Lab 1: Linear models for quantitative genetics

BMI 206

Abolfazl Arab - (or Abe)

10/16/2024

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

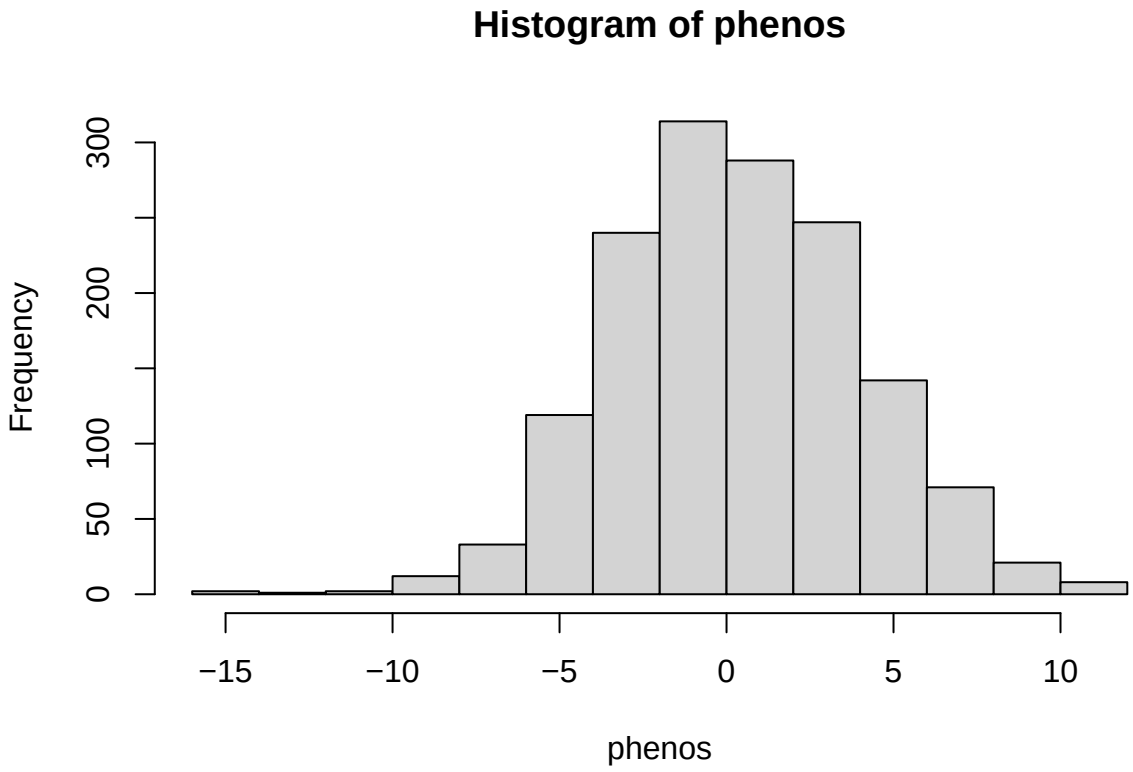
PART1: Analyzing provided genotype and phenotype data.

Prepare the data. Read in the genotype and phenotype matrices.

```
genos = as.matrix(read.table("./genos.txt"))
phenos = as.matrix(read.table("./phenos.txt"))
```

Make a histogram of the phenotypes. Do they look normally distributed?

```
hist(phenos)
```



I think these looks like a normal distribution.

How are the genotypes encoded?

```
table(genos)
```

```
## genos
##       0       1       2
## 4773842 5447131 4779027
```

grouped into "0", "1", and "2"

How many individuals are there in the dataset and how many SNPs? (Save them in `N` and `M`, respectively.)

- N = 1500 # How many individuals
- M = 10000 # how many SNPs

```
dim(genos)
dim(phenos)
N = 1500  # How many individuals
M = 10000 # how many SNPs
```

Compute the *minor* allele frequency for every SNP. Check MAFs are <0.5.

```
MAFs = array(0,M)
for(i in 1:M) {
  ind = table(genos[,i]) %>% data.frame

  freq_A = sum( (ind$Freq / (2 * N) ) * c(0,1,2))

  MAFs[i] = min(c(freq_A, 1 - freq_A))

}
```

```
MAFs[1:10]
max(MAFs)
```

```
MAFs %>% length
```

Run a GWAS under an additive model and save the p-values, z-scores, and effect sizes.

z-scores: estimate divided by standard-deviation

```
pvalues = array(0,M)
zscores = array(0,M)
betas = array(0,M)

for(i in 1:M) {
    g = genos[,i]
    res = summary(lm(phenos~g))
    zscores[i] = res$coefficients[2,'t value']
    pvalues[i] = res$coefficients[2,'Pr(>|t|)']
    betas[i]   = res$coefficients[2,'Estimate']

}
```

Summarize the effect sizes.

```
summary(betas)
hist(betas)
```

Are there any significantly associated SNPs? If so, which SNPs are they?

```
assoc = which(pvalues<0.05 / length(pvalues))
assoc
```

How big are their effect sizes? How significant are they?

```
betas[assoc] %>% min; betas[assoc] %>% max
zscores[assoc]
pvalues[assoc] %>% min; pvalues[assoc] %>% max
```

- *effect sizes range from -3.4 to 2.2.*
- *-log10(p-values) 07 to 86*

Draw a QQ plot for log10(p) values.

```
obsLogPvs = sort(-log10(pvalues))
expLogPvs = sort(-log10(seq(1/M,1,1/M)))
plot(expLogPvs,obsLogPvs,main='QQ plot')
abline( a=0, b=1 )
#label the significant SNPs red
points(expLogPvs[(M-length(assoc)):M],obsLogPvs[(M-length(assoc)):M],col="red")
```

Is there inflation? Use the chi-square statistics to check.

```
chis = zscores^2
lambdaGC = median(chis)/0.454 # why .454?
lambdaGC
```

This `lambdaGC` suggests that there is some p-value inflation. I think the number of variants highlighted in red is biologically reasonable and we are not seeing huge number of significant data points.

Plot the phenotype predictions for the most significant SNP.

```
topSNP = genos[,order(pvalues)[1]]
plot(topSNP,phenos)
abline(lm(phenos~topSNP)$coeff,col="red")
```

Build a linear predictor of the phenotype using the associated SNPs.

```
ypred = array(0,N)
for(i in 1:N) {
      ypred[i] = genos[i,assoc] %*% betas[assoc]
}

plot(ypred,phenos)
```

What is the correlation between the predicted phenotype and the true phenotype?

```
cor(ypred,phenos)
```

BONUS: Test each of the associated SNPs for non-linearity.

```
hp = array(0,length(assoc))

for (i in 1:length(assoc)) {
  g = genos[,assoc[i]]
  h = g
  h[h==2]=0
  #Hint: can use anova(lm(?),lm(?)) or summary(lm(?))
  hp[i] <- anova( lm(phenos~g), lm(phenos~g*h) )$Pr[2]
  #skip multiple test correction for now
}

hp
```

BONUS: Visualize a linear SNP and a non-linear SNP.

```
par( mfrow=c(1,2) )


hp = array(0,M)

for (i in 1:M) {
  g = genos[,i]
  h = g
  h[h==2]=0
  #Hint: can use anova(lm(?),lm(?)) or summary(lm(?))
  hp[i] <- anova( lm(phenos~g), lm(phenos~g*h) )$Pr[2]
  #skip multiple test correction for now
}


linSNP = genos[,which.max(hp)]
nonlinSNP = genos[,which.min(hp)]

plot(linSNP,phenos)
points( c(0,1,2), tapply(phenos,linSNP, mean ), col=2, pch=16, cex=3 )
lines( c(0,1,2), tapply(phenos,linSNP, mean ), col=2, lwd=2  )
```

```
plot(nonlinSNP,phenos)
points( c(0,1,2), tapply( phenos,nonlinSNP, mean ), col=2, pch=16, cex=3 )
lines( c(0,1,2), tapply( phenos,nonlinSNP, mean ), col=2, lwd=2  )
```

Repeat the GWAS to test for recessive rather than additive genetic effects.

```
genos2 = genos
genos2[genos<1]=1 # (AA)[1], (AG)[1], (GG)[2]
pvalues2 = array(0,M)
zscores2 = array(0,M)
betas2 = array(0,M)
for(i in 1:M) {
  g = genos2[,i]
  res = summary(lm(phenos~g))
  zscores2[i] = res$coefficients[2,'t value']
  pvalues2[i] = res$coefficients[2,'Pr(>|t|)']
  betas2[i]   = res$coefficients[2,'Estimate']
}
```

Are the same SNPs significant or not?

```
assoc2 = which(pvalues2<0.05 / length(pvalues2))
assoc2
```

How did the effect sizes change?

```
plot(betas,betas2)
```

```
summary(betas)
summary(betas2)
```

The scatter plot and summary stats show the difference!

PART2: Simulating genotypes with LD.

Establish some important simulation parameters.

```
N = 1000 #number of individuals
M = 30   #number of non-causal SNPs
gs = matrix(0,nrow=N,ncol=M)
```

Simulate a GWAS data set. First, simulate the causal variant.

```
set.seed = (42) #set random seed so we all get the same numbers
MAF = 0.5
gC = rbinom(N,1,MAF) #causal variant
```

Then, simulate the phenotypes given the causal variant.

```
beta = 0.3 #association of causal variant
pheno = gC*beta + rnorm(N)
```

Generate 10 SNPS in tight LD with the causal SNP.

```
rho = 0.9

for(i in 1:10) {
  # idx: the chance they are going to be re-defined
  idx = rbinom(N,1,rho)
```

```
  gs[,i]=gC*idx+rbinom(N,1,MAF)*(1-idx)


  # test they have the right LD empirically
  cat( 'Observed LD = ', cor( gs[,i], gC ), '\n' )
  # Bonus: prove they have the right LD theoretically
}
```

```
## Observed LD =  0.9098785
## Observed LD =  0.8979044
## Observed LD =  0.9181811
## Observed LD =  0.8858927
## Observed LD =  0.9138863
## Observed LD =  0.889867
## Observed LD =  0.8878456
## Observed LD =  0.8838631
## Observed LD =  0.9020109
## Observed LD =  0.8898495
```

Do the same for 10 moderate LD partners (rho=0.6).

```
rho = 0.6

for(i in 11:20) {
  idx = rbinom(N,1,rho)

  gs[,i]=gC*idx+rbinom(N,1,MAF)*(1-idx)

  # test they have the right LD empirically
  cat( 'Observed LD = ', cor( gs[,i], gC ), '\n' )
  # Bonus: prove they have the right LD theoretically
}
```

Do the same for 10 independent SNPs (rho=0).

```
rho = 0

for(i in 21:30) {
  idx = rbinom(N,1,rho)

  gs[,i]=gC*idx+rbinom(N,1,MAF)*(1-idx)

  # test they have the right LD empirically
  cat( 'Observed LD = ', cor( gs[,i], gC ), '\n' )
  # Bonus: prove they have the right LD theoretically
}
```

Run GWAS on the causal variant. Then run GWAS on the other variants. Keep track of the zscores only.

```
zsC = summary(lm(pheno~gC))$coef[2,3]
zs = sapply( 1:M, function(i) summary(lm(pheno~gs[,i]))$coef[2,3] )
```

Visualize the relationship between the mean z-scores at the tag SNPs and the z-score at the causal SNP.

```
par( mfrow=c(2,2) )
breaks = hist(c(0,zsC,zs),plot=F)$breaks
```

```
hist(zs[1:10],breaks=breaks, col=1, main='LD partners')
abline(v=zsC)
hist(zs[11:20],breaks=breaks, col=2, main='Low-LD partner SNPs')
abline(v=zsC)
hist(zs[21:30],breaks=breaks, col=3, main='Independent SNPs')
abline(v=zsC)
```

BONUS: Perform LD score regression. First, calculate the LD scores. There should be M+1 of them.

```
ldscores = ?
ldscores
```

BONUS: Visualize LD score regression.

```
chis = c( ?, ? )^2
plot( ?, chis, ylab=expression(chi^2) )
#test for inflation
lambdaGC = median(chis)/0.454
lambdaGC
```

BONUS: Estimate heritability.

```
summary( lm( ? )$coef[2,1] * M/N
```

BONUS: What is the true heritability?

```
var(?) / var(?)
```