Laboratorium: Lista 6

Lokalizacja genów z wykorzystaniem testu Analizy Wariancji i testu niezależności chi-kwadrat

Statystyka w zastosowaniach

Sylwia Majchrowska
Matematyka

1 maja 2016

Spis treści

1 Wstęp

Zadania z tej listy dotyczą lokalizacji genów z wykorzystaniem testów analizy wariancji oraz niezależności chi-kwadrat. Zostały one przeprowadzone na podstawie krzyżówki typu intercross (bazującej na populacji F2). W związku z tym, że dla każdego odcinka DNA jedną z nici dziedziczymy po matce, a drugą po ojcu, to w danym miejscu markera mogą wystąpić 3 możliwości: aa, aA (ewentualnie Aa, ale kolejność nie jest istotna) oraz AA. Nie upraszczając analizy poprzez sprowadzenie krzyżówki do połączenia osobników krewniaczych (krzyżówka wsteczna), dla każdego organizmu możemy podać ciąg genotypów kodowanych za pomocą liczb 0, 1 i 2. Znamy też wartość interesującej nas cechy.

2 Zadanie pierwsze

Aby wygenerować macierz genotypów dla $n = 500$ osobników z krzyżówki typu intercross (populacji F2) na jednym chromosomie o długości 200 cM i odstępie między sąsiednimi markerami $\Delta = 1$ cM posłużono się przykładowym kodem do generacji genotypów na jednym chromosomie krzyżówki wstecznej omawianym na wykładzie generując dwa chromosomy (i dodając do siebie dwie macierze genotypów). Oznaczmy przez $M_{j,k}$ genotyp w j-tym markerze u k-tego osobnika, taki, że

$$M_{j,k} = \begin{cases} 0 & \text{aa}, \\ 1 & \text{Aa}, \\ 2 & \text{AA}. \end{cases}$$

Kolejnym krokiem było wygenerowanie wektora wartości cechy Y_i, którego wartości ściśle zależą od genotypu genu zlokalizowanego pośrodku chromosomu (tzn. w $\delta = 100$ cM), w taki sposób, że

$$Y_i \sim \begin{cases} N(\beta_1, 1) & \text{dla } M_{100,i} = 0, \\ N(\beta_2, 1) & \text{dla } M_{100,i} = 1, \\ N(0, 1) & \text{dla } M_{100,i} = 2, \end{cases}$$

gdzie $(\beta_1, \beta_2) \in \{(0.5, 0.25), (0.5, 0.5)\}$.

Estymacji położenia genu dokonano zgodnie ze wzorem

$$\hat{\delta} = \arg\max_{i \in \{0,\ldots,200\}} |t_i|,$$

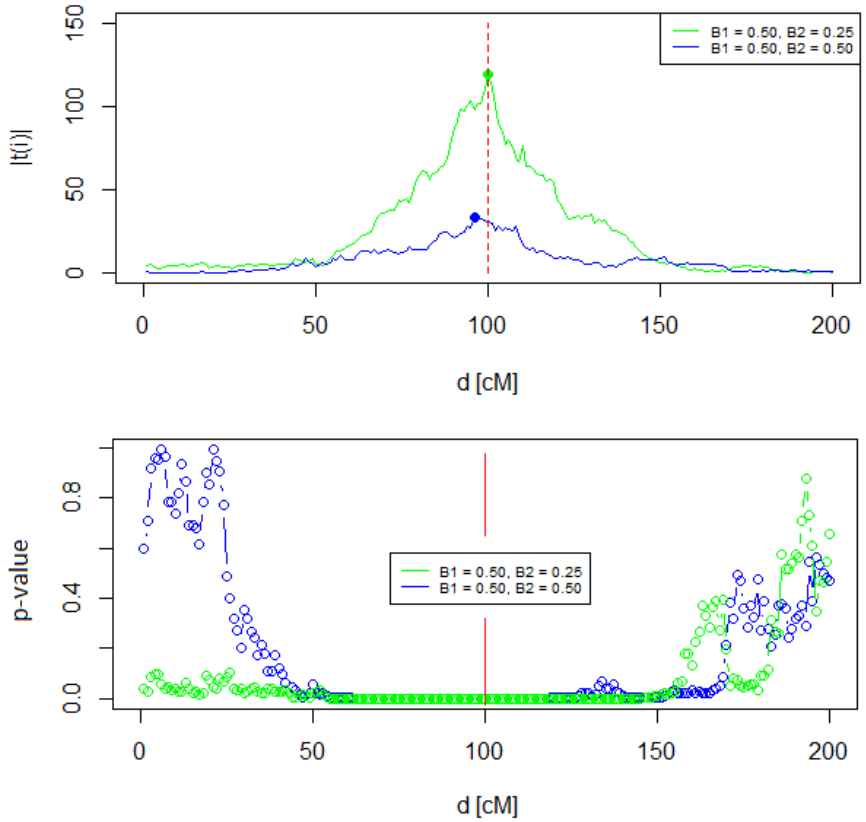
gdzie t_i jest statystyką testu Analizy Wariancji po całym chromosomie.

Wyestymowane położenia genów:

(β_1, β_2)	$\hat{\delta}$
(0.50, 0.25)	100
(0.50, 0.50)	96

Wyznaczone, na podstawie maksymalizacji statystyki testu analizy wariancji, położenia genu są zbliżone do rzeczywistego ulokowania skorelowanego odcinka DNA. Na wykresach 1 widać wyraźne wybrzuszenie (rozkład statystyk) lub wgłębienie (rozkład p-wartości) w okolicach środka chromosomu. Ta zależność jest mocniejsza w przypadku większych różnic między średnimi (różne β_i) w poszczególnych grupach genotypowych, co wiąże się z charakterem testu ANOVA

(H_0: Średnie w grupach (populacjach) są równe. vs H_1: Nie wszystkie średnie są równe.)



Rysunek 1: Wykresy zależności bezwzględnej wartości statystyk oraz p-wartości od odległości liczonej od lewego końca chromosomu. Czerwone linie pokazują położenie 100. markera, a pełne punkty wyestymowane położenia genu.

Kod źródłowy zadania 1 podpunktów a) - b) w języku R:

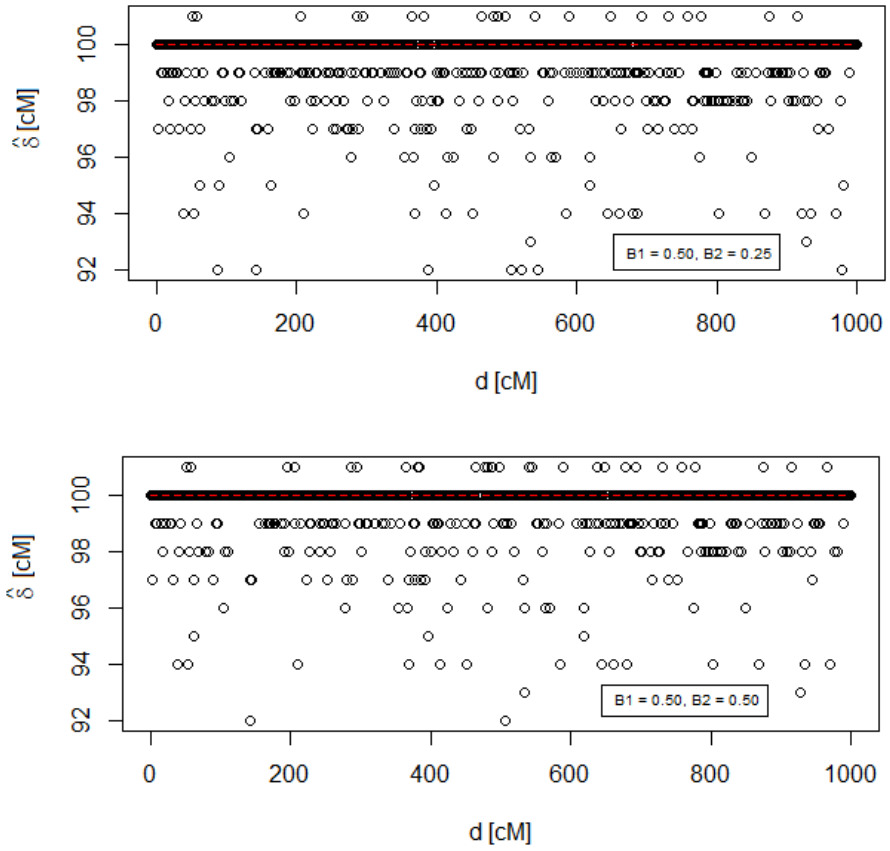
```
# ------------------------------------------------
# Autor: M. Bogdan
# 500 elementowa proba
set.seed(12)
# Generacja genotypow na jednym chromosomie z krzyzowki
    wstecznej
n <- 500 # liczba osobnikow - wiersze
L <- 200 # dlugosc chromosomu w cM - liczba markerow,
    kolumny;
d <- 1 # odleglosc miedzy sasiednimi markerami w cM
r <- 0.5*(1-exp(-0.02*d)) # p-stwo rekombinacji miedzy
    sasiednimi markerami
P <- rbinom(n,1,0.5) # genotypy w pierwszym markerze
```

```
R <- rbinom(n*(L-1),1,r)
R <- matrix(R,nrow=n,ncol=(L-1)) # macierz rekombinacji
    miedzy sasiednimi markerami
X <- cbind2(P,R)
X <- apply(X,1,'cumsum')
X <- t(X)
X <- X%%2
set.seed(1222)
# Generacja genotypow na jednym chromosomie z krzyzowki
    wstecznej
n <- 500 # liczba osobnikow - wiersze
L <- 200 # dlugosc chromosomu w cM - liczba markerow,
    kolumny;
d <- 1 # odleglosc miedzy sasiednimi markerami w cM
r <- 0.5*(1-exp(-0.02*d)) # p-stwo rekombinacji miedzy
    sasiednimi markerami
P <- rbinom(n,1,0.5) # genotypy w pierwszym markerze
R <- rbinom(n*(L-1),1,r)
R <- matrix(R,nrow=n,ncol=(L-1)) # macierz rekombinacji
    miedzy sasiednimi markerami
Y <- cbind2(P,R)
Y <- apply(Y,1,'cumsum')
Y <- t(Y)
Y <- Y%%2
M = X + Y  # finalna macierz genotypow
# ------------------------------------------------------
# a)
set.seed(251)
beta11 = 0.50
beta12 = 0.25
beta21 = 0.50
beta22 = 0.50
wiersz = M[,100]
ce = rnorm(500)
for(ind in 1:500){
  if(wiersz[ind] == 1){
    wiersz[ind] = ce[ind] + beta12
    ce[ind] = ce[ind] + beta22
  }
  if(wiersz[ind] == 0){
    wiersz[ind] = ce[ind] + beta11
    ce[ind] = ce[ind] + beta21
  }
}
```

```
# b)
Stat01 = sapply(1:L, function(i){anova(lm(wiersz~M[,i]))$
    'F value'[1]})
pwar01 = sapply(1:L, function(i){anova(lm(wiersz~M[,i]))$
    'Pr(>F)'[1]})
d01 = which.max(abs(Stat01))  # 100
Stat02 = sapply(1:L, function(i){anova(lm(ce~M[,i]))$'F
    value'[1]})
pwar02 = sapply(1:L, function(i){anova(lm(ce~M[,i]))$'Pr
    (>F)'[1]})
d02 = which.max(abs(Stat02))  # 96
# plot(abs(Stat01), xlim=c(0,200), ylim=c(0,150), xlab =
    'd [cM]', ylab = '|t(i)|', type = 'l', col = 'green')
# lines(abs(Stat02), xlim=c(0,200), xlab = 'd [cM]', ylab
    = '|t(i)|', col = 'blue', type = 'l')
# legend("topright", c('B1 = 0.50, B2 = 0.25', 'B1 =
    0.50, B2 = 0.50'), col = c('green','blue'), text.col =
    "black", lty = 1, cex=0.6)
# lines(rep(100,301), seq(0,150, by = 0.5), type="l", pch
    =22, lty=2, col="red")
```

2.1 Podpunkt c)

Aby wyznaczyć obciążenie oraz odchylenie standardowe wyznaczonych estymatorów położenia powyższe doświadczenie powtórzono 1000 dla każdej pary $\{\beta_1, \beta_2\}$.

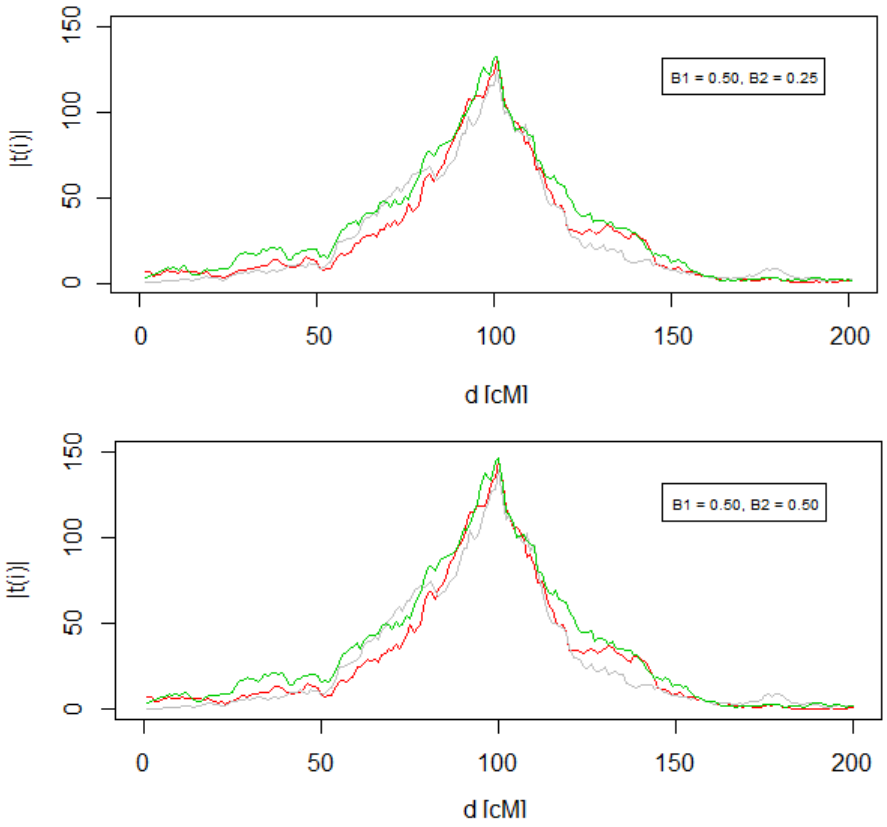


Rysunek 2: Wykresy zależności wyestymowanych położeń genu od nr eksperymentu. Czerwoną przerywaną linią oznaczono rzeczywiste umiejscowienie genu.

Jak można zauważyć na wykresach 2 większość wyznaczonych estymatorów idealnie trafia w znane umiejscowienie genu. Pomimo to rozrzut estymatorów jest dość duży - $\hat{\delta} \in [92, 102] \cap \mathbb{N}$ - dla obu par parametrów (β_1, β_2).

Uzyskane wyniki:

(β_1, β_2)	$\sigma(\hat{\delta})$	$E(\hat{\delta}) - \delta$
(0.50, 0.25)	1.3562	-0.5810
(0.50, 0.50)	1.164548	-0.428000



Rysunek 3: Wykresy zależności bezwzględnej wartości statystyk od odległości liczonej od lewego końca chromosomu.

Wraz ze wzrostem parametru β_2 maleje odchylenie standardowe estymatora położenia, co jest związane ze zwiększeniem siły sygnału, co dość dobrze obrazuje wykres 3 - większe wartości bezwzględne statystyk testu jednoczynnikowej analizy wariancji (bardziej *strzeliste wybrzuszenia*). Mimo to obserwujemy podobny rozrzut wartości $\hat{\delta}$ w obu przypadkach. Oszacowane obciążenie dla tego parametru jest bliskie zeru, a jego ujemna wartość wskazuje na delikatne zaniżenie wielkości estymatora w stosunku do szacowanego parametru (na wykresie 2 widać przewagę estymatorów zaniżonych w stosunku do zawyżonych).

Kod źródłowy zadania 1 podpunktu c) w języku R:

```
doswiadczenie2 <- function(b1,b2, sumNum, genotypy, nr_
    markera){
  set.seed(6532)
  osobniki = length(genotypy[,1])  # 500
  L = length(genotypy[1,])  # 200 = dlugosc chromosomu w
      cM
  cecha = matrix(rnorm(osobniki * sumNum), osobniki,
      sumNum)
  wiersze = cbind(genotypy[,nr_markera],genotypy[,nr_
      markera])
  for (ind in 3:sumNum){
    wiersze = cbind(wiersze,genotypy[,nr_markera])
  }

  for (i in 1:sumNum){
    for(ind in 1:osobniki){
      if(wiersze[ind,i] == 1){
        wiersze[ind,i] = cecha[ind,i] + b2
      }
      if(wiersze[ind,i] == 0){
        wiersze[ind,i] = cecha[ind,i] + b1
      }
    }
  }
  rm(cecha)
  Stat = matrix(0,L,sumNum)
  for(j in 1:sumNum){
    Stat[,j] = sapply(1:L, function(i){anova(lm(wiersze[,
        j]~genotypy[,i]))$'F value'[1]})
  }
  # lines(abs(Stat[,100]), col = 2, type='l')
  # lines(abs(Stat[,200]), col = j, type='l')
  # lines(abs(Stat[,1000]), col = 3, type='l')

  d = apply(abs(Stat), 2, which.max)

  # return (d)
  return (c(sd(d), mean(d) - nr_markera))
}

w1 = doswiadczenie2(0.5, 0.25, 1000,M,100)
w2 = doswiadczenie2(0.5, 0.5, 1000,M,100)
```

3 Zadanie drugie

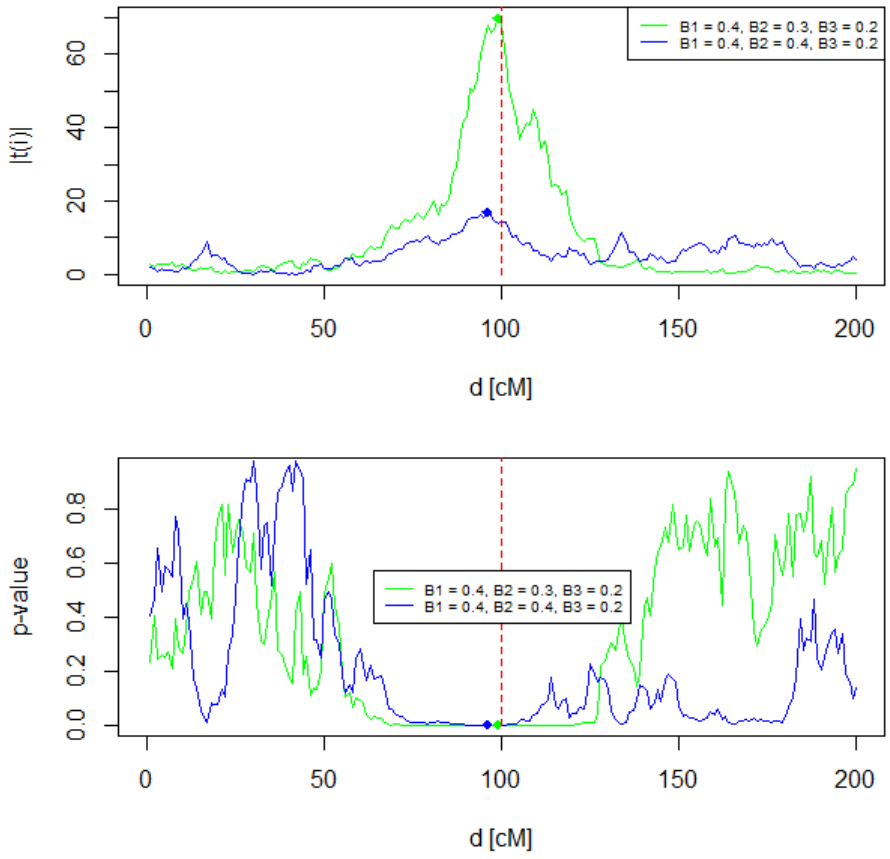
W tym zadaniu należało wygenerować wektor cechy binarnej Y_i, którego wartości ściśle zależą od genotypu genu zlokalizowanego pośrodku chromosomu (tzn. w $\delta = 100\,\text{cM}$), w taki sposób, że

$$Y_i \sim \begin{cases} B(1,\beta_1) & \text{dla } M_{100,i} = 0, \\ B(1,\beta_2) & \text{dla } M_{100,i} = 1, \\ B(1,\beta_3) & \text{dla } M_{100,i} = 2, \end{cases}$$

gdzie $(\beta_1, \beta_2, \beta_3) \in \{(0.4, 0.3, 0.2), (0.4, 0.4, 0.2)\}$. Estymacji położenia genu dokonano maksymalizując statystykę testu chi-kwadrat niezależności po całym chromosomie.

Wyestymowane położenia genów:

$(\beta_1, \beta_2, \beta_3)$	$\hat{\delta}$
(0.4, 0.3, 0.2)	99
(0.4, 0.4, 0.2)	96



Rysunek 4: Wykresy zależności bezwzględnej wartości statystyk oraz p-wartości od odległości liczonej od lewego końca chromosomu. Czerwone linie pokazują położenie 100. markera, a pełne punkty wyestymowane położenia genu.

Test niezależności chi-kwadrat został użyty w celu zbadania zależności pomiędzy wygenerowaną cechą, a fragmentem DNA. Wyznaczone wartości estymatorów położenia genu są bliskie 100, dodatkowo na podstawie uzyskanych wykresów 4 możemy zauważyć, że wyliczone statystyki (p-wartości) są większe (mniejsze) pośrodku chromosomu.

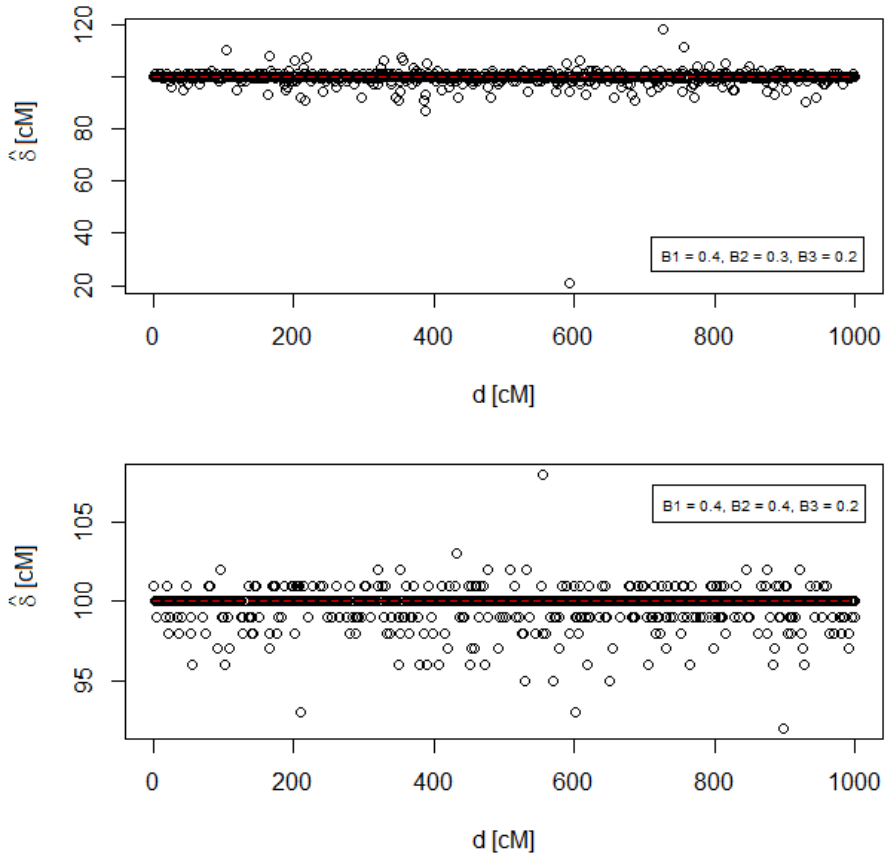
Kod źródłowy zadania 2 podpunkty a) - b) w języku R:

```
# a)
set.seed(25134)
beta11 = 0.4; beta12 = 0.3; beta13 = 0.2
beta21 = 0.4; beta22 = 0.4; beta23 = 0.2
wiersz = M[,100]
ce = M[,100]
for(ind in 1:n){
  if(wiersz[ind] == 1){
    wiersz[ind] = rbinom(1,1,beta12)
    ce[ind] = rbinom(1,1,beta22)
  }
  if(wiersz[ind] == 0){
    wiersz[ind] = rbinom(1,1,beta11)
    ce[ind] = rbinom(1,1,beta21)
  }
  if(wiersz[ind] == 2){
    wiersz[ind] = rbinom(1,1,beta13)
    ce[ind] = rbinom(1,1,beta23)
  }
}
# b)
Stat1 = rep(0,L); Stat2 = rep(0,L)
pwar1 = rep(0,L); pwar2 = rep(0,L)
for (i in 1:L){
  Stat1[i] = chisq.test(wiersz,M[,i], correct = FALSE)$
      statistic
  Stat2[i] = chisq.test(ce,M[,i], correct = FALSE)$
      statistic

  pwar1[i] = chisq.test(wiersz,M[,i], correct = FALSE)$p.
      value
  pwar2[i] = chisq.test(ce,M[,i], correct = FALSE)$p.
      value
}
d1 = which.max(abs(Stat1)) # 99
d2 = which.max(abs(Stat2)) # 96
```

3.1 Podpunkt c)

Aby wyznaczyć obciążenie oraz odchylenie standardowe wyznaczonych estymatorów położenia doświadczenie powtórzono 1000 dla każdej trójki $\{\beta_1, \beta_2, \beta_3\}$.



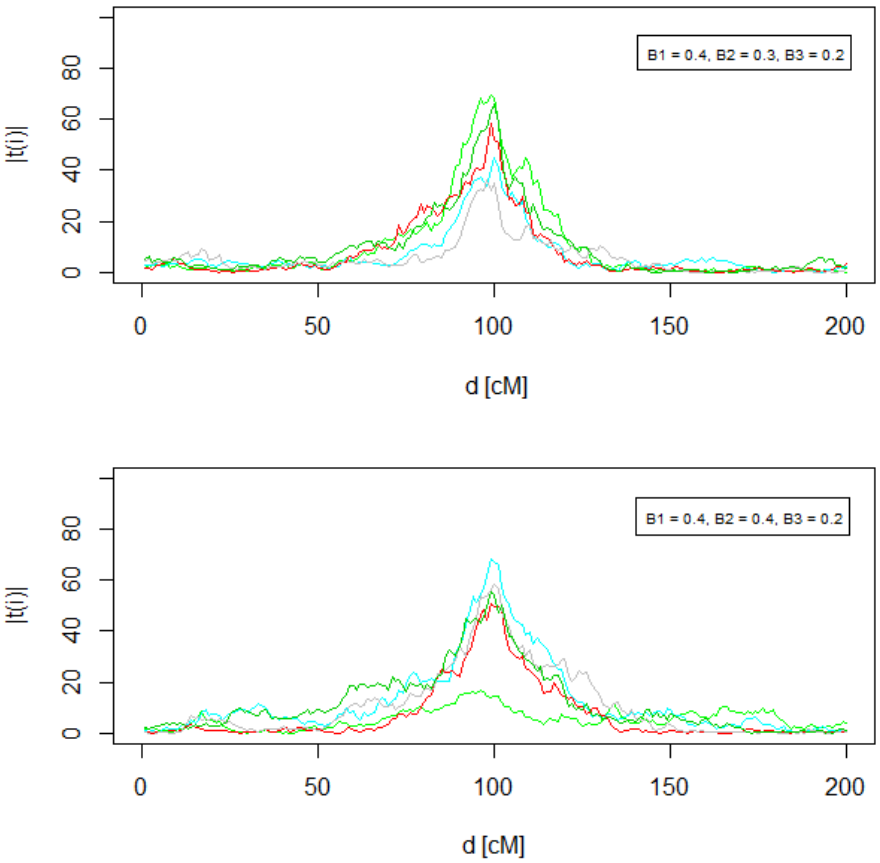
Rysunek 5: Wykresy zależności wyestymowanych położeń genu od nr eksperymentu. Czerwoną przerywaną linią oznaczono rzeczywiste umiejscowienie genu.

Uzyskane wyniki:

$(\beta_1, \beta_2, \beta_3)$	$\sigma(\hat{\delta})$	$E(\hat{\delta}) - \delta$
(0.4, 0.3, 0.2)	3.1368710	-0.391
(0.4, 0.4, 0.2)	0.9932863	-0.213

Na wykresach nr 5 zaznaczono wartości uzyskanych estymatorów $\hat{\delta}$ w zależności od nr eksperymentu. Wyniki w obu przypadkach oscylują wokół wartości rzeczywistej parametru położenia genu. Dla trójki (0.4, 0.3, 0.2) rozrzut parametrów jest większy niż dla drugiego rozważanego przypadku, co wiąże się z większą siłą (prawdopodobieństwem) sygnału dla poszczególnych genotypów. Tę obserwację potwierdzają uzyskane oszacowania odchyleń standardowych. Uzyskane obciążenia i w tym zadaniu są ujemne, co widać przy lekko niesymetrycznym rozłożeniu estymatorów wokół 100 (w przypadku (0.4, 0.3, 0.2) może nie być to

tak widoczne ze względu na występujące obserwacje odstające - należy zwrócić szczególna uwagę na skalę na osi Y).



Rysunek 6: Wykresy zależności bezwzględnej wartości statystyk od odległości liczonej od lewego końca chromosomu.

Kod źródłowy zadania 2 podpunktu c) w języku R:

```
# c)
doswiadczenie3 <- function(b1, b2, b3, sumNum, genotypy,
    nr_markera){
  set.seed(65632)
  osobniki = length(genotypy[,1])  # 500
  L = length(genotypy[1,])  # 200 = dlugosc chromosomu w
      cM
  wiersze = cbind(genotypy[,nr_markera],genotypy[,nr_
      markera])
  for (ind in 3:sumNum){
    wiersze = cbind(wiersze,genotypy[,nr_markera])
  }
```

```
  for (i in 1:sumNum){
    for(ind in 1:osobniki){
      if(wiersze[ind,i] == 1){
        wiersze[ind,i] = rbinom(1,1,b2)
      }
      if(wiersze[ind,i] == 0){
        wiersze[ind,i] = rbinom(1,1,b1)
      }
      if(wiersze[ind,i] == 2){
        wiersze[ind,i] = rbinom(1,1,b3)
      }
    }
  }
  Stat = matrix(0,L,sumNum)
  for(j in 1:sumNum){
    for (i in 1:L){
      Stat[i,j] = chisq.test(wiersze[,j],M[,i], correct =
          FALSE)$statistic
    }}

  d = apply(abs(Stat), 2, which.max)
  # return (d)
  return (c(sd(d), mean(d) - nr_markera))
}
w1 = doswiadczenie3(0.4, 0.3, 0.2, 1000,M,100)
w2 = doswiadczenie3(0.4, 0.4, 0.2, 1000,M,100)
```