Quantitative Genomics and Genetics - Spring 2021
BTRY 4830/6830; PBSB 5201.01

Homework 5 (version 1)

Assigned April 7; Due 11:59PM April 16

Problem 1 (Easy)

a. Provide a symbolic formula that defines a causal polymorphism AND provide a definition of a causal polymorphism in words.

b. For the genetic regression model:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \tag{1}$$

$$\epsilon \sim N(0, \sigma^2_\epsilon) \tag{2}$$

what is the $\mathrm{E}(Y|X_a = x_a, X_d = x_d)$ (i.e. the expected value of the phenotype for an individual with a specific genotype: x_a, x_d? HINTS: 1. Remember the rule of the Expected Value for a sum of random variables, 2. Make use of the expected value of a normal distribution.

Problem 2 (Medium)

Note that for this question, your answer should include R code that generates the appropriate answers. Use Rmarkdown and submit your .Rmd script (HTML optional) and note there will be penalties for scripts that fail to compile. Please note that you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!).

Consider the data in the files (‘QG21 - hw5_phenotypes.txt’; ‘QG21 - hw5_genotypes.txt’) of the scaled height phenotypes and SNP genotype data (respectively) collected in a GWAS. Note that in the ‘phenotypes’ file the column lists the individuals in order (1st entry is the phenotype for individual 1, the nth is for individual n). Also note that for each of the SNPs, there are two total alleles, i.e. two letters indicate each SNP and there are three possible states per SNP genotype: two homozygotes and a heterozygote. In the ‘genotypes’ file, each column represents a specific SNP (column 1 = genotype 1, column 2 = genotype 2) and each consecutive pair of rows represent all of the genotype states for an individual for the entire set of SNPs (rows 1 and 2 = all of individual 1’s genotypes, rows 3 and 4 = individual 2’s genotypes). Also note that the genotypes in the file are listed in order along the genome such that the first genotype is ‘genotype 1’ and the last is ‘genotype N’.

a. Write R code that inputs the phenotype, plus an (additional) line of code that calculates the number of samples n and report the number n. *Note that you do not have to output the phenotypes (!!) just provide the R code and report the value for* n.

b. Write R code that produces a histogram of your phenotype data. *Provide the R code and the histogram.*

c. Write R code that inputs the genotype data plus a line of code that outputs the number of genotypes N and report the value of N. *Note that you do not have to output the genotypes (!!) just provide the R code and report the value for N you obtained from these data.*

d. Write R code that converts your genotype data input in part [c] into two new matrices, the first a matrix where each genotype is converted to the appropriate X_a value and the second where each genotype is converted to the appropriate X_d value. *Note that you do not have to output the matrices (!!) just provide the R code that will create the matrices if we run it.*

e. Write R code to calculate $MLE(\hat{\beta}) = [\hat{\beta}_{\mu}, \hat{\beta}_a, \hat{\beta}_d]$ for each genotype in the dataset, an F-statistic for each genotype, and a p-value for each genotype using the R function pf(F-statistic, df1, df2, lower.tail = FALSE). *Note that you may NOT use an existing R function for ANY of these calculations other than the calculation of the p-value (=you must write code that calculates each component except the p-value). Also, same: you do not have to output anything (!!) just provide the R code.*

f. Write R code to produce a Manhattan plot (i.e., genotypes in order on the x-axis and -log(p-values) on the y-axis. *Note do NOT use an R function (=write your own code to produce the Manhattan plot) and DO provide your Manhattan plot.*

g. Write R code to produce a histogram of the p-values *Provide the R code and the histogram.*

h. What distribution would you expect the histogram in [g] to resemble IF the null hypotheses you tested (many times) in part [e] were correct? Your histogram in [g] should deviate from this expectation in a particular way - describe how your histogram deviates and provide an explanation for why it is deviating?

i. Using an overall study controlled Type I error of 0.05, write code that uses a Bonferroni correction to assess whether to reject the null hypothesis for each genotype, where your code also outputs the number of each genotype for which you rejected the null (remember: the genotypes are provided in order along the genome). Report the numbers of the genotypes for which you rejected the null.

j. Assuming the set of genotypes for which you rejected the null hypothesis in part [i] do indeed indicate the positions of causal genotypes in the genome, how many causal genotypes do you think these significant genotypes are indicating overall (and explain your answer)?

Problem 3 (Difficult)

In quantitative genomics, the null hypothesis of interest for a genotype X and phenotype Y can be stated in the general form:

$$H_0 : Cov(X, Y) = 0 \tag{3}$$

where if we consider the genetic linear regression model with random variables X_a and X_d, this null hypothesis can be stated as:

$$H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \tag{4}$$

or even more precisely as:

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \tag{5}$$

To see the connection between (5) and (6), demonstrate that $Cov(X_a, Y) = 0$ and $Cov(X_d, Y) = 0$ when $\beta_a = 0$ and $\beta_d = 0$. HINT: note that for arbitrary random variables X_1, X_2, and X_3 that $Cov(X_1, X_2 + X_3) = Cov(X_1, X_2) + Cov(X_1, X_3)$ and that for a linear regression $Pr(X_a, \epsilon) = Pr(X_a)Pr(\epsilon)$ and $Pr(X_d, \epsilon) = Pr(X_d)Pr(\epsilon)$. Show the steps of the derivation and explain the rules you use where appropriate.