BSG-MDS Practical 3 Statistical Genetics

Biel Caballero and Gerard Gomez

05/12/2023, submission deadline 12/12/2023

Population substructure

1. Perform the alleles test for this data set. Provide the p-value and the odds ratio and comment on the results.

```
df <- matrix(c(112,278,150,206,348  ,150),byrow=TRUE,ncol=3)
colnames(df) <- c("AA","Aa","aa")
rownames(df) <- c("Cases","Controls")
df
```

```
##           AA  Aa  aa
## Cases    112 278 150
## Controls 206 348 150
```

```
alleles<-cbind(2*df[,1]+df[,2],2*df[,3]+df[,2])
colnames(alleles) <- c("A","a")
fisher.test(alleles)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  alleles
## p-value = 0.000231
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6295932 0.8709495
## sample estimates:
## odds ratio
##  0.7406221
```

The p-value is 0.000231 and the odds ratio is 0.7406221. This means that we accept the alternative hypothesis meaning that the true odds ratio is not equal to 1. Therefore we can say that there is a trend in the proportions of the SNP and the Alzheimer's disease.

2. Test for association using a codominant, a dominant and a recessive model. Provide the p-values for all the tests and comment on the results.

```
#Codominant
codominant <- matrix(c(112,278,150,206,348,150),byrow=TRUE,ncol=3)
colnames(codominant) <- c("AA","AB","BB")
rownames(codominant) <- c("Cases","Controls")
codominant
```

```
##           AA  AB  BB
## Cases    112 278 150
## Controls 206 348 150
```

```
results<-chisq.test(codominant)
results
```

```
##
##  Pearson's Chi-squared test
##
## data:  codominant
## X-squared = 14.241, df = 2, p-value = 0.0008085
```

```
results$expected
```

```
##                 AA       AB       BB
## Cases    138.0386 271.7363 130.2251
## Controls 179.9614 354.2637 169.7749
```

```
fisher.test(codominant)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  codominant
## p-value = 0.0007783
## alternative hypothesis: two.sided
```

```
#Dominant
dominant <- cbind(codominant[,1],codominant[,2]+codominant[,3])
colnames(dominant) <- c("AA","AB or BB")
rownames(dominant) <- c("Cases","Control")
dominant
```

```
##          AA AB or BB
## Cases   112      428
## Control 206      498
```

```
results<-chisq.test(dominant)
results
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dominant
## X-squared = 11.216, df = 1, p-value = 0.0008108
```

```
results<-chisq.test(codominant, correct = FALSE)
results
```

```
##
##  Pearson's Chi-squared test
##
## data:  codominant
## X-squared = 14.241, df = 2, p-value = 0.0008085
```

```
#Recessive
recessive <- cbind(codominant[,1]+codominant[,2],codominant[,3])
colnames(recessive) <- c("AA or AB","BB")
rownames(recessive) <- c("Cases","Control")
recessive
```

```
##          AA or AB  BB
## Cases         390 150
## Control       554 150
```

```r
results<-chisq.test(recessive)
results
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  recessive
## X-squared = 6.6433, df = 1, p-value = 0.009953
```

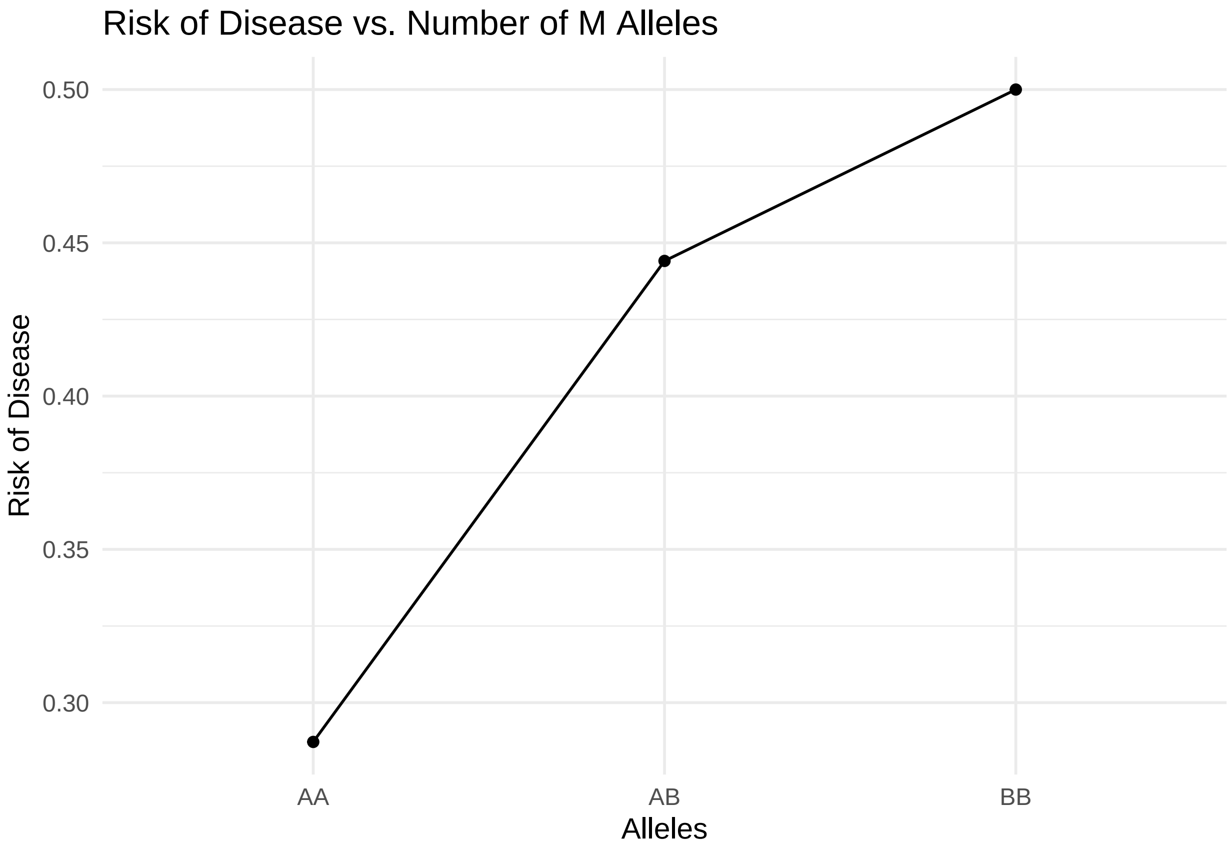
```r
fisher.test(recessive)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  recessive
## p-value = 0.009094
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5377586 0.9218594
## sample estimates:
## odds ratio
##  0.7041877
```

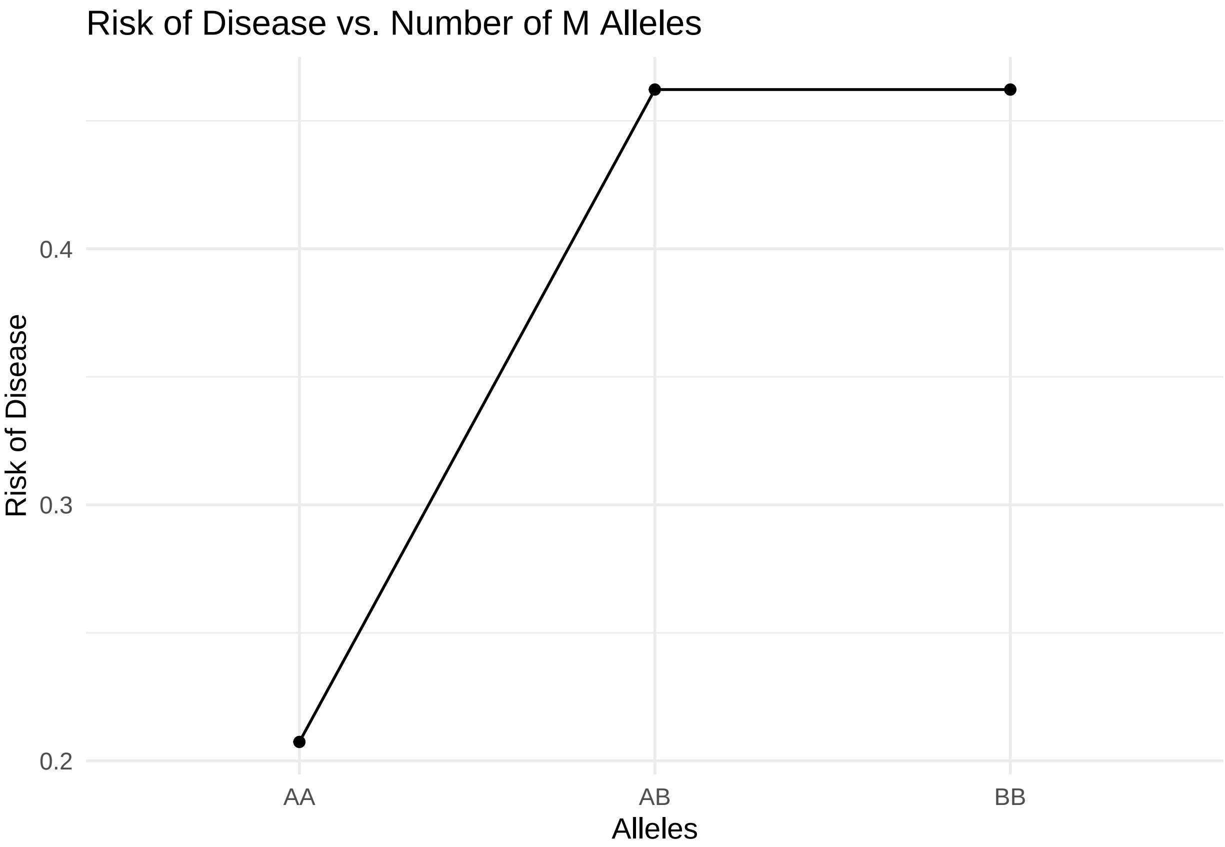
The results for the Codominant case, show that we reject the null hypothesis that the probability of disease with all the genotypes is the same.
The results for the Dominant case, show that we reject the null hypothesis that the probability of disease does not depend on B.
The results for the Codominant case, show that we reject the null hypothesis that the probability of disease does not depend on being homozygote BB.

3. Plot the risk of disease as a function of the number of m alleles. Comment on the results. Which model seems most appropriate?

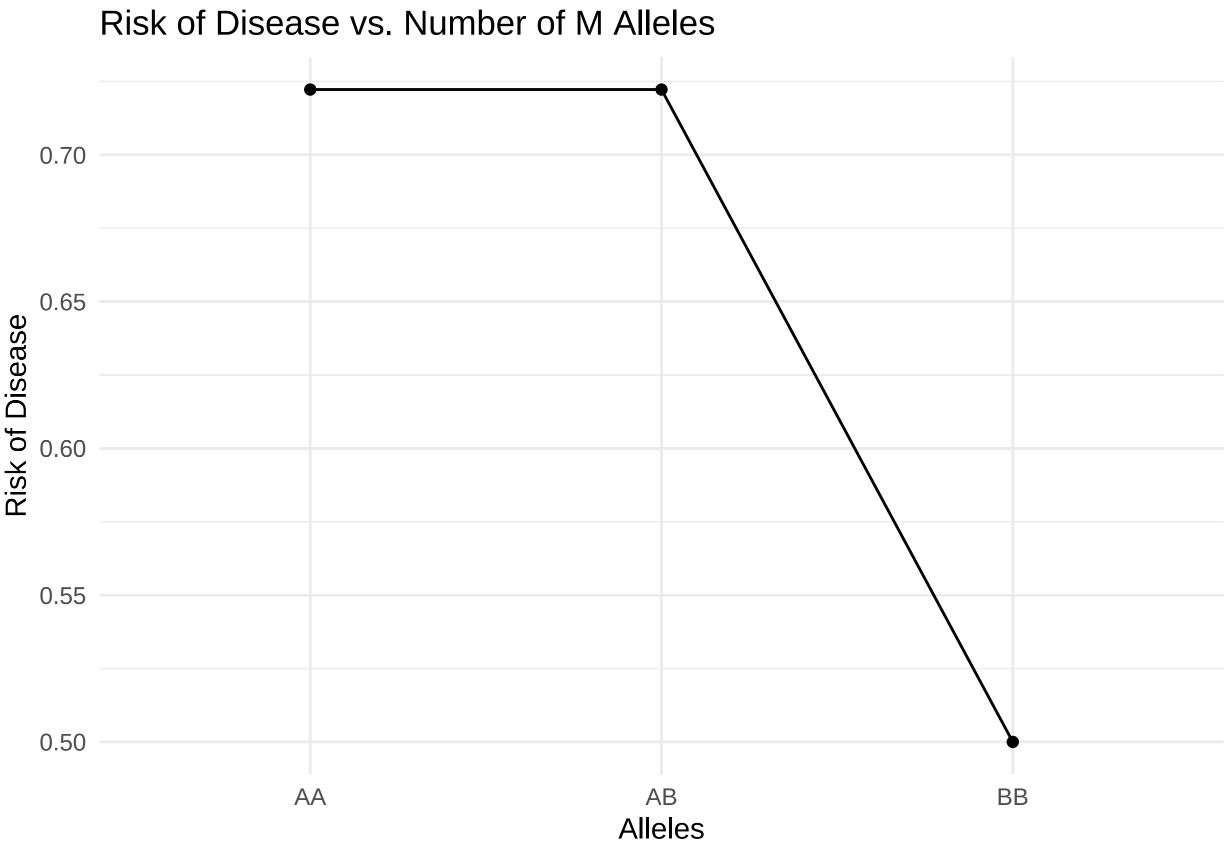
```r
#Codominant
m_alleles <- c("AA", "AB", "BB")
risk_of_disease <- c(codominant[1,1]/(codominant[1,1]+codominant[1,2]), codominant[1,2]/(codominant[1,2

data <- data.frame(m_alleles, risk_of_disease)

ggplot(data, aes(x = m_alleles, y = risk_of_disease,group = c(1,1,1))) +
  geom_point() + geom_line() +
  labs(x = "Alleles", y = "Risk of Disease") +
  ggtitle("Risk of Disease vs. Number of M Alleles") +
  theme_minimal()
```



```
#Dominant
m_alleles <- c("AA", "AB", "BB")
risk_of_disease <- c(dominant[1,1]/(dominant[1,1]+dominant[1,2]), dominant[1,2]/(dominant[1,2]+dominant

data <- data.frame(m_alleles, risk_of_disease)

ggplot(data, aes(x = m_alleles, y = risk_of_disease,group = c(1,1,1))) +
  geom_point() +
  geom_line() +
  labs(x = "Alleles", y = "Risk of Disease") +
  ggtitle("Risk of Disease vs. Number of M Alleles") +
  theme_minimal()
```



```
#Recessive
m_alleles <- as.factor(c("AA", "AB", "BB"))
risk_of_disease <- c(recessive[1,1]/(recessive[1,1]+recessive[1,2]), recessive[1,1]/(recessive[1,1]+rec

data <- data.frame(m_alleles, risk_of_disease)

ggplot(data, aes(x = m_alleles, y = risk_of_disease,group = c(1,1,1))) +
  geom_point() +
  geom_line() +
  labs(x = "Alleles", y = "Risk of Disease") +
  ggtitle("Risk of Disease vs. Number of M Alleles") +
  theme_minimal()
```



In these plots we can see the risk of having the disease given the alleles they have in the genotype. The model that seems to be the most appropiate would be the codominant model, because it makes sense regarding the number for each genotype to have these amount of risk of having the disease. The other two plots, have the same probability for AB and BB (in the case of dominant) and AA and AB (in the case of recessive), because we grouped them in the previous exercise as the same outcome, so it was easier to work with.

4. Perform Armitage trend test for this data set. Does the null hypothesis beta1 = 0 hold? Comment on your response.

```
n <- sum(df)

cas <- rep(c(0,1,2), df[1,])
con <- rep(c(0,1,2), df[2,])

y <- c(rep(1,sum(df[1,])),rep(0,sum(df[2,])))
x <- c(cas,con)

r <- cor(x,y)
A <- n*(r^2)

pval <- pchisq(A,df=1,lower.tail = FALSE)
pval
```

```
## [1] 0.0002000008
```

No, the null hypothesis doesn't hold. This means that there is a a trend in the proportions of the SNP and the Alzheimer's disease

5. Is there evidence for association of this marker with the disease? Argument your response

We can say there is a clear association between the marker and the disease. This can be clearly seen in the after the allele and Armitage test have been performed, where we got a pvalue lower than 0.05 for both tests, this means that there is a trend in the proportions of the SNP with the Alzheimers