Final Exam (Please See RMD File for Code Blocks)

November 14, 2020

---

1 Problem

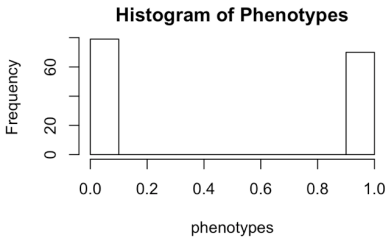
1.1 Part A: Importing the Phenotype Data (Chunk 1)

```
phenotypes <- read.csv("final2019_pheno.csv")
```

1.2 Part B: Calculating the sample size n (Chunk 1)

```
nrow(phenotypes)
> 149
```

1.3 Part C: Phenotypes Histogram (Chunk 1)



$(85).png$

2 Problem

2.1 Part A: Importing the Genotype Data (Chunk 1)

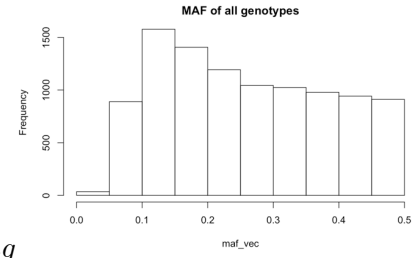
```
genotypes <- read.csv("final2019_genotypes.csv")
```

2.2 Part B: Reporting the number of SNPs N (Chunk 1)

```
N_genotypes <- ncol(genotypes);
> 10000
```

2.3 Part C: Calculating the MAF, MAF Plot (Chunk 2)

```
maf_vec = apply(genotypes,2, maf_calc)
```



(86).*png*

3 Problem

3.1 Part A: Applying a Logistic Regression with No Covariates (Chunk 4)

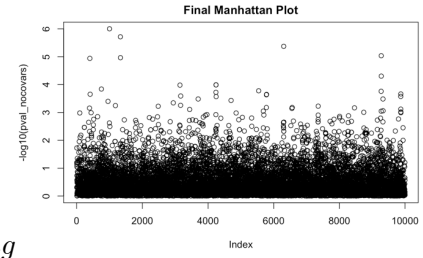
With the code from the IRLS Algorithm, we can calculate the LRT for each p-value, by running the code below.

```
    library(MASS)
library(ggplot2)

Y_binary <- matrix(phenotypes, ncol = 1)
logl_nocovars<-vector(length = N_genotypes)

for(j in 1:N_genotypes){
  logistic_nocovars <-logistic.IRLS(X = cbind(xa_matrix[,j], xd_matrix[,j]), Covars = NULL,Y=Y_bin
  logl_nocovars[j] <- logistic_nocovars[[2]]
  #cat("Processing genotype =", j,"\n")
}

# log likelihood for NULL hypothesis
logl_H0_nocovars <- logistic.IRLS(Y=Y_binary, X= NULL, Covars=NULL, beta.initial.vec = c(0))[[2]]

LRT_nocovars <-2*logl_nocovars-2*logl_H0_nocovars #likelihood ratio test statistic
#likelihood ratio test statistic for every genotype

pval_nocovars <- pchisq(LRT_nocovars, 2, lower.tail = FALSE)
```

3.1.1 p values vector output

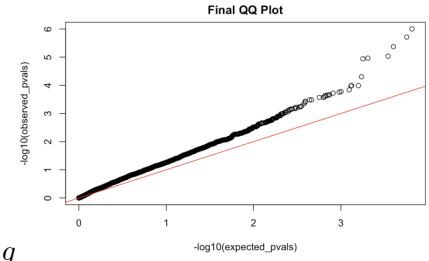
```
    > pval_nocovars
   [1] 4.910436e-01 6.291158e-02 6.291158e-02 1.875348e-02 7.361430e-02 9.598369e-01 8.914101e-02
   [9] 1.840138e-01 2.840807e-01 8.093166e-01 7.801604e-01 3.322038e-01 9.928469e-01 7.503981e-01
  [17] 4.622421e-01 3.229075e-01 6.797912e-01 1.937457e-01 4.414211e-02 1.346931e-01 4.674891e-01
  [25] 2.420868e-01 2.420868e-01 2.461041e-01 4.693336e-01 3.077064e-01 7.697710e-01 3.257051e-01
  .......
```

3.2 Part B: Manhattan Plot (Chunk 6)



(87).*png*

3.3 Part C: QQ Plot (Chunk 6)



(88).*png*

4 Problem

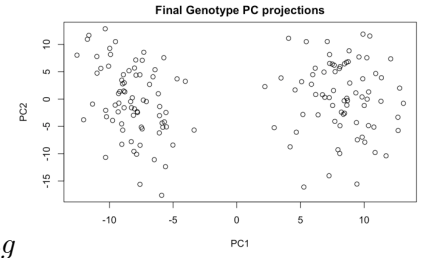
4.1 Part A: Performing a PCA (Chunk 7)

```
geno_pca <- prcomp(genotypes)
```

4.2 Part B: Projection Plot for PC 1 and PC 2 (Chunk 7)

With the prcomp method in R, we are able to 'perform' a PCA by looking at the individual principal components, that we can view from the data files.

```
plot(geno_pca$x[,1], geno_pca$x[,2], main = "Final Genotype PC projections", xlab = "PC1", yla
```



(89).*png*

5 Problem

5.1 Part A: Calculating the p-values for a Logistic Regression with Covariates (Chunk)

5.1.1 LRT Implementation (Chunk 8)

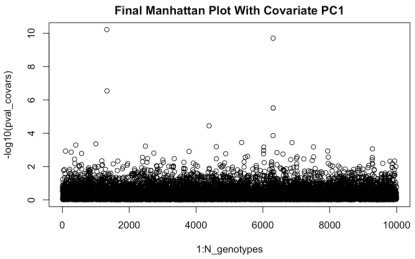
```
lr_likelihood <- function(y, x_input = NULL){
```

```
    n_samples <- length(y)

    X_mx <- cbind(matrix(1, nrow = n_samples, ncol = 1), x_input)

    MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% y

    y_hat <- X_mx %*% MLE_beta

    var_hat <- sum((y - (y_hat))^2) / (n_samples - 1)

    log_likelihood <- -((n_samples / 2) * log(2 * pi * var_hat) ) - ((1/ (2*var_hat)) * sum((y - (

    return(log_likelihood)
}

LRT_test <- function(logl_H0, logl_HA, df_test){

    LRT<-2*logl_HA-2*logl_H0 #likelihood ratio test statistic
    #likelihood ratio test statistic for every genotype
    pval <- pchisq(LRT, df_test, lower.tail = F)
    return(pval)
}


h0_withPC <- lr_likelihood(phenotypes, cbind(geno_pca$x[,1]))

pval_covars <-vector(length =N_genotypes)

for (i in 1 : N_genotypes){

 #   cat("Testing Gentoype = ", i, "\n")
    ha_withPC <- lr_likelihood(phenotypes, cbind(xa_matrix[,i], xd_matrix[,i], geno_pca$x[,1]))
    pval_covars[i] <- LRT_test(h0_withPC, ha_withPC, df_test = 2)
}
```

5.1.2 Output of p values vector

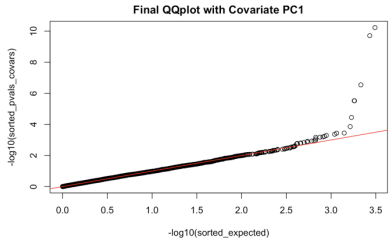
```
    > pval_covars
   [1] 0.1520381216 0.2872408664 0.2872408664 0.2110624041 0.9099776833 0.2338996806 0.1366777315
   [9] 0.0588023845 0.2646853292 0.4374557971 0.6427561254 0.6228036029 0.9655646497 0.6947020409
  [17] 0.3698471199 0.2877714941 0.3822955443 0.5124401138 0.0918964693 0.4683129465 0.7222273620
```

5.2 Part B: Manhattan Plot with Covariates (Chunk 9)



(90).*png*

5.3 Part C: QQ Plot with Covariates (Chunk 9)



(91).*png*

6 Problem

6.1 Part A: Analyzing the p-values (Chunk 10)

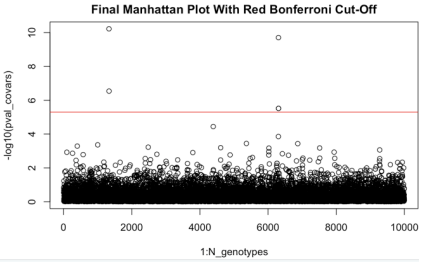
When enforcing the Bonferroni correction

$$\text{BC} = \frac{0.05}{\text{Numberofgenotypes}} \ ,$$

in the usual way in R, we see that the relevant p-values, after introducing the first principal component as a covariate, are $1331, 1333, 6305, 6306, 6308$.

6.2 Part B: Reporting Separate Peaks with the Bonferroni Correction (Chunk 10)

From the red line, we report 2 peaks that are greater than the Bonferroni corrected cutoff. In terms of criteria, we have determined these peaks by looking at which p-values, with the first principal component as a covariate, that are higher than the cutoff that we provided a numerical formula for in **A**, in which we looked at all p-values above $-\log_{10}(BC)$.



(92).*png*

7 Problem

7.1 Part B: Positively Correlated Samples

The samples $(1, 2), (3, 4)$ are positively correlated.

7.2 Part C: Negatively Correlated Samples

There are no negatively correlated samples.

7.3 Part D: Uncorrelated Samples

The samples $(1, 3), (1, 4), (2, 3), (2, 4)$ are uncorrelated.

8 Problem

8.1 Part A: Relationship Between Equations 6 and 7

To determine the relationship, we recall that $h^2 = \frac{V_A}{V_P}$, with $V_A = 0$, implies that solving the equation

$$2\text{MAF}(1 - \text{MAF})\beta_\alpha^2 = 0 \ ,$$

would give us the relationship between different β's. From our knowledge that these alleles are segregating with non-zero generic effects, not only so that the relationship exists but also that $V_p = 0$ justifies why solving this equation gives the answer. Moreover, recalling that

$$2p_1(1 - p_1)\beta_\alpha^2 = 0 \ ,$$

or, as directly stated on the last lecture slide Additive Genetic Variance II,

$$V_A = 2p(1 - p)\beta_\alpha^2 \ ,$$

implies that, after substitution,

$$2p_1(1 - p_1)(\beta_a(1 + \frac{\beta_d}{2}(p_1 - p_2))^2 \ .$$

It turns out that the relationship is of the form,

$$\begin{aligned}
&\Rightarrow \beta_a(1 + \frac{\beta_d}{2}(p_1 - p_2))^2 = 0 \\
&\Leftrightarrow \beta_a(1 + \frac{\beta_d}{2}(p_1 - p_2))^2 = 0 \\
&\Leftrightarrow 1 + \frac{\beta_d}{2}(p_1 - p_2) = 0 \\
&\Leftrightarrow -2 = \beta_d(p_1 - p_2) \\
&\Leftrightarrow -2 = \beta_d(p_1 - 1 - p_1) \\
&\Leftrightarrow \frac{-2}{\beta_d} + 1 = 2p_1 \\
&\Leftrightarrow \frac{-1}{\beta_d} + \frac{1}{2} = p_1 \ ,
\end{aligned}$$

which is precisely the relationship that we want.

To be as clear as possible, from what was mentioned in the hint given in the last lecture, we know that this is the relationship that we want; the equality above explicitly show that the relationship that we want is, precisely,

$$\beta_\alpha = \beta_a(1 + \frac{\beta_d}{2}(p_1 - p_2)) \ ,$$

again, from solving the system of equations,

$$\begin{aligned}\beta_\mu - \beta_a - \beta_d &= 0 \ , \\ a + d &= \beta_\mu + \beta_d \ , \\ 2a &= \beta_\mu + \beta_a - \beta_d \ .\end{aligned}$$

8.2 Part B

When $MAF = 0$,

$$\begin{aligned}h^2 = \frac{V_A}{V_P} = \frac{2\text{MAF}(1-\text{MAF})\beta_\alpha^2}{V_P} = \frac{0}{V_P} = 0 \ , \\ \Rightarrow h = 0 \ .\end{aligned}$$

9 Problem

9.1 Part A

The 9 possible genotype combinations are $A_1A_1B_1B_1$, $A_1A_2B_1B_1$, $A_2A_2B_1B_1$, $A_1A_1B_1B_2$, $A_1A_2B_1B_2$, $A_2A_2B_1B_2$, $A_1A_1B_2B_2$, $A_1A_2B_2B_2$, $A_2A_2B_2B_2$.

9.2 Part B

The values are

$$X_{a,1}(A_1A_1B_1B_1) = -1 \ ,$$

$$X_{d,1}(A_1A_1B_1B_1) = -1 \ ,$$

$$X_{a,2}(A_1A_1B_1B_1) = -1 \ ,$$

$$X_{d,2}(A_1A_1B_1B_1) = -1 \ .$$

9.3 Part C

With the assumption that a linear regression model is the 'correct' model, we could calculate the expected phenotypic value of an individual as follows,

$$\begin{aligned}\mathbf{E}(Y|g = A_1A_1B_1B_1) = \beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} + X_{a,1}X_{a,1}\beta_{a_1,a_2} \\ + X_{a,1}X_{d,2}\beta)a_1, d_2 + X_{d_1}X_{a_2}\beta_{d_1,a_2} + X_{d_1}X_{d_1}\beta_{d_1,d_2} \\ = 0.2 + (-1)0.1 + (-1)0.2 + (-1)(-0.3) + (-1)0.17 + \\ (1)(-0.11) + (1)0.32 + (1)0.08 + (1)(-0.03) \\ \approx 0.29 \ .\end{aligned}$$

10 Problem

10.1 Part A

$\Omega = \{H, T\}$

10.2 Part B

$\mathcal{F} = \emptyset, \{H\}, \{T\}, \{H, T\}$

10.3 Part C

The probability function for this sigma algebra is $\Pr(\emptyset) = 0, \Pr(\{H\}) = 0.5, \Pr(\{T\}) = 0.5, \Pr(\{H, T\}) = 1.0$.

10.4 Part D

From the Bernoulli distribution, we would define such a random variable by either setting $X(\{H\}) = 0, X(\{T\}) = 1$, or $X(\{H\}) = 1, X(\{T\}) = 0$.

10.5 Part E

The expected value is

$$\mathbf{E}X = \sum_{i=0}^{1} (X = i)\Pr(X_i) = 0 \times 0.5 + 1 \times 0.5 = 0.5 \ .$$

10.6 Part F

The power set 2^{10} captures all possible outcomes of that sample.

10.7 Part G

One statistic that would not be great to use is

$$T(x_1, \cdots, x_{10}) = -1 \ ,$$

because for this particular statistic T, the samples take constant values that are outside of the possible range from the parameter.

10.8 Part H

The MLE is

$$\frac{1}{10} \sum_{i=1}^{10} x_i \ .$$

10.9 Part I

The p-value, in the case that the observed value of the statistic $T(\mathbf{x}) = 10 \times \text{MLE}(\hat{p}) = 8$, is influenced by

$$T(\mathbf{x}) = T(x_1, \cdots, x_{10}) = 10 \times \text{MLE}(\hat{p}) = \sum_{i=1}^{10} x_i = t = 8 \ ,$$

which from the rigorous definition of a p-value, equals

$$\int_{t=8}^{\infty} \Pr(|T(\mathbf{x})||\mathcal{H}_0 = \text{true})\text{d}t = \int_{t=8}^{\infty} \Pr(|8||\mathcal{H}_0 = \text{true})\text{d}t = \int_{t=8}^{\infty} \Pr\bigg(\bigg(\text{MLE}(\hat{p}) = \frac{4}{5}\bigg)|\mathcal{H}_0 = \text{true}\bigg)\text{d}t \ ,$$

which allows us to conclude that,

$$\int_{t=8}^{\infty} \Pr(|8||\mathcal{H}_0 = \text{true}) = 1 \ .$$

10.10 Part J

For $\alpha = 0.05$, we would not reject the null $\mathcal{H}_0$ because the p-value that we calculated in **I** is greater than 0.05.