą rzędowi macierzy L



Testowanie istotności efektów stałych (Test oparty na ilorazie wiarygodności - porównywanie modeli z różną ilością efektów stałych):

- ▶ Hipotezy $H_0 : \beta \in \Omega_0$, gdzie $\Omega_0 \in \Omega$
- ▶ Statystyka testowa $-2 \log \lambda_N = -2 \log \frac{L_{\Omega_0}}{L_{\Omega}}$
- ▶ Statystyka przy prawdziwości H_0 ma rozkład χ^2 z liczbą stopni swobody równą różnicy parametrów pomiędzy Ω i Ω_0



Estymacja parametrów wariancji:

$$-2 \log L_{REML} = -2 \log L + \log |X^T V^{-1} X|$$

Metody estymacji parametrów wariancji:

- ▶ EM
- ▶ Newton-Raphson
- ▶ Fisher scoring



Testowanie istotności efektów losowych (Test oparty na ilorazie wiarygodności):

- ▶ Hipotezy $H_0 : \sigma_u^2 = 0$
- ▶ Statystyka testowa
$$-2 \log \lambda_N = -2 \log L_{\text{zredukowanej}} + 2 \log L_{\text{penej}}$$
- ▶ Statystyka przy prawdziwości H_0 jest mieszaniną rozkładów χ^2 z 0 i 1 stopniem swobody ($\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$)



Porównywanie modeli:

- ▶ $AIC = -2 \log L + 2s$
- ▶ $BIC = -2 \log L + s \log(n^*)$

gdzie

- ▶ s - liczba wszystkich parametrów w modelu
- ▶ $n^* = n - p$ dla funkcji L_{REML}
- ▶ $n^* = n$ dla funkcji L



Predykcja efektów losowych:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

Warto przypomnieć sobie funkcje `cbind()` i `rbind()`



Przykład

Zbiór danych o nazwie "milkgeneż" pakietu "PBImisc". Kolejne kolumny zawierają:

- ▶ numer krowy
- ▶ numer laktacji
- ▶ gen BTM3a1
- ▶ wydajność mleczną
- ▶ wydajność tłuszczu



Chcemy sprawdzić, czy obserwowana mutacja w genie *BTN3a1* wpływa na mleczność krów. Mamy 915 pomiarów mleczności dla osobników o genotypie 1 i 85 pomiarów dla genotypu 2. Nie można zastosować w tym przypadku testu t-studenta dla dwóch grup, ani modelu z jednym efektem stałym genu ponieważ część osobników była mierzona więcej niż jeden raz (przez kilka laktacji).

Aby uwzględnić podobieństwo w pomiarach do modelu dołożymy efekt losowy osobnika. Wybieramy efekt losowy, ponieważ krów jest dużo i nie chcemy wprowadzać tak wielu parametrów do modelu. Dodatkowo nie interesuje nas efekt pojedynczych osobników, a efekt genu.



Przykład

Model:

$$y_{mleko} = X_{btn3a1}\beta_{btn3a1} + X_{laktacja}\beta_{laktacja} + Z_{krowa}u_{krowa} + \epsilon$$

- ▶ u_{krowa} - efekt losowy o rozkładzie $\mathcal{N}(0, \sigma_{krowa}^2)$
- ▶ macierz Z_{krowa} ma wymiary 1000×409 , $z_{i,j} = 1$ gdy wiersz i dotyczy j tego osobnika oraz 0 w przeciwnym razie.



Formuła:

- ▶ `require(lme4)`
- ▶ `m1=lmer(milk btn3a1+lactation+(1|cow.id),data=milkgene)`

Otrzymujemy wyniki:

- ▶ Efekt stały genu –244.07. O tyle średnio wyższa jest mleczność osobników o genotypie 2 od osobników o genotypie 1.
- ▶ Parametr $\sigma^2_{krowa} = 1240403$ i jest praktycznie równy wariancji błędu.

Należy sprawdzić założenia modelu tzn. Normalność reszt i efektów losowych oraz ich niezależność.



Istotność efektów stałych:

- ▶ `summary(m1)$coefs` - testy t-studenta
- ▶ test oparty na ilorazie wiarygodności tzn. tworzymy dwa modele z i bez efektu genu `btn3a1`, następnie obliczamy logarytmy funkcji wiarygodności dla każdego z modeli i wyznaczamy statystykę λ . Rozkład lambdy to χ^2_1 .

Istotność efektów losowych:

- ▶ test oparty na ilorazie wiarygodności tzn. tworzymy dwa modele z i bez efektu genu osobniczego, następnie obliczamy logarytmy funkcji wiarygodności dla każdego z modeli i wyznaczamy statystykę λ . Rozkład lambdy to $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$.



Przykład

Istotność efektów stałych:

- ▶ Efekt genu *BTN3a1* wynosi -1.03 z p-wartością $p = 0.299$

Istotność efektów losowych:

- ▶ Statystyka $\lambda = 243.92$ z p-wartością $p = 5.49e - 55$



Model:

$$y = X\beta + Zu + \epsilon$$

Predykcja efektów losowych:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

Warto przypomnieć sobie funkcje `cbind()` i `rbind()`



Przykładowy program

```
sigma_a=0.3 #starting value for additive polygenic variance
sigma_e=0.7 #starting value for error variance

nme=function(y,X,Z1,A,sigma_a,sigma_e) #Mixed model equations
{
  G1=ginv(A)*c(sigma_e)/c(sigma_a)
  C=rbind(cbind(t(X)%*%X,t(X)%*%Z1),
          cbind(t(Z1)%*%X,t(Z1)%*%Z1+G1)) #coefficients matrix
  r=rbind(t(X)%*%y,t(Z1)%*%y)
  nme=ginv(C)%*%r #solution for fixed and random effects
  list(C=C,nme=nme)
}
```



Estymatory parametrów wariancji uzyskane przy pomocy algorytmu EM:

$$\blacktriangleright \sigma_u^{2[t+1]} = \frac{\hat{\mathbf{u}}'^{[t]} \mathbf{A}^{-1} \hat{\mathbf{u}}^{[t]} + \text{tr}(\mathbf{A}^{-1} \mathbf{C}_u^{-1[t]}) \sigma_\epsilon^{2[t]}}{NA}$$

$$\blacktriangleright \sigma_\epsilon^{2[t+1]} = \frac{\hat{\boldsymbol{\epsilon}}'^{[t]} \hat{\boldsymbol{\epsilon}}^{[t]} + \text{tr}([\mathbf{X}, \mathbf{Z}] \mathbf{C}^{-1[t]} [\mathbf{X}, \mathbf{Z}]') \sigma_\epsilon^{2[t]}}{N}$$

Przykładowy program

```
EM=function(y,X,Z1,A,sigma_a,sigma_e) #EM estimators for variance components
{
  n=dim(X)[1] #number of animals
  n_X=dim(X)[2] #number of fixed effects

  t=1 #iteration No. 1
  tmp=0.1 #test of convergence
  print(t)

  while (tmp>0.001)
  {
    mme_new=mme(y,X,Z1,A,sigma_a,sigma_e)
    C_new=ginv(mme_new$C)
    Ca=C_new[(n_X+1):(n_X+n),(n_X+1):(n_X+n)]
    mme2=mme_new$mme

    a=as.matrix(mme2[(n_X+1):(n_X+n)])
    sigma_a_new=(t(a)%*%ginv(A)%*%a+
      sum(diag(ginv(A)%*%Ca))*c(sigma_e))/n
    res=as.matrix(y-X%*%as.matrix(mme2[1:n_X])-
      Z1%*%as.matrix(mme2[(n_X+1):(n_X+n)]))
    sigma_e_new=(t(res)%*%res+
      sum(diag(cbind(X,Z1)%*%C_new%*%t(cbind(X,Z1))))*c(sigma_e))/n
```



Przykładowy program

```
tmp=max(c(abs(sigma_a-sigma_a_new),abs(sigma_e-sigma_e_new))) #test of convergence
print(tmp)

sigma_a=sigma_a_new
sigma_e=sigma_e_new

t=t+1
print(t)
}
V=Z1%*%(A*c(sigma_a))%*%t(Z1)+diag(n)*c(sigma_e)
gV=ginv(V)
logL=n*log(2*pi)+log(det(V))+
  t(y-X%*%as.matrix(mme2[1:n_X]))%*%gV%*%(y-X%*%as.matrix(mme2[1:n_X]))+
  log(det(t(X)%*%gV%*%X))
write.table(rbind(sigma_a,0,sigma_e,logL,t),paste(path,"Results/",
  trait,"_H0.txt",sep=""),
  col.names=F,row.names=F,sep=";") #writing results to file
}
```



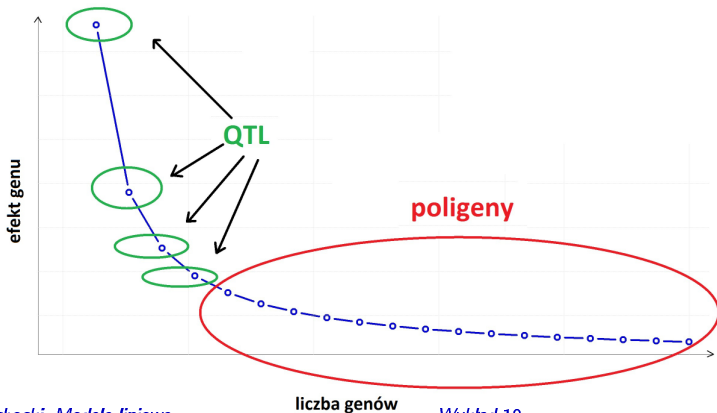
Modelowanie efektów genów zlokalizowanych na chromosomie szóstym *Bos taurus* jako zmiennych w czasie przy użyciu modeli mieszanych z regresjami losowymi

- ▶ **QTL** (ang. **Quantitative Trait Locus**) – region na chromosomie związany z siłą ekspresji danej cechy ilościowej (np. wydajność mleka u krów). Na jedną cechę może wpływać kilka QTL z różną siłą.



1. Zmienność genetyczna

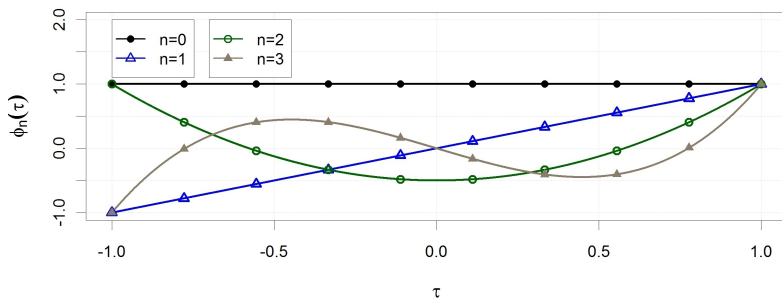
- ▶ wiele loci z małymi efektami (efekt poligenów)
- ▶ kilka loci z dużymi efektami (QTL)





2. Model mieszany z regresjami losowymi

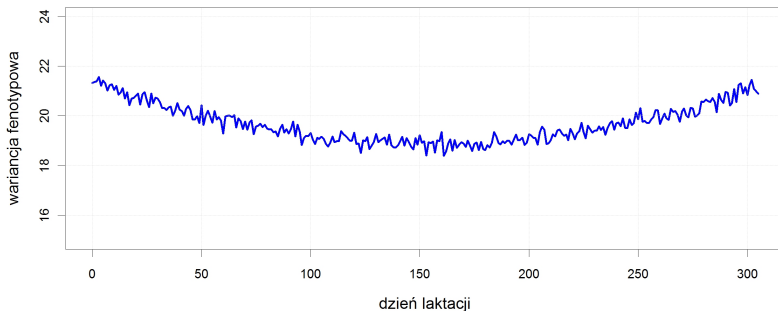
- ▶ wiele obserwacji dla każdego osobnika
- ▶ modelowanie struktury kowariancji na poziomie genetycznym
- ▶ efekty zmienne w czasie





3. Chromosom szósty *Bos taurus*

- ▶ obecność wielu QTL
- ▶ dotychczas efekt QTL stały w czasie
- ▶ brak mutacji funkcjonalnych





1. Sprawdzenie nowego modelu pod kątem przydatności do detekcji QTL.
2. Sprawdzenie, czy istotne QTL są stałe, czy zmienne w czasie.
3. Próba estymacji nowych pozycji QTL na chromosomie szóstym *Bos taurus*.



Zbiór danych - zwierzęta

- ▶ 716 krów rasy holsztyńsko-fryzyjskiej
 - ▶ 23 ojców
 - ▶ 9 stad
 - ▶ wiele obserwacji fenotypowych (6 497)
- ▶ rozważane cechy
 - ▶ dzienna wydajność mleka



źródło: <http://www.wikipedia.org>

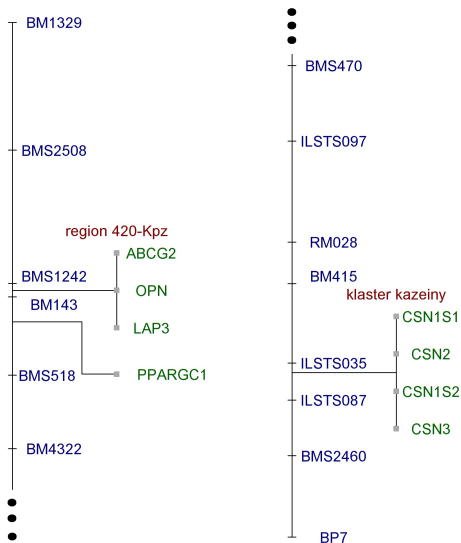


N	Marker	Położenie (cM)	Allele
1	BM1329	0.00	6
2	BMS2508	8.54	7
3	BMS1242	17.44	7
4	BM143	18.33	7
5	BMS518	23.57	4
6	BM4322	28.47	7
7	BMS470	32.00	4
8	ILSTS097	37.04	2
9	RM028	43.79	4
10	BM415	46.56	7
11	ILSTS035	51.87	9
12	ILSTS087	54.34	3
13	BMS2460	58.06	4
14	BP7	63.50	5

MARC map (Ihara et al., 2004)



Zbiór danych - chromosom 6





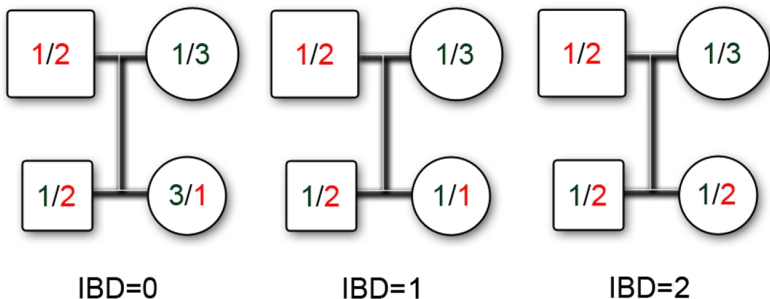
$$y_i = \mu + \beta + \alpha_i + \rho_i + \epsilon_i,$$

- ▶ y_i – skumulowana 305 – dniowa wartość cechy dla osobnika i
- ▶ β – efekty stałe
- ▶ α_i – losowy efekt addytywnie poligeniczny, $\alpha \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\sigma_\alpha^2)$
- ▶ ρ_i – losowy efekt QTL, $\rho \sim \mathcal{N}(\mathbf{0}, \mathbf{IBD}\sigma_\rho^2)$
- ▶ ϵ_i – residua, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N\sigma_\epsilon^2)$



Wykorzystane modele - macierz IBD

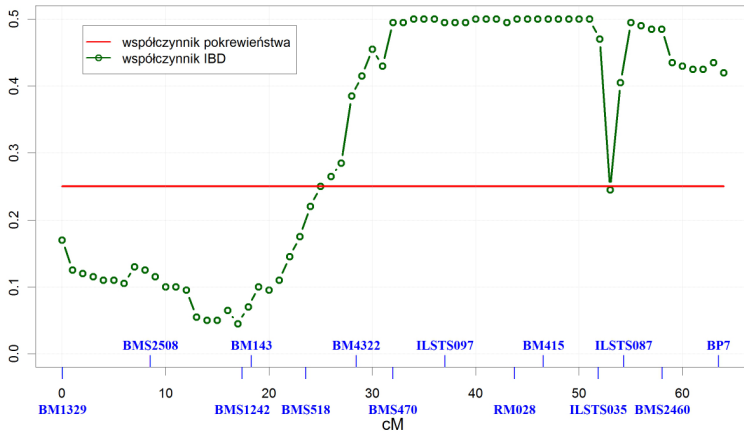
IBD (ang. **I**dentical **B**y **D**escent) – oszacowanie podobieństwa genetycznego pomiędzy osobnikami w danym miejscu genomu.



$$\vartheta = \frac{1}{2}P(\text{IBD} = 1) + P(\text{IBD} = 2)$$



Wykorzystane modele - macierz A a macierz IBD





$$y_{ij} = \beta + \sum_{m=0}^2 \alpha_{im} \phi_m(\tau_{ij}) + \rho_i + \sum_{m=0}^2 \xi_{im} \phi_m(\tau_{ij}) + \epsilon_{ij},$$

- ▶ y_{ij} – j ta obserwacja dla osobnika i
- ▶ $\phi_m(\tau_{ij})$ – m ty wielomian Legendre'a w czasie τ_{ij}
- ▶ α_{im} – m ty współczynnik regresji losowej dla efektu addytywnie poligenicznego, $\alpha \sim \mathcal{N}(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_\alpha)$
- ▶ ξ_{im} – m ty współczynnik regresji losowej dla efektu trwale środowiskowego, $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{NA} \otimes \mathbf{P}_\xi)$



$$y_{ij} = \beta + \sum_{m=0}^2 \alpha_{im} \phi_m(\tau_{ij}) + \sum_{m=0}^2 \rho_{im} \phi_m(\tau_{ij}) + \sum_{m=0}^2 \xi_{im} \phi_m(\tau_{ij}) + \epsilon_{ij},$$

- ▶ ρ_{im} – m -ty współczynnik regresji losowej dla efektu QTL,
 $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}, \mathbf{IBD} \otimes \mathbf{G}_{\rho})$



- ▶ Hipotezy dla modeli 1 i 2:

$$H_0 : \sigma_\rho^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_\rho^2 \neq 0$$

$$\Lambda \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

- ▶ Hipotezy dla modelu 3:

$$H_0 : \mathbf{G}_\rho = 0 \quad \text{vs.} \quad H_1 : \mathbf{G}_\rho \text{ jest macierzą nieujemnie} \\ \text{określoną o wymiarze } 3 \times 3$$

$$\Lambda \sim \chi_6^2$$



- Hipotezy do porównywania modeli 2 i 3:

$$H_0 : \mathbf{G}_\rho = \begin{pmatrix} \sigma_\rho^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{vs.} \quad H_1 : \mathbf{G}_\rho = \begin{pmatrix} \sigma_\rho^2 & \mathbf{G}_{\rho 12} \\ \mathbf{G}_{\rho 12}^T & \mathbf{G}_{\rho 22} \end{pmatrix}$$

$$\Lambda \sim \chi_2^2$$



Wykorzystane modele - ilość parametrów do estymacji

- ▶ Model 1 – 3 parametry ($\times 65 \text{ cM} \times 4 \text{ cechy} = 780$)

$$\sigma_{\alpha}^2, \sigma_{\rho}^2 \text{ i } \sigma_{\epsilon}^2$$

- ▶ Model 2 – 14 parametrów (3 640)

$$\mathbf{G}_{\alpha}, \sigma_{\rho}^2, \mathbf{P}_{\xi} \text{ i } \sigma_{\epsilon}^2$$

- ▶ Model 3 – 19 parametrów (4 940)

$$\mathbf{G}_{\alpha}, \mathbf{G}_{\rho}, \mathbf{P}_{\xi} \text{ i } \sigma_{\epsilon}^2$$

$\mathbf{G}_{\alpha}, \mathbf{G}_{\rho}, \mathbf{P}_{\xi}$ są wymiaru 3×3



1. ASReml

- ▶ estymacja parametrów wariancji

2. LOKI

- ▶ estymacja macierzy IBD

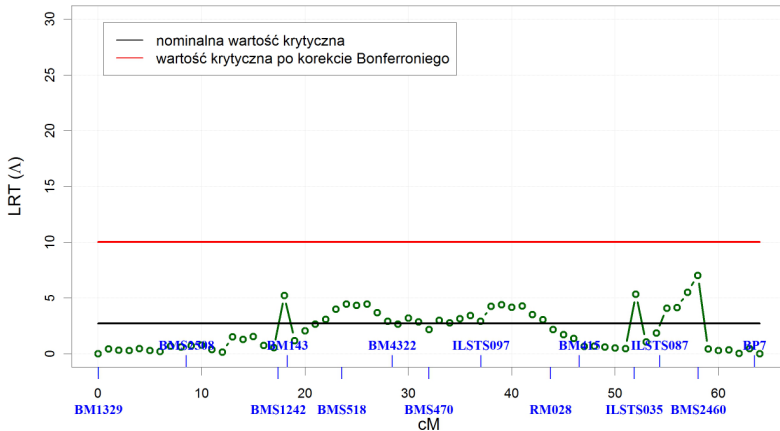
3. R - samodzielnie napisane programy do:

- ▶ wyznaczania macierzy spokrewnień A
- ▶ rozwiązywania układu równań modelu mieszanego
- ▶ wyznaczania wartości funkcji wiarygodności
- ▶ wyznaczania wartości statystyki testu opartego na ilorazie wiarygodności



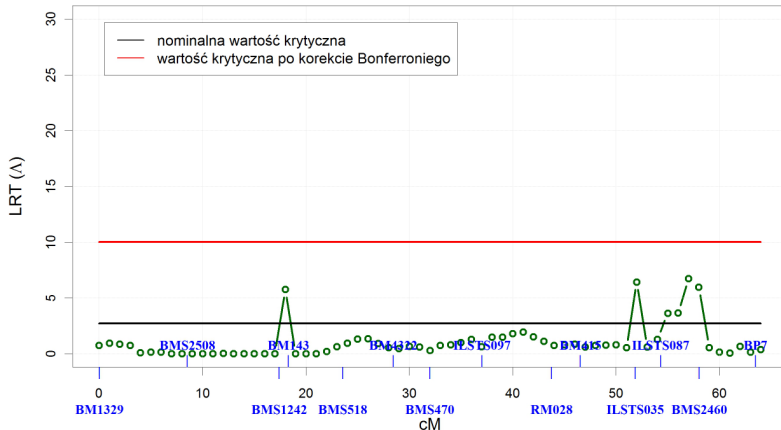


Wyniki - wydajność mleka (model 1)



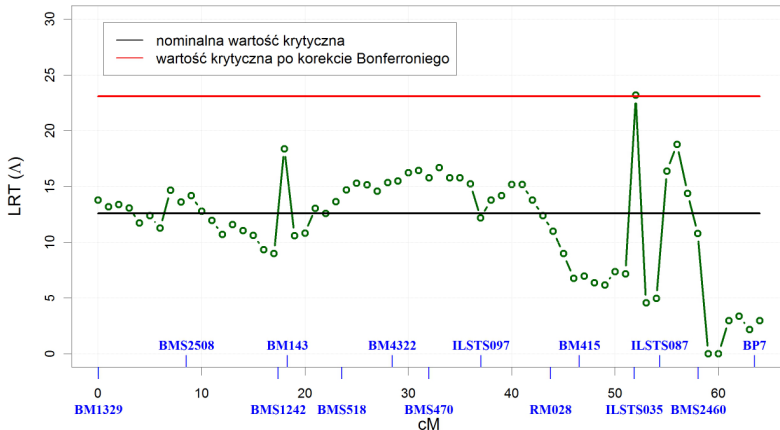


Wyniki - wydajność mleka (model 2)



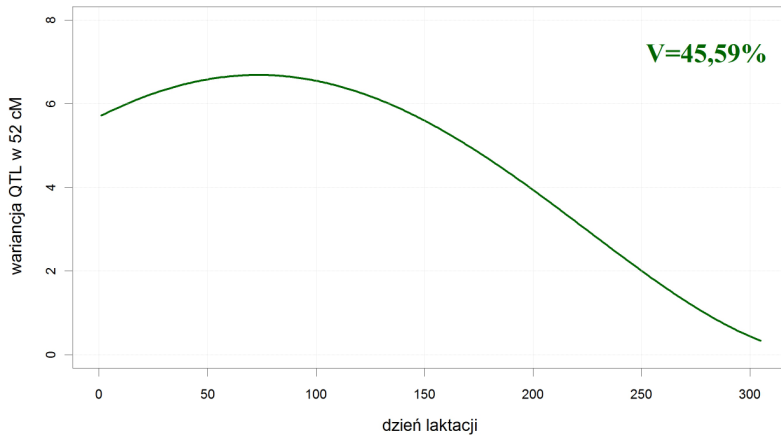


Wyniki - wydajność mleka (model 3)





Wyniki - wydajność mleka (wariancja QTL)





- ▶ Modele uwzględniające zmienny w czasie efekt QTL mają większą moc detekcji niż modele z efektem QTL stałym w czasie.
- ▶ Modele ze stałym efektem QTL nie są w stanie wykryć QTL o wyraźnie zmiennym efekcie .



Dziękuję za uwagę

