yw867_Midterm

Yuanyuan Wu

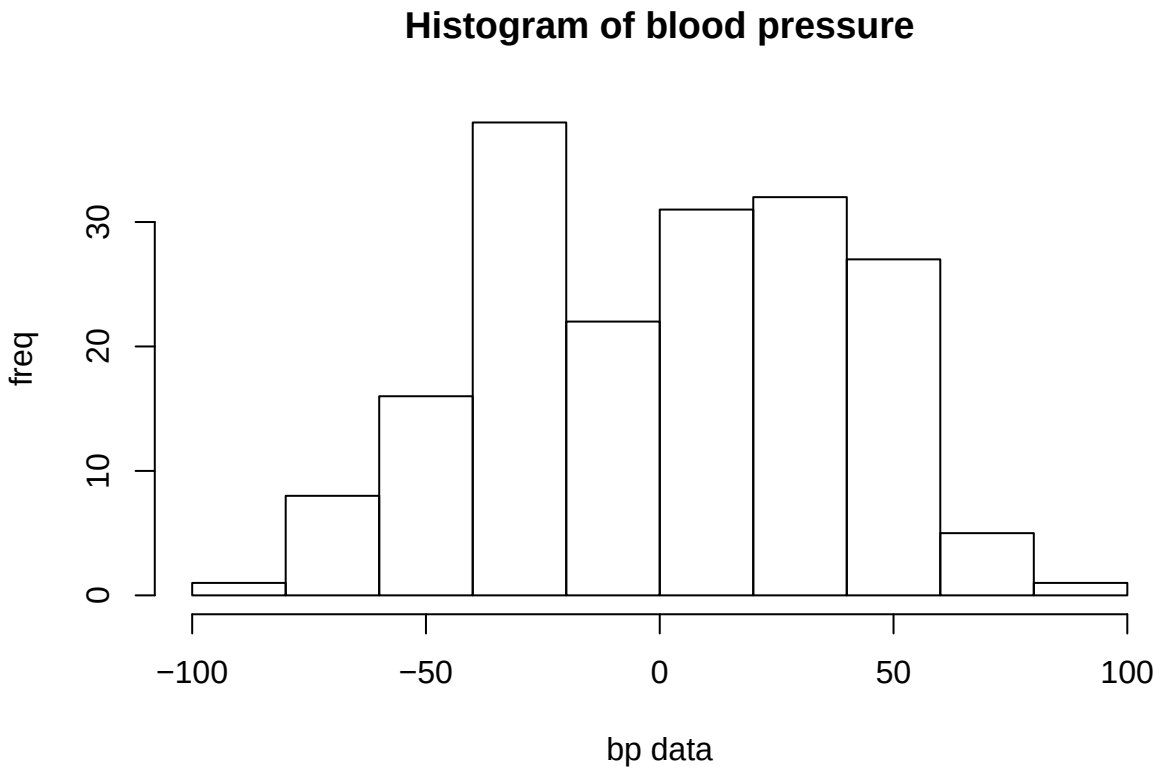
4/14/2019

1.

```
#import id and phenotype(bp)
id_bp <- read.csv('midterm2019_pheno+pop.csv',header = FALSE)

#b.calculate and report total sample size n
n <-  nrow(id_bp)
(n)
```

```
## [1] 181
```

```
#c. plot histogram, label
hist(id_bp[,2],xlab = 'bp data',ylab = 'freq',main = paste("Histogram of" , 'blood pressure'))
```



1d.

The shape of the histgram has two peaks while the 'normal distribution' has only one peak. The two peaks probably result from the two populations with differenr means that are mixed together. The two mixed populations are likely of normal distribution seperately and linear regression with covariate can be used.

1e.

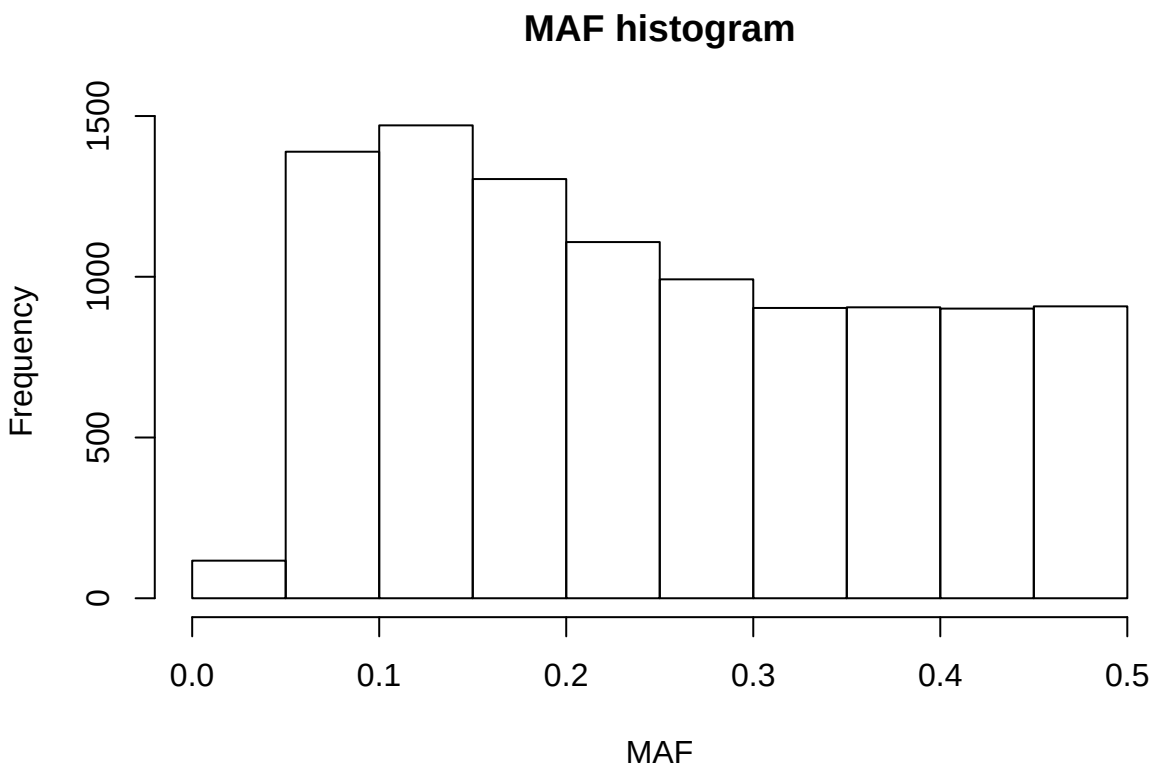
The phenotype(blood pressure) is not case/control situation with multiple scattered values.

2.

```
#a. import genotype data
geno <- read.csv('midterm2019_genotypes_v2.csv',header = FALSE)

#b. number of SNPs:N
N <- ncol(geno)
(N)
```

```
## [1] 9998
```

```
#c.MAF
#calculate all the allefrequency for the allele represented by '0'
#initiate a1 to store freq of allele '0'
a1 <- rep(0,N)
a1 <- (colSums(geno==0)*2+colSums(geno==1))/(2*n)

#change into '2' as minor allele if frequency >0.5

a2_ind <- which(a1>0.5)

a1[a2_ind] <- 1-a1[a2_ind]

#eliminate <- which(a1<0.1)
#hist MAF
hist(a1,main = 'MAF histogram',xlab = 'MAF')
```



2d.

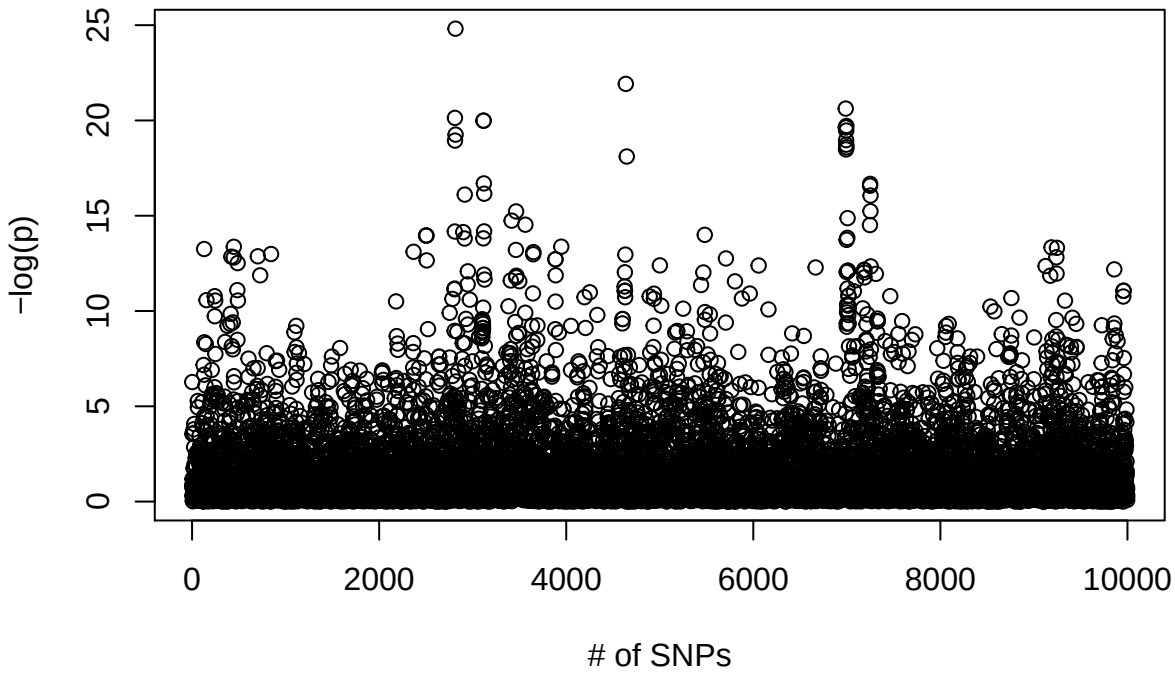
'power' of a hypothesis test: the probability of the test that rejects the NULL when the alternative hypothesis is true. Type 1 error is the probability of incorrectly rejecting the null hypothesis when it is correct. ###2e. if the MAF of a specific SNP differ in the two populations, the p of the SNP is likely low, which result in the rejection of the NULL when it is not a causal SNP(false positive)

3

```
#3a
#calculcate p, without cvariate
#geno to XaXd
XaXd <- function(geno){
  #number of samples
  n <- nrow(geno)
  #number of SNPs
  N <- ncol(geno)

  Xa <- matrix(NA,nrow = n,ncol = N)
  #initializa Xd with -1s for homozygotes
  Xd <- matrix(-1,nrow = n,ncol = N)

  #assign values to the heterozygotes
  Xa[which(geno==1)] <- 0
  Xd[which(geno==1)] <- 1

  ##assume 0 is the homozygote for minor allele
```

```
  Xa[which(geno==0)] = 1
  Xa[which(geno==2)] = -1

  #calculate all the allefrequency for the allele represented by '0'
  #initiate a1 to store freq of allele '0'
  a1 <- rep(0,N)
  a1 <- (colSums(geno==0)*2+colSums(geno==1))/(2*n)

  #change into '2' as minor allele if frequency >0.5
  a2_ind <- which(a1>0.5)

  ##correct by allele frequency, if 2 is the minor allele
  Xa[,a2_ind] <- -Xa[,a2_ind]

  return(list(Xa=Xa,Xd=Xd))

}

library(MASS)
library(ggplot2)

pval_calculator <- function(pheno_input, xa_input, xd_input){

  n_samples <- length(xa_input)

  X_mx <- cbind(1,xa_input,xd_input)

  MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% pheno_input
  y_hat <- X_mx %*% MLE_beta

  SSM <- sum((y_hat - mean(pheno_input))^2)
  SSE <- sum((pheno_input - y_hat)^2)
  df_M <- 2
  df_E <- n_samples - 3

  MSM <- SSM / df_M
  MSE <- SSE / df_E

  Fstatistic <- MSM / MSE

  pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)

  return(pval)
}




xaxd <- XaXd(geno)
Xa <- xaxd$Xa
Xd <- xaxd$Xd
```

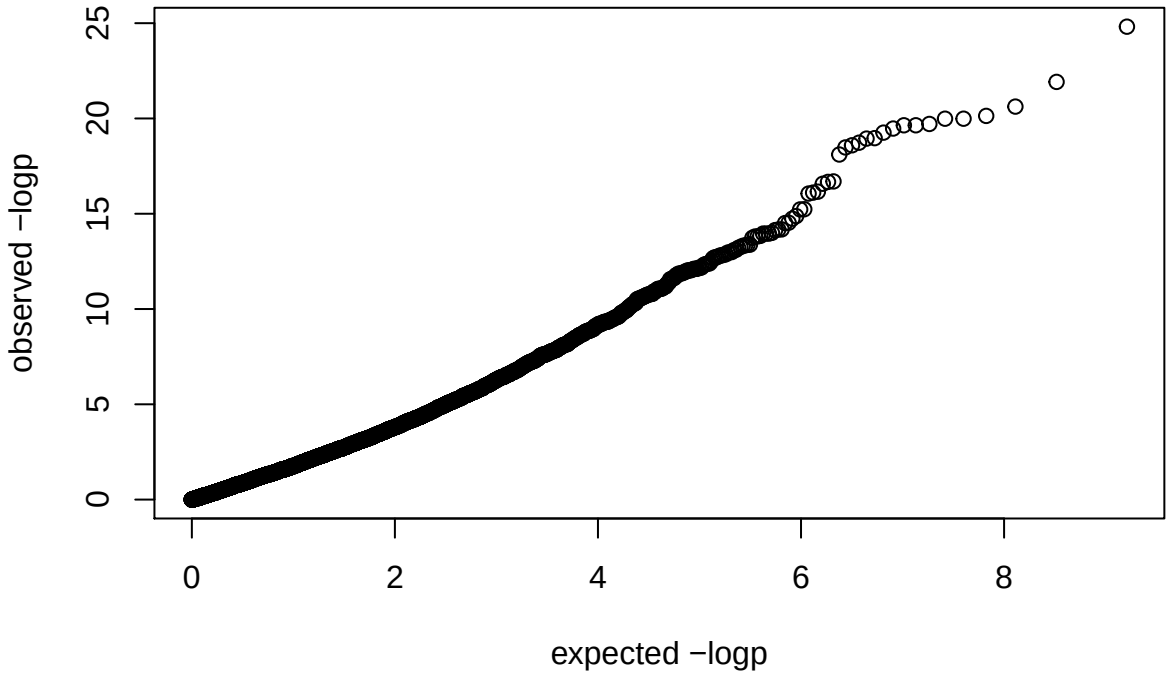
```
pval_mx <- rep(0,N)
    for (i in 1 : N){
        pval_mx[i] <- pval_calculator(id_bp[,2], Xa[,i], Xd[,i])
    }



#3b.
plot(-log(pval_mx),ylab = '-log(p)',xlab = '# of SNPs')
```



##4

QQ plot

```
Unif = seq(0,1,length.out = N+1)[-1]
p_Unif = -log(Unif)
#order from smallest to largest for p and p_Unif
p_Unif <- sort(p_Unif)
logp2 <- sort(-log(pval_mx))
plot(p_Unif,logp2,xlab = 'expected -logp',ylab = 'observed -logp')
```



4b.

The shape of the QQ plot is quite linear and this indicates appropriate model fits for the expected p and observed p.

5

```
#a. report n1,n2
pop1 <- id_bp[which(substr(id_bp[,1],1,2)=='HG'),]
pop2 <- id_bp[which(substr(id_bp[,1],1,2)=='NA'),]
n1 <- nrow(pop1)
n2 <- nrow(pop2)
(c(n1,n2))
```

```
## [1] 89 92
```

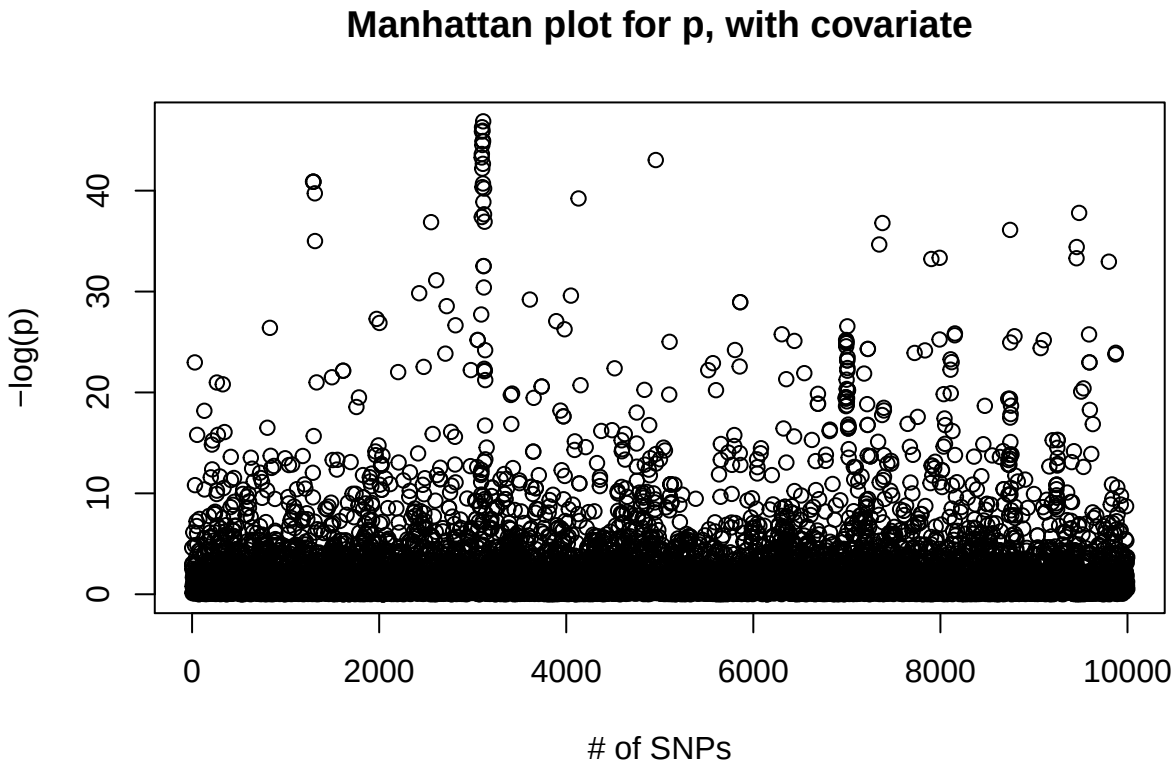
5b.

a PCA will tell the major differnces factors in the population and grouping the population. Population of differnt ancestries can be distinguished with PCA since they ar e likely to be distinguished in the PCA.

6

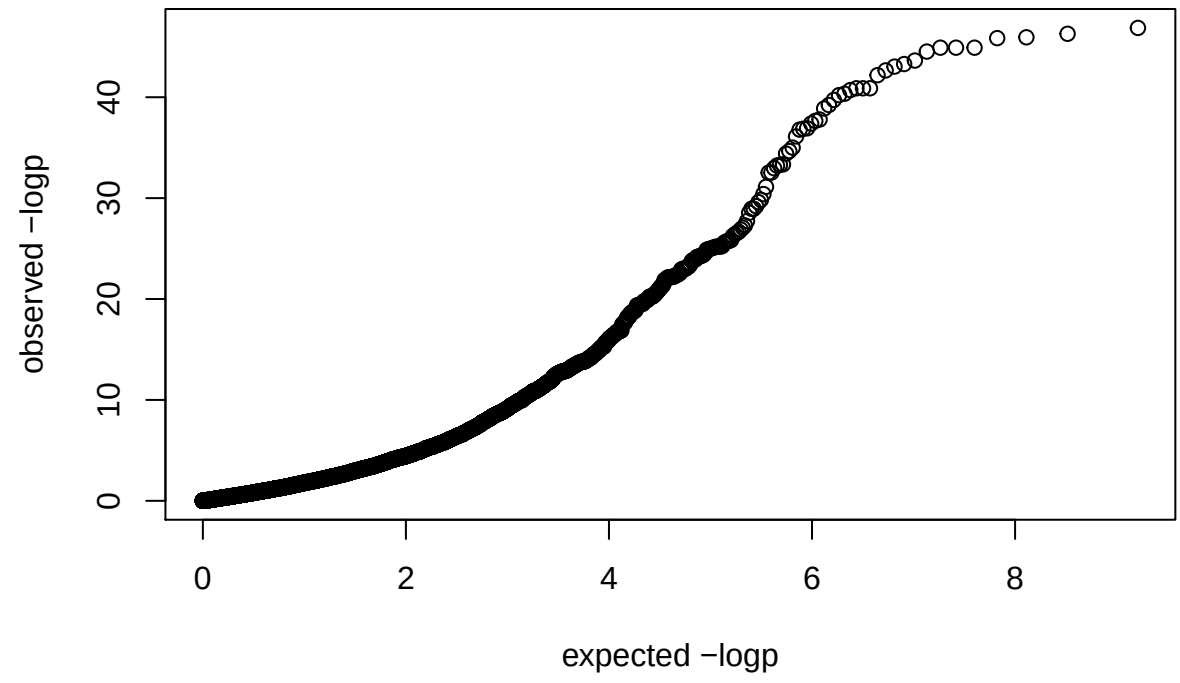
```
#linear regression with covariates
#set up Xz
Xz <- matrix(1,nrow = n, ncol = 1)
#n1 population, Xz=-1
Xz[which(substr(id_bp[,1],1,2)=='HG'),1] <- -1
```

```r
pval_calculator_xz <- function(pheno_input, xa_input, xd_input,xz){

  n_samples <- length(xa_input)

  X_mx <- cbind(1,xa_input,xd_input,xz)

  MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% pheno_input
  y_hat1 <- X_mx %*% MLE_beta
  y_hat0 <- MLE_beta[1] + xz*MLE_beta[4]

  SSE0 <- sum((pheno_input-y_hat0)^2)
  SSE1 <- sum((pheno_input - y_hat1)^2)


  df_M <- 2
  df_E <- n_samples - 3

  Fstatistic <- ((SSE0-SSE1)/df_M) / (SSE1 / df_E)

  pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)

  return(pval)
}
p_xz <-  rep(0,N)
for (i in 1 : N){
    p_xz[i] <- pval_calculator_xz(id_bp[,2], Xa[,i], Xd[,i],Xz)
  }

#6b. Manhattan plot
plot(-log(p_xz),ylab = '-log(p)',xlab = '# of SNPs',main = 'Manhattan plot for p, with covariate')
```



##7

```
Unif = seq(0,1,length.out = N+1)[-1]
p_Unif = -log(Unif)
#order from smallest to largest for p and p_Unif
p_Unif <- sort(p_Unif)
logp3 <- sort(-log(p_xz))
plot(p_Unif,logp3,xlab = 'expected -logp',ylab = 'observed -logp')
```



7b.

After adding the covariates,the fit gets worse. The covariate chosen specific on the population may not be optimal

8

```
#a.
p_Bon <- 0.05/N

#b. report peaks
hits <- which(-log(p_xz)>(-log(p_Bon)))
#let peaks sit in an approximate of +-250 loci

#import hits that report the index of the SNPs that are significant
#import p values
peaksearch <- function(hits,p){
  #the peak that stores the index of the highet p value
  peak <- c()
  hits_p <- p[hits]

  #find reginal maximum p
  for(i in 2:(length(hits_p)-1)){

    if(hits_p[i]>hits_p[i-1]&&hits_p[i]>hits_p[i+1]){
      peak <- c(peak,which(p==hits_p[i]))
    }

  }
  return(peak)
}

peaks <- peaksearch(hits,-log(p_xz))
```

There are probably 87 seperate peaks across the all the sites. The peak I identified is the reginol maxium across the significant sites

8c

likely less than 1 causal SNPs for 1 seperate peak since lots of them can be either false positve or the non-optimal regression model used. ##8d Due to linkage equilibrium, it is hard to locate the exact site of the causal sites. Also, some other factors including very rare allele, copy numbers, etc. will interfere with the GWAS result. ##9 ###9a the peak i identified is the reginal optimal and thus has the biggest p value in the window

```
(peaks)
```

```
##  [1]  131  265  328  832  995 1312 1495 1784 1973 2003 2203 2428 2555 2610
## [15] 2722 2816 2980 3091 3096 3100 3103 3105 3112 3122 3127 3133 3408 3421
## [29] 3611 3893 4050 4131 4517 4626 4753 4836 4887 4957 5040 5105 5568 5653
## [43] 5804 6040 6304 6352 6439 6691 6992 6994 6996 7005 7008 7015 7020 7184
## [57] 7346 7380 7472 7650 7724 7903 7992 8035 8109 8120 8155 8475 8725 8734
## [71] 8744 8729 8725 8734 8744 8745 8747 8751 8755 8792 9105 9200 9251 9255
```

```
## [85] 9483 9589 9801
```

9b

the significant level of a SNP will be affected by the MAF in the population. In the non-optimal model, it is not a good indicator of the distance to the causal SNP ###9c.

```
#calculate correlation in Xa to the closet SNPs
cor_value <- rep(0,)
```

9d.

The nearby SNPs are linked together and transfered to the descents as trunks.(LD) ##10 ###10a definition of a causal polympphism: A mutation(difference) in the causal mutaion is likely to result in the change of the phenotypes.

10b.

ideal experimenr: when the strong association of the genotype and phenotype is uncovered, it is necessary to mutate a s

10c.

definition of p-val: p-value is defined as the probability, computed under the null hypothesis, that the test statis- tic would be equal to or more extreme than its observed value

10d.

1)A type I error of p will occur based on the definition of p-value 2)complex population structure in the data that are not corrected 3)error in GWAS could occur