Genomic_Data

Gautham Meenakshisundaram

3/4/2020

Introduction

Based on multiple studies, the researchers looked at some genes at a certain genomic region. They hypothesised that 5 genes could be potentially involved in weight gain of animals. The researchers generated an experimental design with 10 heterozygote sires, where each sire had 40 offspring with randomly selected females that had a matching genetic background. The sires and the offspring were genotyped for all 5 markers and phenotypic measures were recorded.

We need to test association between these markers and the phenotypes.

First Dataset

```
data1 = read.table("siredata01.txt", header=T, sep="\t")
data1
```

```
##        id weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 1   sire1 334.14  M2  M1  M3  M2  M3  M4  M4  M2  M4  M2
## 2   sire2 364.81  M3  M2  M3  M2  M2  M3  M2  M4  M3  M1
## 3   sire3 383.95  M2  M4  M2  M4  M3  M2  M3  M4  M1  M4
## 4   sire4 349.88  M2  M1  M1  M2  M4  M3  M2  M1  M4  M3
## 5   sire5 357.87  M1  M3  M2  M1  M3  M1  M3  M4  M3  M2
## 6   sire6 364.87  M3  M2  M1  M3  M4  M2  M3  M1  M1  M3
## 7   sire7 361.36  M4  M1  M4  M1  M2  M1  M1  M2  M2  M4
## 8   sire8 357.56  M3  M4  M2  M4  M4  M1  M2  M4  M1  M3
## 9   sire9 333.49  M1  M2  M4  M3  M3  M2  M1  M2  M3  M2
## 10 sire10 360.92  M2  M3  M3  M1  M3  M2  M2  M3  M3  M2
```

```
dim(data1)
```

```
## [1] 10 12
```

There are no missing values in the data.

```
summary(data1)
```

```
##        id          weight        m11    m12    m21    m22    m31    m32
##  sire1  :1   Min.   :333.5   M1:2   M1:3   M1:2   M1:3   M2:2   M1:3
##  sire10 :1   1st Qu.:351.8   M2:4   M2:3   M2:3   M2:3   M3:5   M2:4
##  sire2  :1   Median :359.4   M3:3   M3:2   M3:3   M3:2   M4:3   M3:2
##  sire3  :1   Mean   :356.9   M4:1   M4:2   M4:2   M4:2          M4:1
##  sire4  :1   3rd Qu.:363.9
##  sire5  :1   Max.   :383.9
##  (Other):4
##  m41    m42    m51    m52
##  M1:2   M1:2   M1:3   M1:1
##  M2:4   M2:3   M2:1   M2:4
##  M3:3   M3:1   M3:4   M3:3
##  M4:1   M4:4   M4:2   M4:2
##
##
##
```

The data looks fine as for each marker there are two alleles. Each allele can be of M1, M2, M3 or M4 type.

Second Dataset

```
data2 = read.table("progdata01.txt", header=T, sep="\t")
head(data2)

##    id  sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 1 id1 sire1   F 293.61  M5  M6  M2  M5  M3  M6  M2  M5  M4  M4
## 2 id2 sire1   M 335.43  M1  M4  M3  M1  M4  M5  M4  M3  M2  M2
## 3 id3 sire1   M 340.09  M2  M3  M2  M6  M3  M1  M4  M3  M4  M3
## 4 id4 sire1   M 343.08  M2  M3  M2  M1  M4  M6  M2  M3  M4  M5
## 5 id5 sire1   F 410.08  M1  M3  M2  M5  M3  M2  M2  M4  M4  M4
## 6 id6 sire1   F 302.17  M2  M2  M2  M5  M3  M5  M2  M4  M4  M1

dim(data2)

## [1] 400  14
```

The dataset contains 400 records (10 sires × 40 offspring) but also has missing values.

Searching for missing genotypes using ‘grep’ which searches for pattern matches.

```
index=grep("m",names(data2))
missing=numeric()
for (i in 1:length(index))
  missing=c(missing,which(data2 [,index[i]]=="-"))
print(data2[missing,])

##          id   sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 70     id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301   id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367   id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.1   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.1 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.1 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.2   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.2 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.2 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.3   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.3 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.3 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.4   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.4 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.4 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.5   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.5 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.5 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.6   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.6 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.6 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.7   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.7 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.7 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.8   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.8 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.8 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
## 70.9   id70  sire2   M 372.45   -   -   -   -   -   -   -   -   -   -
## 301.9 id301  sire8   F 305.34   -   -   -   -   -   -   -   -   -   -
## 367.9 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -
```

When we print the indices ‘missing’, we see that ‘id70’, ‘id301’, ‘id367’ repeated across all the markers which we are removing now.

```
data2[which(data2$id=="id70"),]

##      id  sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 70 id70 sire2   M 372.45   -   -   -   -   -   -   -   -   -   -

data2=data2[-which(data2$id=="id70"),]
data2[which(data2$id=="id301"),]

##        id  sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 301 id301 sire8   F 305.34   -   -   -   -   -   -   -   -   -   -

data2=data2[-which(data2$id=="id301"),]
data2[which(data2$id=="id367"),]

##        id   sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 367 id367 sire10   F 313.62   -   -   -   -   -   -   -   -   -   -

data2=data2[-which(data2$id=="id367"),]

data2$id=as.character(data2$id)
```

Searching for missing values under ‘Weight’ column and removing them.

```
data2[which(data2$weight=="-"),]
```

```
##      id  sire sex weight m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
## 8   id8 sire1   M      -  M1  M1  M3  M6  M3  M4  M4  M2  M2  M6
## 90 id90 sire3   M      -  M4  M1  M2  M5  M2  M5  M4  M2  M1  M1
## 94 id94 sire3   F      -  M4  M2  M4  M4  M2  M4  M4  M3  M4  M5

data2=data2[-which(data2$weight=="-"),]

data2$weight=as.numeric(as.character(data2$weight))
```

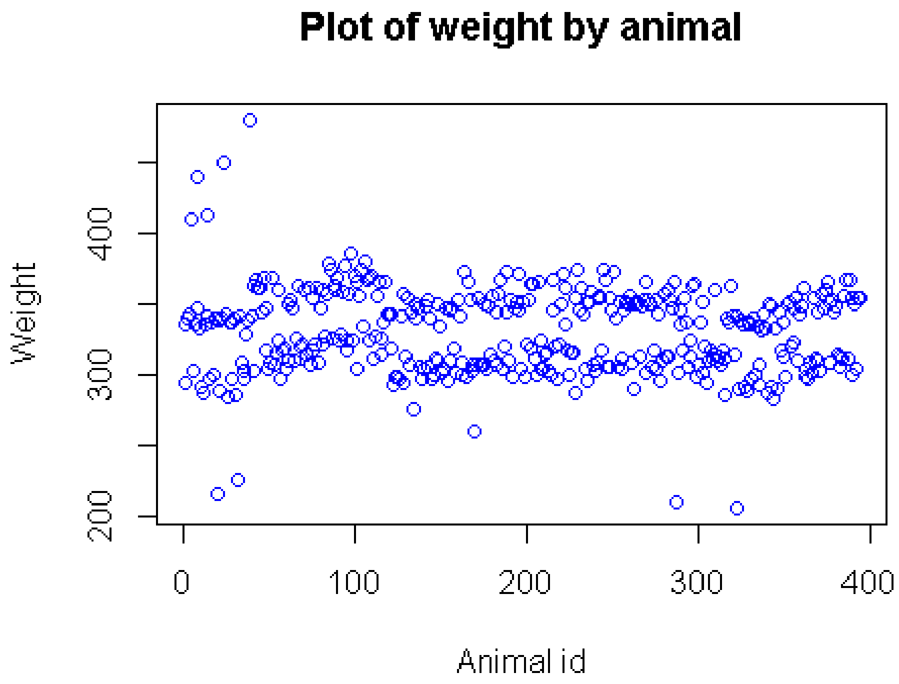
The updated shape of the second dataset is shown below.

```
dim(data2)

## [1] 394  14

summary(data2)

##       id                  sire      sex         weight       m11       m12
##  Length:394          sire4  : 40   F:197   Min.   :205.7   - :  0   - : 0
##  Class :character    sire5  : 40   M:197   1st Qu.:307.2   M1: 97   M1:67
##  Mode  :character    sire6  : 40           Median :334.1   M2:135   M2:66
##                      sire7  : 40           Mean   :330.4   M3: 99   M3:68
##                      sire9  : 40           3rd Qu.:352.2   M4: 58   M4:63
##                      sire1  : 39           Max.   :480.5   M5:  5   M5:73
##                      (Other):155                                    M6:57
##  m21      m22     m31      m32     m41      m42     m51      m52
##  - :  0   - : 0   - :  0   - : 0   - :  0   - : 0   - :  0   - : 0
##  M1: 94   M1:62   M1: 54   M1:65   M1: 76   M1:56   M1: 81   M1:57
##  M2:129   M2:60   M2:112   M2:68   M2:142   M2:62   M2: 85   M2:62
##  M3: 89   M3:78   M3:134   M3:70   M3: 81   M3:76   M3:133   M3:59
##  M4: 78   M4:58   M4: 88   M4:58   M4: 89   M4:77   M4: 89   M4:75
##  M5:  4   M5:59   M6:  6   M5:72   M5:  6   M5:71   M5:  6   M5:73
##           M6:77            M6:61            M6:52            M6:68
```

The data looks fine as for each marker there are two alleles. Each allele can be of M1, M2, M3, M4, M5 or M6 type.

Visualization

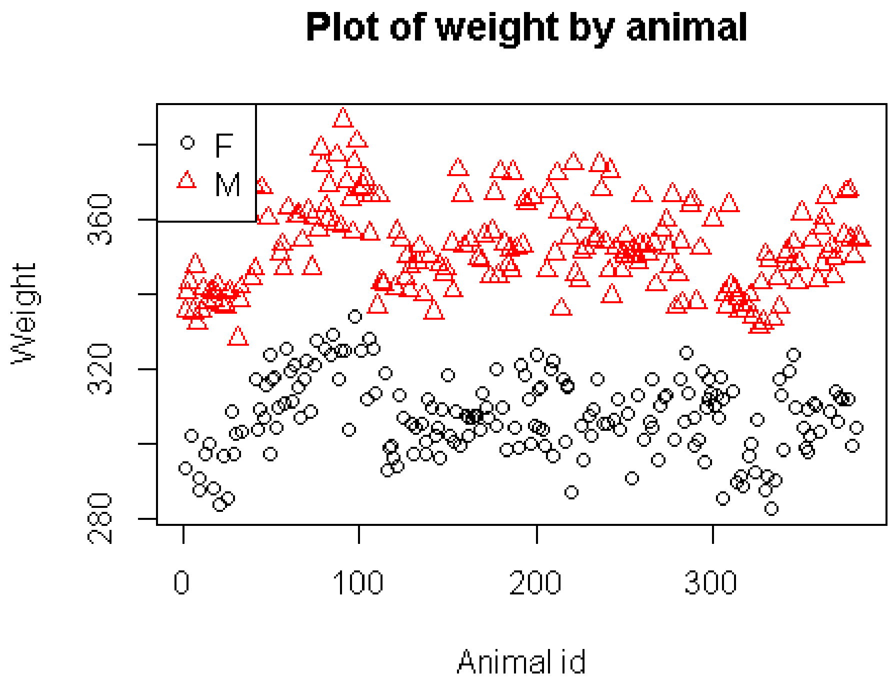
```
plot(data2$weight,main="Plot of weight by animal", xlab="Animal id",
ylab="Weight",col="blue")
```



Outliers are seen above the weight of 400 and below the weight of 280 along with two groups of animals in the plot.

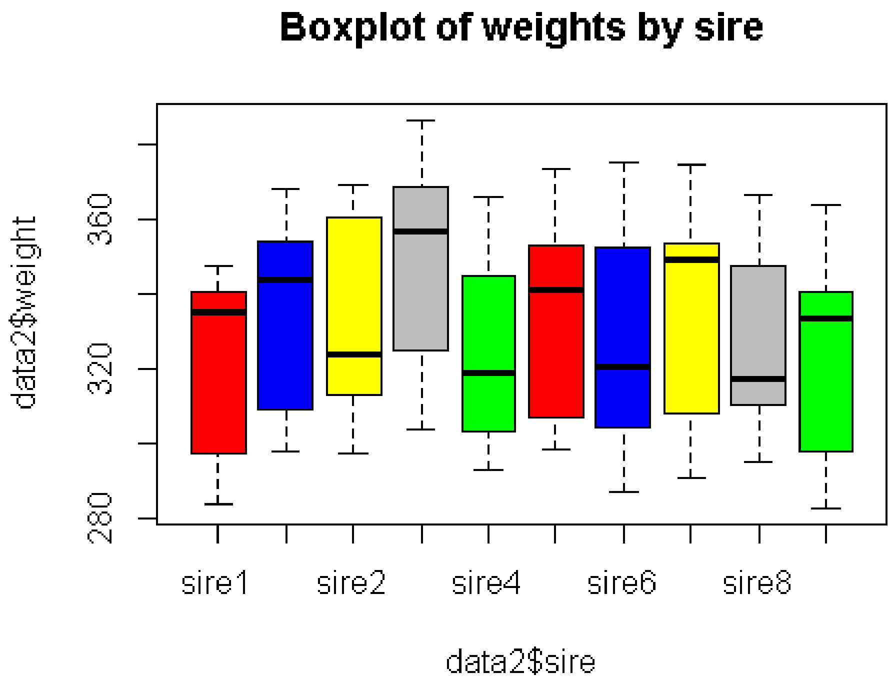
Outliers seen above the weight of 400 and below the weight of 280 are removed using ‘which’.

```r
data2=data2[-which(data2$weight>400),]
data2=data2[-which(data2$weight<280),]

plot(data2$weight,col=data2$sex, pch=as.numeric(data2$sex), main=" Plot of weight by animal", xlab="Animal id",ylab="Weight")
legend("topleft",levels(data2$sex),col=1:2,pch=1:2)
```



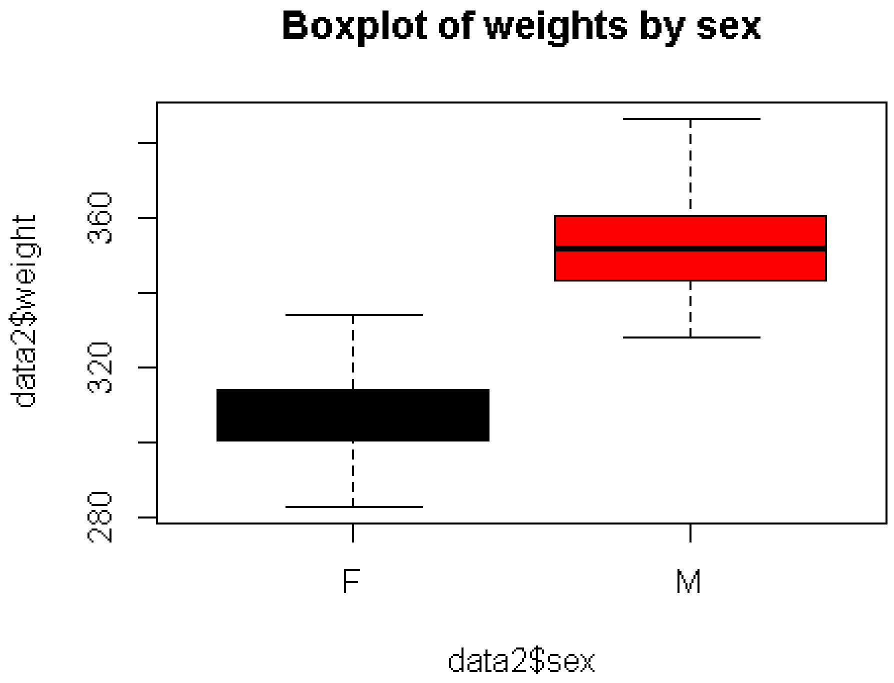
Outliers are removed and only two groups of animals are seen.

```
boxplot(data2$weight~data2$sire, col = c("Red","Blue","yellow","grey","green"
), main="Boxplot of weights by sire")
```



Boxplot of weights by sire

```
boxplot(data2$weight~data2$sex, col=1:length(levels(data2$sex)),
main="Boxplot of weights by sex")
```



Boxplot of weights by sex

The boxplots in the second boxplot figure shows that ‘sex’ factor is splitting the weight data into two groups whereas the boxplots in the first boxplot figure illustrate a varying weights for sires.