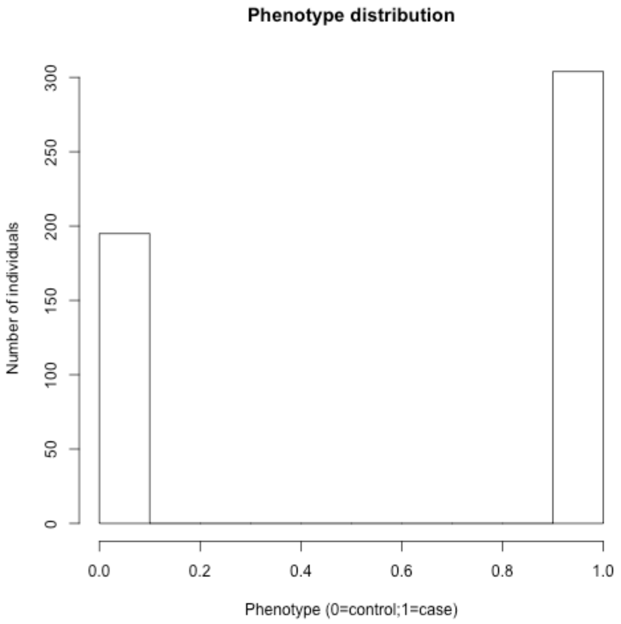
Final Exam

Dan Jin. Dec. 12, 2012

Problem 1:

1. Histogram of the phenotypes:



2. The sample **size (n) = 499**; the number of genotypes per individual **(N) = 2480**.
3. There is **no missing genotype data**. I checked whether there are none "A", "T", "G", "C" characters. I also checked whether there are missing phenotype by comparing the number of phenotype and the sample size. R scripts:

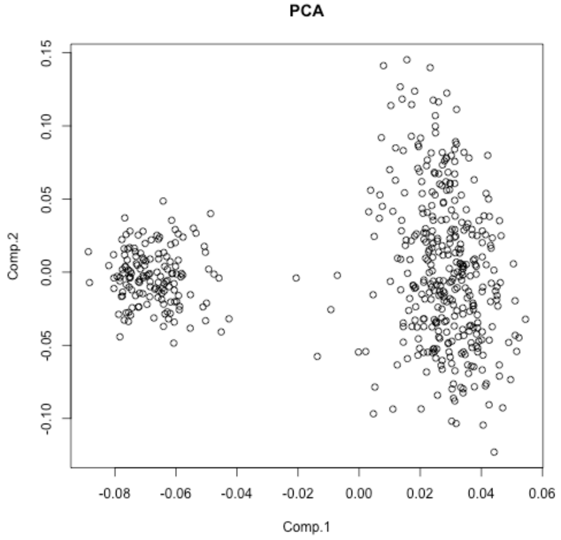
```
if ((sum(genodata.df != "A" & genodata.df != "T" & genodata.df !=
"G" & genodata.df != "C") == 0) & length(sample.ls$y) ==
sample.size) {
             cat("No missing genotype data.")
       }
```

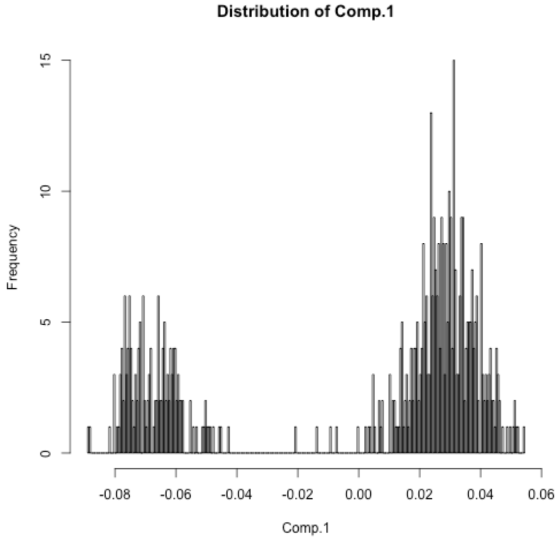
Problem 2:

1. Remove genotypes with a MAFH<0.05: Please refer to "FinalScript.r" for R code.
2. **2436** genotypes are left after removing genotypes with a MAF <0.05.

Problem 3:

1. Code Xa and Xd: Please refer to "FinalScript.r" for R code.
2. PCA plot: Please refer to "FinalScript.r" for R code.





3. The samples are clustered into two distinct groups on PCA plot, which indicates that they are from two different populations.
4. There are **147** individuals in one group and **352** individuals in the other one.
5. -0.03 is a reasonable split point because if we project all the points on PCA plot onto x-axis, there are two peaks separated by a gap around -0.03.

Problem 4:

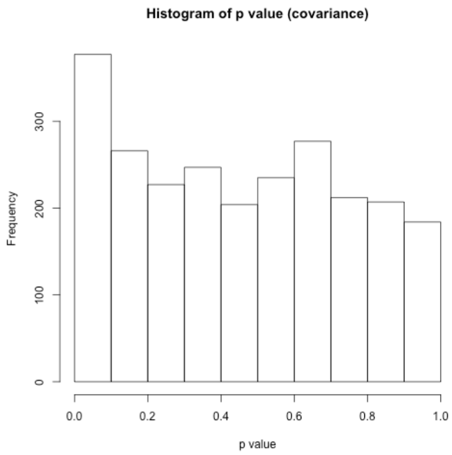
1. IRLS algorithm for four β parameters: Please refer to "FinalScript.r" for R code.
4. **mean(|$\hat{\boldsymbol{\beta}}_z$|) = 1.189981**; **mean(|$\hat{\boldsymbol{\beta}}_a$|) = 0.7216282**; **mean(|$\hat{\boldsymbol{\beta}}_d$|) = 0.3656255.**
5. The mean of the absolute values of $\hat{\beta}_z$ is greater than the mean values of the absolute values of the $\hat{\beta}_a$ and $\hat{\beta}_d$. This implies that the population structure affects the overall phenotype more significantly than β_a and β_d.

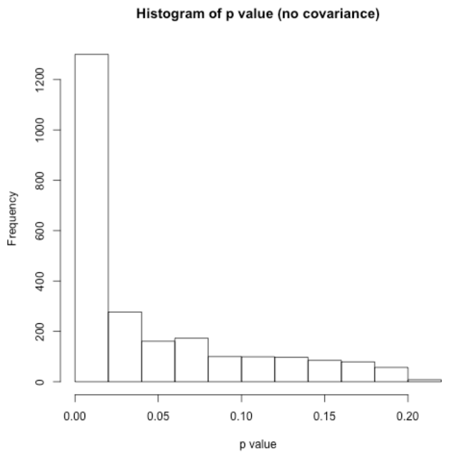
Problem 5:

1. IRLS algorithm for three β parameters: Please refer to "FinalScript.r" for R code.
2. **mean(|$\hat{\boldsymbol{\beta}}_{a-}$|) = 0.816693**. It is greater than **mean(|$\hat{\boldsymbol{\beta}}_a$|)**
3. **mean(|$\hat{\boldsymbol{\beta}}_{d-}$|) = 0.3796283**. It is greater than **mean(|$\hat{\boldsymbol{\beta}}_d$|)**
4. Since we didn't consider the population structure in the model, the affect of population structure is "incorporated" into β_a and β_d.

Problem 6:

1. p-values: Please refer to "FinalScript.r" for R code.
2. Histograms for the two lists of p-value.

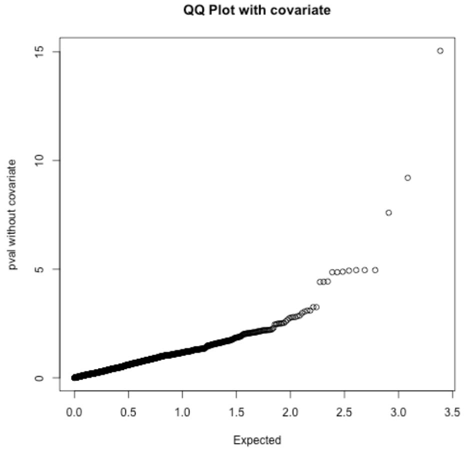


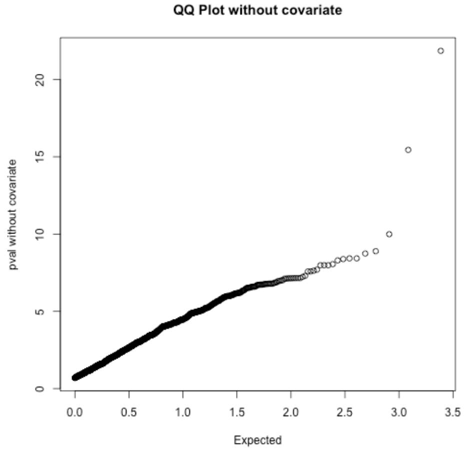


3. The p-values calculated with covariance are more closed to uniform distribution. Based on PCA plot, we know the samples are from two distinct populations. So the model with covariance is more closed to the true distribution. The p values calculated with this model are more uniform distributed under the null hypotheses.

Problem 7:

1. QQ plots for the two lists of p-values:





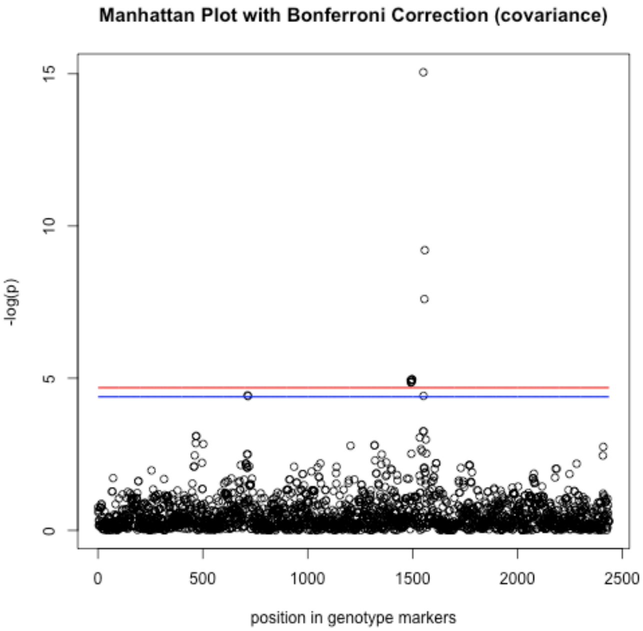
2. The QQ plot for the analysis considering the covariate looks appropriate, because it looks like a line with a tail. Most genotypes are not in LD with a causal polymorphism, so the null hypothesis is correct, and their p-values follow a uniform distribution. But the few that are in LD with a causal polymorphism produce significant p-value and these are in the “tail”.

3. The QQ plot for the analysis without the covariate indicates something is wrong, because the plot looks like a convex curve with a tail. This means many genotypes that are not in LD with the causal polymorphism have more significant p values than expected. In this case, we should not trust the genotypes with significant p-values as potential hits. Unaccounted for

covariates can cause this issue and the most frequent issue is unaccounted for population structure.

Problem 8:

1. Manhattan Plot for the p-values obtained with covariate.



2. The **red** line indicates a Bonferroni corrected threshold for a single test α = 0.05. There are **10** association hits indicated at this cutoff (10 genotypes that have p values more significant than the threshold).

3. The **blue** line indicates a Bonferroni corrected threshold for a single test α = 0.1. There are **13** association hits indicated at this cutoff (13 genotypes that have p values more significant than the threshold.

4. Type I error α represents the probability of incorrectly reject null hypothesis when it's true (false positive). The choice of α value is quite arbitrary. When we set a threshold as α=0.05, it means there will be 0.05 probability that we reject the null hypothesis incorrectly (5% false positives). α=0.1 threshold means there will be 0.1 probability that we reject the null hypothesis incorrectly (10% false positives). Unfortunately, we do not know which hits are true positives, which ones are false positives. So the hits we found at either cutoff may both be true positive, or both may be false positives (although 5% is pretty low, there's still a chance we get false positives.) And for the three hits only indicated at α=0.1 cutoff, they could be true positives, with relatively less significant p values; or they could be false positives due to lower threshold. We should be more confident in the hits at the more

conservative cutoff (α=0.05 in this case), since lower type I error will give fewer false positives.

Problem 9:

1. List the number of the two SNP markers corresponding to the two most significant hits indicating different location in the genome (Please refer to "FinalScript.r" for R code):

```
$position
[1] 1496 1551

$pval
[1] 1.101063e-05 8.906336e-16
```

2. Nine dummy variables for the first five individuals:

Individual	#1	#2	#3	#4	#5
X_{mu}	1	1	1	1	1
$X_{a,1}$	1	0	1	0	0
$X_{d,1}$	-1	1	-1	1	1
$X_{a,2}$	0	0	0	0	-1
$X_{d,2}$	1	1	1	1	-1
$X_{a,1}$ $X_{a,2}$	0	0	0	0	0
$X_{a,1}$ $X_{d,2}$	1	0	1	0	0
$X_{d,1}$ $X_{a,2}$	0	0	0	0	-1
$X_{d,1}$ $X_{d,2}$	-1	1	-1	1	-1

R code for determining the dummy variables:

```
position.1=sig.hit.ls$position[1]
position.2=sig.hit.ls$position[2]

epi.X=list()
for (i in 1:5){
      Xa.1=sample.ls$Xa[i,position.1]
      Xd.1=sample.ls$Xd[i,position.1]
      Xa.2=sample.ls$Xa[i,position.2]
      Xd.2=sample.ls$Xd[i,position.2]
      Xa.1_Xa.2=Xa.1*Xa.2
      Xa.1_Xd.2=Xa.1*Xd.2
      Xd.1_Xa.2=Xd.1*Xa.2
      Xd.1_Xd.2=Xd.1*Xd.2
      epi.X[[i]]=c(1,Xa.1,Xd.1,Xa.2,Xd.2,Xa.1_Xa.2,Xa.1_Xd.2,Xd.1
_Xa.2,Xd.1_Xd.2)
}

> epi.X
[[1]]
[1]  1  1 -1  0  1  0  1  0 -1
```

```
[[2]]
[1] 1 0 1 0 1 0 0 0 1

[[3]]
[1]  1  1 -1  0  1  0  1  0 -1

[[4]]
[1] 1 0 1 0 1 0 0 0 1

[[5]]
[1]  1  0  1 -1 -1  0  0 -1 -1
```

3. **Definition of epistasis**: a case where the effect of an allele substitution at one locus A1 -> A2 alters the effect of a substituting an allele at another locus B1->B2
4. The definition of epistasis doesn't indicate anything about the physical interaction between these two loci or the physical interaction between their transcription (mRNAs) or translation products (proteins). So their protein hypothesis can't be explained by the result of epistasis analysis.

Problem 10: Perform a GWAS using a pseudo-Bayesian approach

1. For the **linear model**, we assume the probability model is normal:
   $\boldsymbol{y} = \boldsymbol{x}\beta + \epsilon, \qquad \epsilon \sim multiN(0, \boldsymbol{I}\sigma_\epsilon^2)$
   a. The complete posterior probability for the genetic model is:
      $\Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) \propto \Pr\,(y|\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2) \Pr\,(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2)$
   b. We are assuming uniform priors with the following joint probability distribution:
      $\Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2) = \Pr(\beta_\mu) \Pr(\beta_a) \Pr(\beta_d) \Pr(\sigma_\epsilon^2)$
      $\Pr(\beta_\mu) = \Pr(\beta_a) = \Pr(\beta_d) = c$
      $\Pr(\sigma_\epsilon^2) = c$
   c. Under this prior, the complete posterior distribution is multivariate normal:
      $$\Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) \propto (\sigma_\epsilon^2)^{-\frac{n}{2}} \exp\left[\frac{(y - x\beta)^T (y - x\beta)}{2\sigma_\epsilon^2}\right]$$
      $$\sim multivariate\ normal$$
   d. We are interested in the marginal posterior probability of the parameters. Under this prior, the marginal posterior probability distribution is a multi-t distribution:
      $$\Pr(\beta_a, \beta_d | y) = \int_{-\infty}^{\infty} \int_0^{\infty} \Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y)\, d\beta_\mu d\sigma_\epsilon^2 \sim multi - t\ distribution$$
   e. We are testing the following hypotheses:
      $H_0: \beta_a \cap \beta_d = 0, \qquad H_A: \beta_a \neq 0 \cup \beta_d \neq 0$

f. We are using pseudo-Bayesian approach to hypothesis testing that makes use of credible intervals:

$$c.i.(\theta) = \int_{-c_\alpha}^{c_\alpha} \Pr(\theta|y) d\theta = 1 - \alpha$$

2. Critical components/steps

a. Code Xa, Xd

b. Calculate the Bayesian estimates of the parameters:

i. Calculate the matrix C:

$$C = \begin{bmatrix} X_a^T X_a & X_a^T X_d \\ X_d^T X_a & X_d^T X_d \end{bmatrix}$$

ii. Construct estimators using the posterior probability distribution ($\hat{\theta} = mean(\theta|y)$). Calculate the posterior means of β_a and β_d:

$$mean(\Pr(\beta_a, \beta_d|y)) = \left[\hat{\beta}_a, \hat{\beta}_d\right]^T = C^{-1}[X_a, X_d]^T y$$

iii. Calculate the marginal posterior covariance matrix with the posterior covariance degrees of freedom = n-6:

$$cov = \frac{\left(y - [X_a, X_d]\left[\hat{\beta}_a, \hat{\beta}_d\right]^T\right)^T \left(y - [X_a, X_d]\left[\hat{\beta}_a, \hat{\beta}_d\right]^T\right)}{n-6} C^{-1}$$

c. Calculate the 0.95 credible interval using qmvt() function in R with the multi-t degrees of freedom = n-4:

d. Decide if the 0.95 credible interval over-laps zero

i. If the 0.95 credible interval overlaps zero, do not reject the null hypothesis.

ii. If the 0.95 credible interval doesn't overlap zero, reject the null hypothesis.

Instruction of using FinalScript.r

To run this script, please copy FinalScript.r to the same fold where containing two input data files, then type the following lines in R.

```
source("FinalScript.r")
result.ls=list()
result.ls=control.fun()
```

Result of this script:

1. Return one list, **result.ls**, which includes all results required.
2. Create **7 plots** in png format. (See below for more details)

Summary of the content in result.ls:

1. $covar: beta.MLE, means of the absolute value of beta and p values calculated with covariance Xz
2. $nocovar: beta.MLE, means of the absolute value of beta and p values calculated without covariance Xz
3. $hits.ls: all genotype marker (marker number and their p value) significant than Bonferroni correction;
4. $hit.set.ls: distinct sets;
5. $sig.hit.ls: the most significant marker in each set (marker number and the associated p value);
6. $epi.X: a list containing the nine dummy variables for the first five individuals for epistasis analysis.

Structure of result.ls:

```
 result.ls
 |-$covar
 |  |-beta.MLE
 |  |-beta.a.mean
 |  |-beta.d.mean
 |  |-beta.z.mean
 |  |-pval
 |
 |-$nocovar
 |  |-beta.minus.MLE
 |  |-beta.a.minus.mean
 |  |-beta.d.minus.mean
 |  |-pval
 |
 |-$hits.ls
 |  |-$position
 |  |-$pval
 |
 |-$hit.set.ls
 |
 |-$sig.hit.ls
 | |-$position
 | |-$pval
 |
 |-$epi.X
```

List of .png image files.

1. Q1_Phenotype distribution.png
2. Q3_PCA.png
3. Q6.1_Histogram of p value (covariance).png
4. Q6.2_Histogram of p value (no covariance).png
5. Q7.1_QQ Plot with covariate.png
6. Q7.2_QQ Plot without covariate.png
7. Q8_Manhattan Plot with Bonferroni correction (covariance).png

Screen output when running the script:

```
Importing data.
The sample size (n) = 499
The number of genotypes (N) = 2480

Q1: Check phenotype distribution.
"Q1_Phenotype distribution.png" created.

Q1: Check genotype data.
No missing genotype data.

Q2: Remove genotype with MAF< 5 %.
2436 SNPs ( 98.22581 %) have MAF >= 5 %.

Coding Xa and Xd...

Q3: PCA.
"Q3_PCA.png" created.
There are 147 individuals in one group and 352 individuals in the other one.

::: GWAS with covariance :::
Q5: Estimating beta parameters and calculating p values...
beta parameters and p values are stored in sample.ls$covar
The mean of the absolute values of beta.a.hat is 0.7216282
The mean of the absolute values of beta.d.hat is 0.3656255
The mean of the absolute values of beta.z.hat is 1.189981

Q6: Histogram of p values.
"Q6.1_Histogram of p value (covariance).png" created.

Q7: QQ plot.
"Q7.1_QQ Plot with covariate.png" created.

Q8: Manhattan plot.
"Q8_Manhattan Plot with Bonferroni correction (covariance).png" created.
10 hits are indicatd at a (single test) alpha=0.05.
13 hits are indicatd at a (single test) alpha=0.1.

::: GWAS without covariance :::
Q5: Estimating beta parameters and calculating p values...
beta parameters and p values are stored in sample.ls$covar
```

```
The mean of the absolute values of beta.a.minus.hat is 0.816693
The mean of the absolute values of beta.d.minus.hat is 0.3796283

Q6: Histogram of p values.
"Q6.2_Histogram of p value (no covariance).png" created.

Q7: QQ plot.
"Q7.2_QQ Plot without covariate.png" created.

Q9: Find the two most significant hits indicating different locations in the
genome.
Finding potential hits...
Clustering hits into sets...
Finding the most significant hits in each sets...
The positions and p values of the two most significant hits from different
locations are stored in result.ls$sig.hit.ls
Position: 1496 1551
Corresponding p values: 1.101063e-05 8.906336e-16

Q9: The nine dummy variables for the first five individuals for epistatic
analysis.
The dummy variables are stored in result.ls$epi.X
Individual 1 : 1 1 -1 0 1 0 1 0 -1
Individual 2 : 1 0 1 0 1 0 0 0 1
Individual 3 : 1 1 -1 0 1 0 1 0 -1
Individual 4 : 1 0 1 0 1 0 0 0 1
Individual 5 : 1 0 1 -1 -1 0 0 -1 -1

FinalScript.r is done.
```

A quick view of the data in result.ls:

```
$covar
$covar$beta.MLE
                 [,1]          [,2]          [,3]      [,4]
   [1,] -0.9786935547  7.964151e-01  1.892851e-01 1.2712466
...

$covar$beta.a.mean
[1] 0.7216282

$covar$beta.d.mean
[1] 0.3656255

$covar$beta.z.mean
[1] 1.189981

$covar$pval
   [1] 1.758578e-01 2.603567e-01 2.278261e-01 6.765007e-01 4.616836e-01
4.941447e-01 7.218077e-01
...

$s0
[1] "############################################"

$nocovar
$nocovar$beta.minus.MLE
                 [,1]          [,2]          [,3]
   [1,]  0.1794364745  0.5701820406  3.024016e-01
...
```

```
$nocovar$beta.a.minus.mean
[1] 0.816693

$nocovar$beta.d.minus.mean
[1] 0.3796283

$nocovar$pval
   [1] 1.255652e-01 1.771274e-01 1.208737e-01 1.540837e-01 1.178842e-01
4.590334e-02 2.221010e-02
...

$s1
[1] "############################################"

$hits.ls
$hits.ls$positon
 [1] 1491 1492 1493 1495 1496 1497 1498 1551 1556 1558

$hits.ls$pval
 [1] 11.35664 11.19048 11.19048 11.24981 11.41665 11.41665 11.41665 34.65460
17.49726 21.19587

$s2
[1] "############################################"

$hit.set.ls
$hit.set.ls[[1]]
[1] 1491 1492 1493 1495 1496 1497 1498

$hit.set.ls[[2]]
[1] 1551 1556 1558

$s3
[1] "############################################"

$sig.hit.ls
$sig.hit.ls$position
[1] 1496 1551

$sig.hit.ls$pval
[1] 1.101063e-05 8.906336e-16

$s4
[1] "############################################"

$epi.X
$epi.X[[1]]
[1]  1  1 -1  0  1  0  1  0 -1

$epi.X[[2]]
[1] 1 0 1 0 1 0 0 0 1

$epi.X[[3]]
[1]  1  1 -1  0  1  0  1  0 -1

$epi.X[[4]]
[1] 1 0 1 0 1 0 0 0 1

$epi.X[[5]]
[1]  1  0  1 -1 -1  0  0 -1 -1
```