Laboratorium: Lista 7

Lokalizacja genów z wykorzystaniem regresji wielorakiej

Statystyka w zastosowaniach

Sylwia Majchrowska
Matematyka

24 maja 2016

Spis treści

1 Wstęp

Zasadniczym problemem testów w pojedynczych markerach jest fakt, że zupełnie ignorujemy wpływ pozostałych markerów. Jeśli związek z cechą ma więcej genów (a zwykle tak jest), to lepszym pomysłem jest próba dopasowania modelu, który wszystkie te istotne geny zawiera. Dodatkowo geny mogą wchodzić ze sobą w interakcje. Wszystko to możemy zamodelować przy pomocy regresji wielokrotnej. Jeśli będziemy rozważać jedynie interakcje drugiego rzędu, to model dla przypadku z dwiema wersjami genotypów jest postaci

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \sum_{1 \leqslant j < l \leqslant m} \gamma_{jl} X_{ij} X_{il} + \epsilon_i. \tag{1}$$

W praktyce, ponieważ m jest duże, ograniczamy się właśnie do interakcji drugiego rzędu, a czasem w ogóle z nich rezygnujemy [3].

2 Zadanie pierwsze

Aby wygenerować macierz genotypów dla $n = 500$ osobników z krzyżówki wstecznej na jednym chromosomie o długości 200 cM i odstępie między sąsiednimi markerami $\Delta = 1\,\text{cM}$ posłużono się przykładowym kodem do generacji genotypów na jednym chromosomie krzyżówki wstecznej omawianym na wykładzie.

Oznaczmy przez $X_{j,k}$ genotyp w j-tym markerze u k-tego osobnika, taki, że

$$X_{j,k} = \begin{cases} \frac{-1}{2} & \text{aa}, \\ \frac{1}{2} & \text{Aa}. \end{cases}$$

Kolejnym krokiem było wygenerowanie wektora wartości cechy Y_i, którego wartości ściśle zależą od genotypu genów zlokalizowanych na 80. i 120. markerze (tzn. w $\delta_{80} = 80\,\text{cM}$ oraz $\delta_{120} = 120\,\text{cM}$), w taki sposób, że

$$Y_i = \beta_1 X_{i,80} + \beta_2 X_{i,120} + \epsilon_i \tag{2}$$

gdzie $(\beta_1, \beta_2) \in \{(0.25, 0.5)\}$, a wektor reszt $\epsilon_i \sim N(0, 1)$.

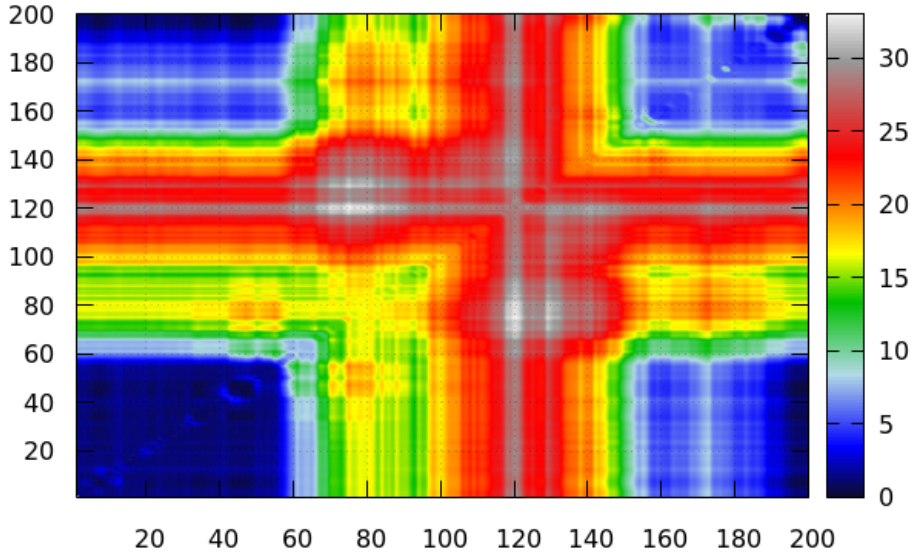
Estymacji położenia genu dokonano maksymalizując statystykę regresji wielorakiej (stosując test F-Snedecora, można stosować także równoważny test F dla analizy wariancji, czy też test Studenta - w przypadku, gdy mamy do czynienia z dwoma genotypami) dla każdej pary markerów po całym chromosomie.

Wyestymowane położenia genów:

$\hat{\delta}_{80}$	$\hat{\delta}_{120}$
75	120

Wyznaczone, na podstawie maksymalizacji statystyki testu F-Snedecora, położenia genów są zbliżone do rzeczywistego ulokowania skorelowanego odcinka DNA. Na wykresie 1 widać wyraźne wybrzuszenie (odpowiadające odcieniom szarości, bieli i czerwieni) w okolicach pary (80, 120) czy też analogicznie (120, 80). Został on sporządzony przy pomocy programu *gnuplot*. Osie X i Y odpowiadają kolejnym markerom na chromosomie (otrzymane wyniki odbito lustrzanie względem prostej y=x, gdyż kolejność genów przy badaniu ich związku z cechą

nie jest istotna), a skala barwna otrzymanym wartościom statystyk dla kolejnych par.



Rysunek 1: Mapy zależności bezwzględnej wartości statystyk dla poprawnego modelu addytywnego od położenia genów mających wpływ na badaną cechę liczonych od lewego końca chromosomu.

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

```
# ------------------------------------------------------------
# Autor: M. Bogdan
set.seed(121680)
# 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 # finalna macierz genotypow
# ------------------------------------------------------------
```

```
# zadanie 1
# a)
Y = Y - 0.5
beta1 = 0.25
beta2 = 0.5
set.seed(12)
cecha = beta1 * Y[,80] + beta2 * Y[,120] + rnorm(500)

# b)
szukanie_pary <- function(ce, genotypy, markery){
  Stat1 = 0.0
  Stat2 = 0.0
  para = rep(0,2)
  for(i in 1:(markery-1)){
    for(j in (i+1):markery){
      Stat2 = summary(lm(ce~genotypy[,c(i,j)]))$
          fstatistic[1]
      if (!is.null(Stat2)){
        if(Stat2 > Stat1){
          Stat1 = Stat2
          para = c(i,j)
        }
      }
    }
  }
  return (para)
}

szukanie_pary(cecha, Y, L) # c(75,120)
```

2.1 Podpunkt c)

Aby wyznaczyć obciążenie, odchylenie standardowe oraz błąd średniokwadratowy wyznaczonych estymatorów położeń powyższe doświadczenie powtórzono 1000 razy.

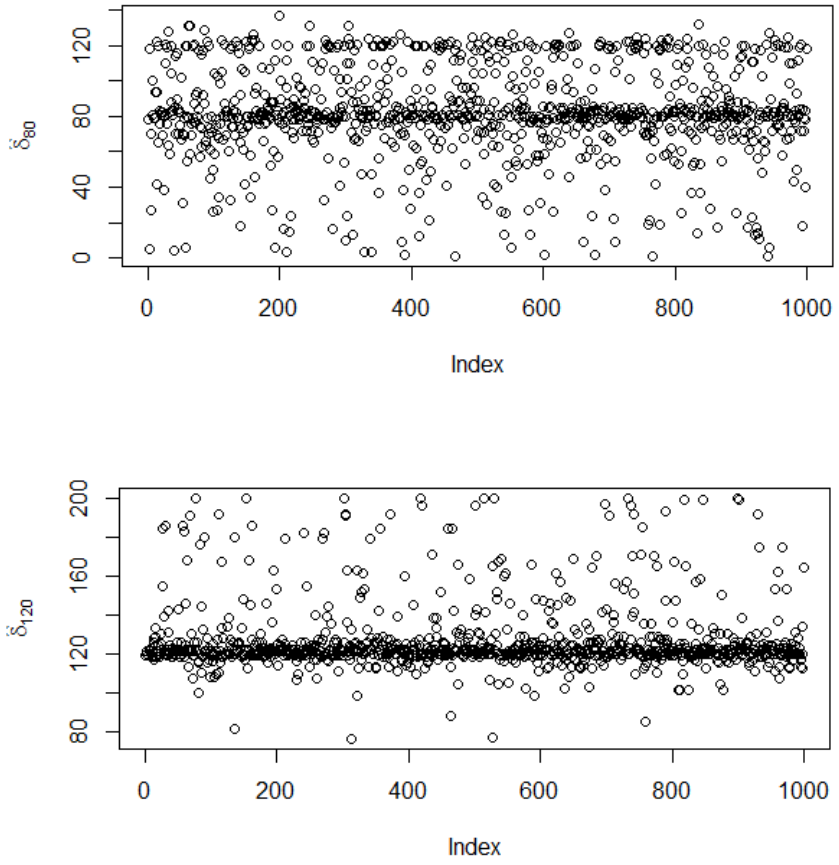
Jak można zauważyć na wykresach 2 większość wyznaczonych estymatorów trafia w znane umiejscowienia genów, trochę lepiej jest to obserwowane dla estymatora $\hat{\delta}_{120}$ - większa wartość β_2. Dodatkowo w przypadku $\hat{\delta}_{80}$ widać silny wpływ genu ulokowanego na 120 cM. Pomimo to rozrzut estymatorów jest dość duży - $\hat{\delta}_{80} \in [1, 137] \cap \mathbb{N}$ oraz $\hat{\delta}_{120} \in [76, 200] \cap \mathbb{N}$.

Uzyskane wyniki:

δ	$\sigma(\hat{\delta})$	$E(\hat{\delta}) - \delta$	$E(\hat{\delta} - \delta)^2$
80	26.18693	2.511	696.672
120	17.42256	6.457	344.935

Im większy parametr β związany jest z konkretnym genem, tym silniejsza będzie jego detekcja. Tym samym odchylenie standardowe estymatora położenia $\hat{\delta}_{80}$ ($\beta_1 < \beta_2$) jest większe niż $\hat{\delta}_{120}$. Oszacowane obciążenia dla obu parametrów są dodatnie, co wskazuje na zawyżenie wielkości estymatorów w stosunku do szacowanych parametrów (na wykresie 2 widać przewagę estymatorów zawyżonych w stosunku do zaniżonych).

Dla obu estymatorów otrzymano dość duży błąd średniokwadratowy. Wartość błędu średniokwadratowego jest rozłożona na błąd związany ze zmiennością estymatora oraz kwadrat błędu systematycznego. Zgadzając się na obciążenie estymatora otrzymujemy estymator o mniejszym błędzie średniokwadratowym. Odległość średniokwadratowa jest miarą dokładności estymacji i informuje, o ile przeciętnie wartości estymatora odchylają się od rzeczywistej wartości parametru. Oznacza to, że im mniejszy MSE (pierwiastek z MSE) tym większa dokładność estymacji. A więc w tym przypadku nie jest ona zbyt dokładna.



Rysunek 2: Wyestymowane na podstawie 1000 symulacji położenia obu genów mających wpływ na cechę.

Kod źródłowy zadania 1 podpunktu c) w języku R:

```
# c
SymNumer = 1000
wyniki = matrix(0, 2, SymNumer)
for(index in 1:SymNumer){
  cecha = beta1 * Y[,80] + beta2 * Y[,120] + rnorm(500)
  wyniki[,index] = szukanie_pary(cecha, Y, L)
}
# 80cM
sd(wyniki[1,]) # odchylenie standardowe
mean(wyniki[1,]) - 80 # obciazenie
mean((wyniki[1,] - 80)^2) # blad sredniokwadratowy

# 120cM
sd(wyniki[2,]) # odchylenie standardowe
mean(wyniki[2,]) - 120 # obciazenie
mean((wyniki[2,] - 120)^2) # blad sredniokwadratowy

plot(wyniki[2,], ylab=expression(hat(delta)[120]))
plot(wyniki[1,], ylab=expression(hat(delta)[80]))
```

3 Zadanie drugie

Model regresji liniowej możemy rozszerzyć wprowadzając do niego sztucznie stworzone predyktory - iloczyny dwóch lub większej liczby zmiennych objaśniających. Pozwala to na uwzględnienie interakcji pomiędzy zmiennymi, czyli zmiany siły wpływu jednej ze zmiennych przy różnych wartościach innej zmiennej.

W tym zadaniu należało wygenerować wektor cechy Y_i zgodnie z modelem z interakcją

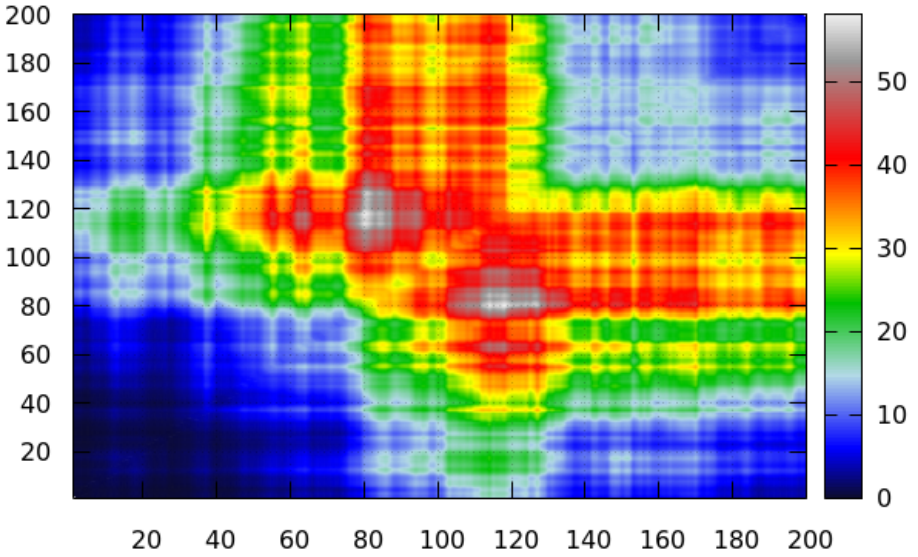
$$Y_i = \beta_1 X_{i,80} + \beta_2 X_{i,120} + \gamma X_{i,80} X_{i,120} + \epsilon_i \tag{3}$$

gdzie $(\beta_1, \beta_2) \in \{(0.25, 0.5)\}$, $\gamma \in \{0.5, 1\}$, a wektor reszt $\epsilon_i \sim N(0, 1)$.

Estymacji położenia genów dokonano w oparciu o odpowiedni model regresji.

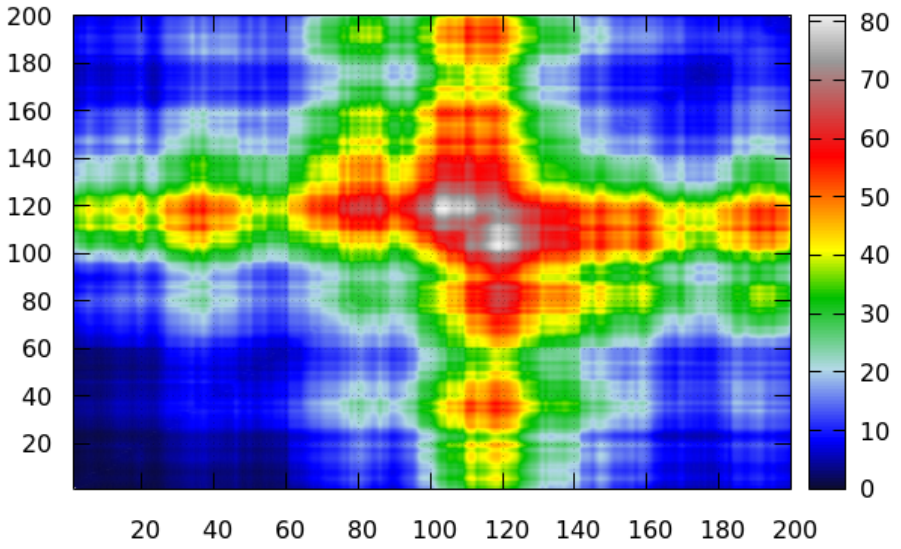
Wyestymowane położenia genów:

γ	$\hat{\delta}_{80}$	$\hat{\delta}_{120}$
0.5	80	118
1	104	119



Rysunek 3: Mapy zależności bezwzględnej wartości statystyk dla modelu z interakcją od położenia genów skorelowanych z cechą liczonych od lewego końca chromosomu dla $\gamma = 1$.

Estymacji położenia genu dokonano maksymalizując statystykę regresji wielorakiej. Wyznaczone wartości estymatorów położenia genu są dość bliskie ich wartościom rzeczywistym. Wyjątek może tu stanowić wielkość 104 odpowiadająca 80. markerowi. Dodatkowo na podstawie uzyskanych map nr 3, 4 możemy zauważyć, że wyliczone statystyki są większe (odpowiadają odcieniom szarości, bieli i czerwieni) w obrębie pary (80, 120) czy też analogicznie (120, 80). Lepiej prezentują się estymatory oraz wyliczone statystyki dla $\gamma = 0.5$, gdyż dla $\gamma = 0.5$ związek z markerem 80. jest mocno *zagłuszany* przez korelację z markerem 120.



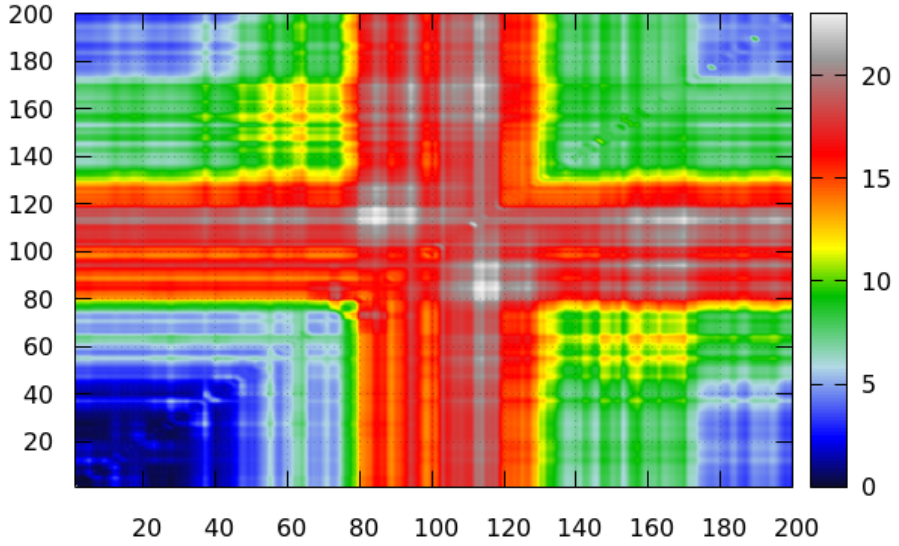
Rysunek 4: Mapy zależności bezwzględnej wartości statystyk dla modelu z interakcją od położenia genów skorelowanych z cechą liczonych od lewego końca chromosomu dla $\gamma = 0.5$.

Estymacji położenia genów dokonano także w oparciu o błędny model addytywny regresji (wstęp do podpunktu b).

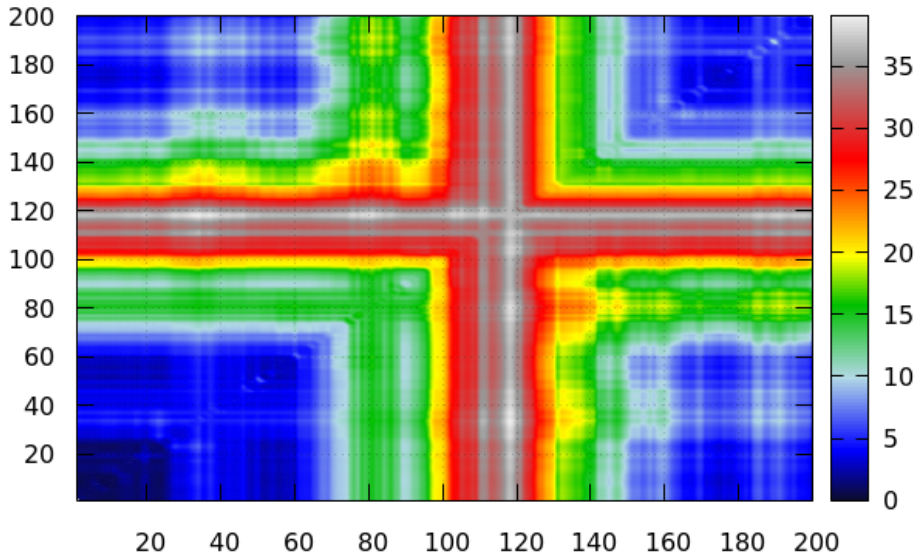
Wyestymowane położenia genów:

γ	$\hat{\delta}_{80}$	$\hat{\delta}_{120}$
0.5	84	114
1	34	118

W tym przypadku na wykresach nr 5 i 6 widać o wiele silniejszy wpływ markera oddalonego o 120 cM licząc od lewej strony chromosomu (większa wartość β_2). Wciąż możemy zaobserwować delikatne wysepki (białe plamy) na skrzyżowaniu wartości 80 i 120, jednakże nie są one już tak wyraźne, jak przy zastosowaniu do lokalizacji poprawnego modelu z interakcją. Dodatkowo w przypadku $\gamma = 1$ oszacowana wartość estymatora położenia dalece odbiega od jego rzeczywistej wartości. Warto zwrócić uwagę na zmieniające się przedziały wartości, w granicach których oscylują wyliczone statystyki.



Rysunek 5: Mapy zależności bezwzględnej wartości statystyk dla błędnego modelu addytywnego od położenia genów skorelowanych z cechą liczonych od lewego końca chromosomu dla $\gamma = 1$.



Rysunek 6: Mapy zależności bezwzględnej wartości statystyk dla błędnego modelu addytywnego od położenia genów skorelowanych z cechą liczonych od lewego końca chromosomu dla $\gamma = 0.5$.

Kod źródłowy zadania 2 w języku R:

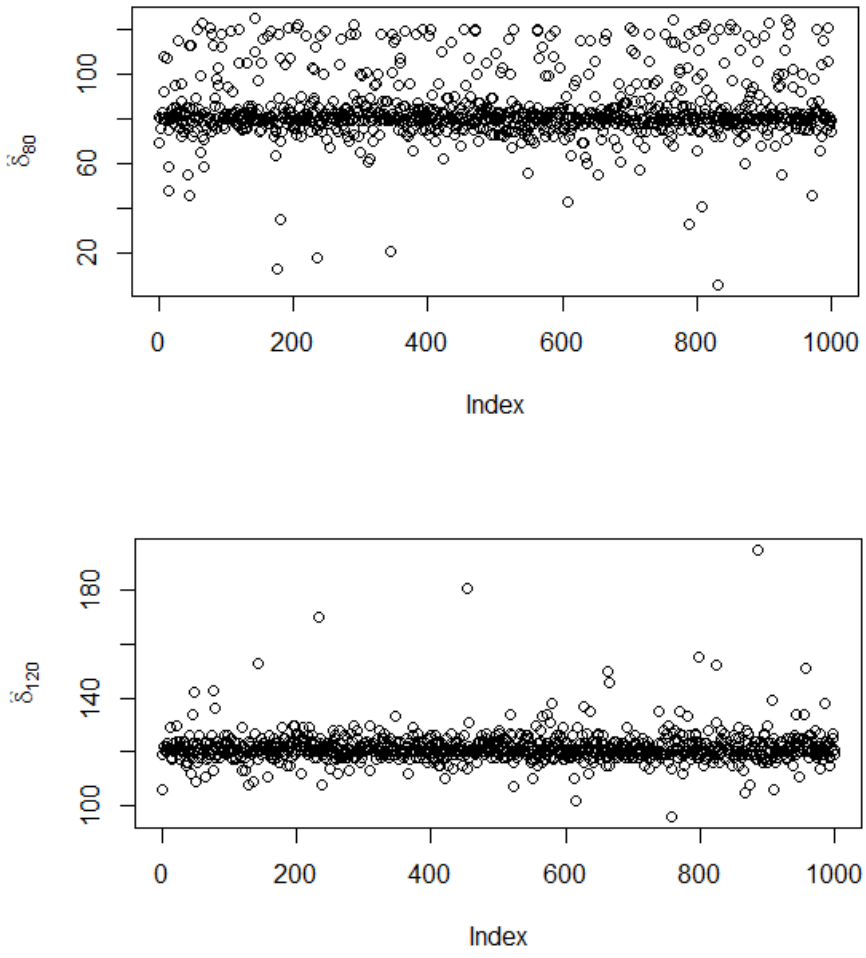
```
# zadanie 2
beta1 = 0.25
beta2 = 0.5
gamma5 = 0.5
gamma1 = 1.0

set.seed(112)
cecha2_5 = beta1 * Y[,80] + beta2 * Y[,120] + gamma5*Y
    [,80] * Y[,120] + rnorm(500)
cecha2_1 = beta1 * Y[,80] + beta2 * Y[,120] + gamma1*Y
    [,80] * Y[,120] + rnorm(500)

szukanie_pary2 <- function(ce, genotypy, markery){
  Stat1 = 0.0
  Stat2 = 0.0
  para = rep(0,2)
  for(i in 1:(markery-1)){
    for(j in (i+1):markery){
      Stat2 = summary(lm(ce~I(genotypy[,i]+genotypy[,j]+
          genotypy[,i]*genotypy[,j])))$fstatistic[1]
      if (!is.null(Stat2)){
        if(Stat2 > Stat1){
          Stat1 = Stat2
          para = c(i,j)
        }
      }
    }
  }
  return (para)
}

szukanie_pary2(cecha2_1, Y, L)  # c(80, 118)
szukanie_pary2(cecha2_5, Y, L)  # c(104, 119)

szukanie_pary(cecha2_1, Y, L)  # c(84, 114)
szukanie_pary(cecha2_5, Y, L)  # c(34, 118)
```

3.1 Podpunkt a)

Aby wyznaczyć obciążenie, odchylenie standardowe oraz błąd średniokwadratowy wyznaczonych estymatorów położeń doświadczenie związane z dopasowywaniem modelu z interakcją dla każdej pary markerów powtórzono 1000 razy.

Uzyskane wyniki:

γ	δ	$\sigma(\hat{\delta})$	$E(\hat{\delta}) - \delta$	$E(\hat{\delta} - \delta)^2$
0.5	80	13.84174	3.912	206.7060
	120	5.835238	1.493	13.84174
1	80	6.671707	0.279	44.54500
	120	3.619411	0.842	13.79600



Rysunek 7: Wyestymowane na podstawie 1000 symulacji położenia obu genów mających wpływ na cechę dla $\gamma = 0.5$.

Jak można zauważyć na wykresach 7 większość wyznaczonych estymatorów trafia w znane umiejscowienia genów. Ponownie dużo lepiej jest to obserwowane dla estymatora $\hat{\delta}_{120}$. W tym przypadku dla $\hat{\delta}_{80}$ silny wpływ genu ulokowanego na 120 cM już tak bardzo nie *zagłusza* markera 80.

O wiele mniejsze odchylenia standardowe, obciążenia i odległości średniokwadratowe dla szacowanych estymatorów obserwujemy dla $\gamma = 1$, czyli większego udziału interakcji. W szczególności widać to po uzyskanych wartościach obciążenia, które w obu przypadkach są nieznacznie większe od 0. Oczywiście im mniejszy jest błąd średniokwadratowy dla modelu tym model jest lepszy.

Kod źródłowy zadania 2 podpunktu a) w języku R:

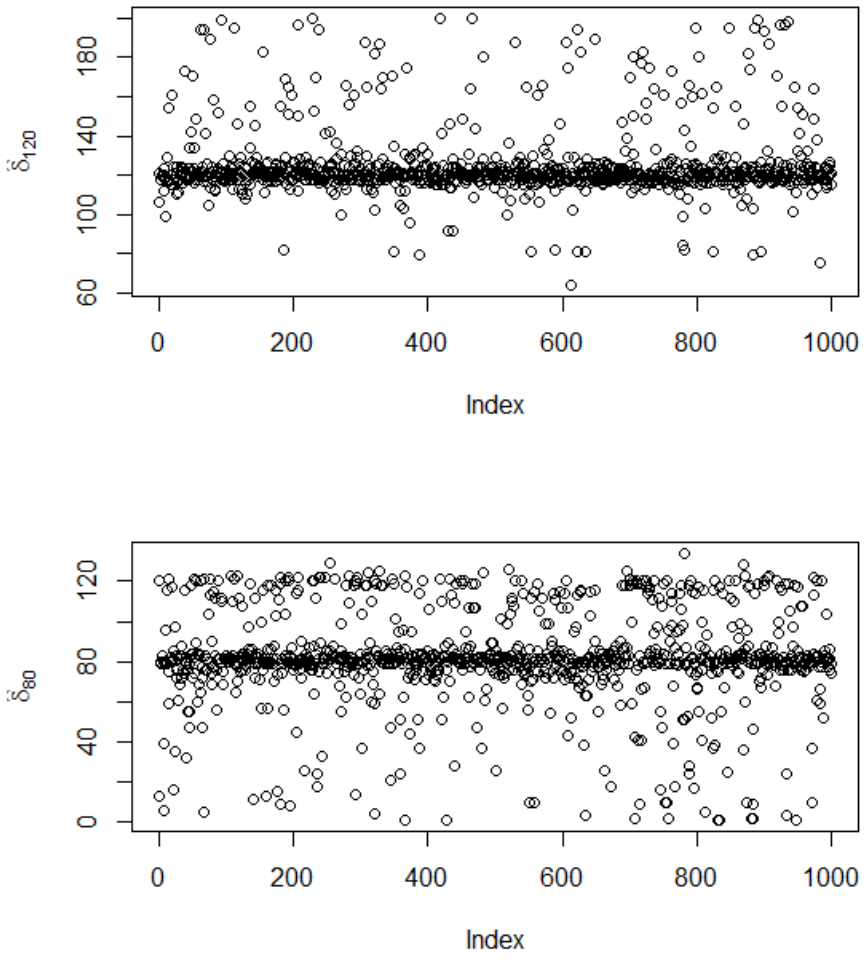
```
# a)
SymNumer = 1000
wyniki2a5 = matrix(0, 2, SymNumer)
for(index in 1:SymNumer){
  cecha2_5 = beta1 * Y[,80] + beta2 * Y[,120] + gamma5*Y
      [,80]*Y[,120] + rnorm(500)
  wyniki2a5[,index] = szukanie_pary2(cecha2_5, Y, L)
}
# 80cM
sd(wyniki2a5[1,]) # odchylenie standardowe
mean(wyniki2a5[1,]) - 80 # obciazenie
mean((wyniki2a5[1,] - 80)^2) # blad sredniokwadratowy

# 120cM
sd(wyniki2a5[2,]) # odchylenie standardowe
mean(wyniki2a5[2,]) - 120 # obciazenie
mean((wyniki2a5[2,] - 120)^2) # blad sredniokwadratowy

wyniki2a1 = matrix(0, 2, SymNumer)
for(index in 1:SymNumer){
  cecha2_1 = beta1 * Y[,80] + beta2 * Y[,120] + gamma1*Y
      [,80]*Y[,120] + rnorm(500)
  wyniki2a1[,index] = szukanie_pary2(cecha2_1, Y, L)
}
# 80cM
sd(wyniki2a1[1,]) # odchylenie standardowe
mean(wyniki2a1[1,]) - 80 # obciazenie
mean((wyniki2a1[1,] - 80)^2) # blad sredniokwadratowy

# 120cM
sd(wyniki2a1[2,]) # odchylenie standardowe
mean(wyniki2a1[2,]) - 120 # obciazenie
mean((wyniki2a1[2,] - 120)^2) # blad sredniokwadratowy
```

3.2 Podpunkt b)

Aby wyznaczyć obciążenie, odchylenie standardowe oraz błąd średniokwadratowy wyznaczonych estymatorów położeń doświadczenie związane z dopasowywaniem błędnego modelu addytywnego dla każdej pary markerów powtórzono 1000 razy.

Uzyskane wyniki:

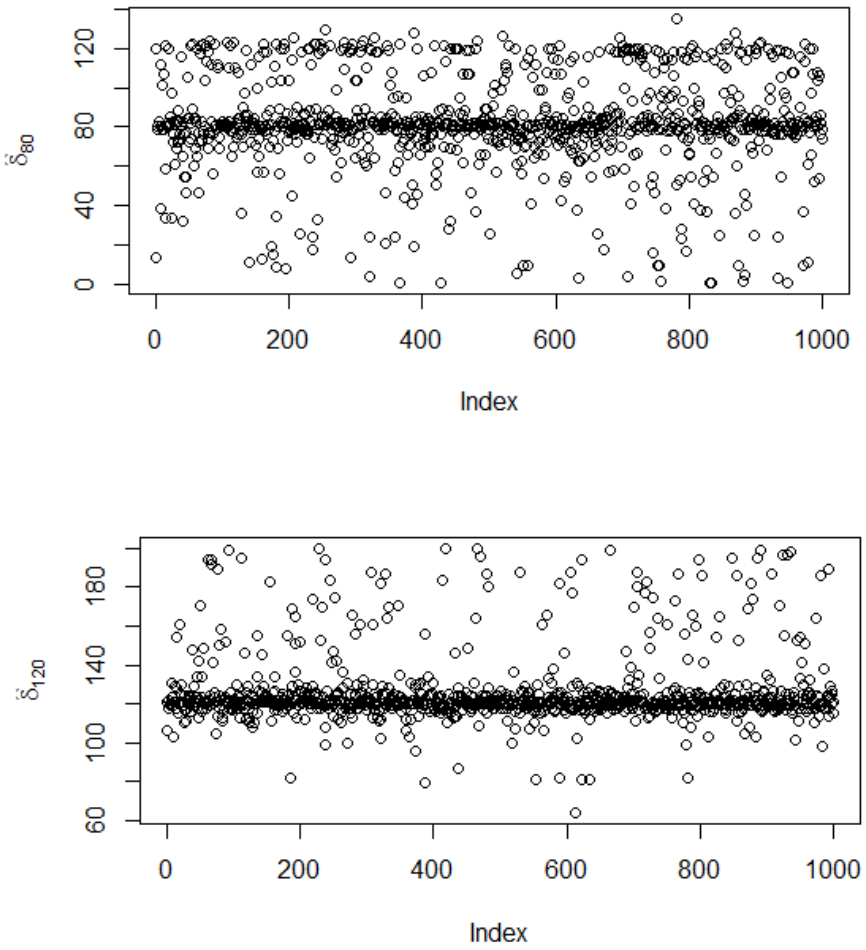
γ	δ	$\sigma(\hat{\delta})$	$E(\hat{\delta})-\delta$	$E(\hat{\delta}-\delta)^2$
0.5	80	21.03705	2.006	581.226
	120	17.81719	5.556	348.004
1	80	23.47258	1.818	553.716
	120	17.36205	4.356	320.114



Rysunek 8: Wyestymowane na podstawie 1000 symulacji położenia obu genów mających wpływ na cechę dla $\gamma = 1$.

Jak można zauważyć na wykresach 8 i 9 większość wyznaczonych estymatorów poprawnie wskazuje na geny skorelowane z cechą, jednakże tym razem oszacowane parametry mają większy rozrzut, niż w przypadku, w którym lokalizacji dokonywaliśmy posługując się poprawnym modelem z interakcją.

Oprócz uzyskanych znacznie większych wartości odchyleń standardowych mamy do czynienia z dużymi wartościami błędu średniokwadratowego. Świadczy to o niezbyt dobrym dopasowaniu modelu do danych.



Rysunek 9: Wyestymowane na podstawie 1000 symulacji położenia obu genów mających wpływ na cechę dla $\gamma = 0.5$.

Kod źródłowy zadania 2 podpunktu b) w języku R:

```
# b)
wyniki2b5 = matrix(0, 2, SymNumer)
for(index in 1:SymNumer){
  cecha2_5 = beta1 * Y[,80] + beta2 * Y[,120] + gamma5*Y
      [,80]*Y[,120] + rnorm(500)
  wyniki2b5[,index] = szukanie_pary(cecha2_5, Y, L)
}
# 80cM
sd(wyniki2b5[1,]) # odchylenie standardowe
mean(wyniki2b5[1,]) - 80 # obciazenie
mean((wyniki2b5[1,] - 80)^2) # blad sredniokwadratowy

# 120cM
sd(wyniki2b5[2,]) # odchylenie standardowe
mean(wyniki2b5[2,]) - 120 # obciazenie
mean((wyniki2b5[2,] - 120)^2) # blad sredniokwadratowy

wyniki2b1 = matrix(0, 2, SymNumer)
for(index in 1:SymNumer){
  cecha2_1 = beta1 * Y[,80] + beta2 * Y[,120] + gamma1*Y
      [,80]*Y[,120] + rnorm(500)
  wyniki2b1[,index] = szukanie_pary(cecha2_1, Y, L)
}
# 80cM
sd(wyniki2b1[1,]) # odchylenie standardowe
mean(wyniki2b1[1,]) - 80 # obciazenie
mean((wyniki2b1[1,] - 80)^2) # blad sredniokwadratowy

# 120cM
sd(wyniki2b1[2,]) # odchylenie standardowe
mean(wyniki2b1[2,]) - 120 # obciazenie
mean((wyniki2b1[2,] - 120)^2) # blad sredniokwadratowy
```

Literatura

[1] M. Bogdan, *Notatki z wykładów Statystyki w zastosowaniach*, Uniwersytet Wrocławski, Wrocław 2016.

[2] K. Dyba, *Notatki z laboratoriów Statystyki w zastosowaniach*, Uniwersytet Wrocławski, Wrocław 2016.

[3] P. Szulc, *Localization of genes*, Mathematica Applicanda Vol. 43(1) 2015, p. 19-35.