Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Homework 6

Assigned April 8; Due 11:59PM April 15

Problem 1 (Easy)

Consider the (slightly idealized) genetics behind behind one of Mendel's famous experiments with pea plants (look Mendel up on wikipedia if you are new to genetics), where for two alleles A_1 and A_2 the phenotype of a pea is guaranteed to be 'yellow' if the genotype is either A_1A_1 or A_1A_2 and guaranteed to be 'green' if the genotype is A_2A_2. If you were to code a random variable for this system $Y(\text{yellow}) = 1$ and $Y(\text{green}) = 0$ and used a (genetic) linear regression to model this case, what would be the true values of the parameters β_a, β_d, and σ_ϵ^2?

Problem 2 (Medium)

For the 'case / contrrol' phenotype data in the file 'QG16_HW6_phenotypes.txt' and the genotype data in the file 'QG16_HW6_genotypes.txt' (where each column is a genotype coded as -1 or 1 for homozygotes and 0 for heterozygotes, each row contains the genotypes for an individual, and assume these genotypes are in the same order along a chromosome as they are presented in the columns) answer the following questions concerning a GWAS analysis of these data. For parts of this question that require R code answers, provide your code in a text file that is easy for us to run to produce the requested output (NOTE THAT FOR FULL CREDIT = easy to run code in a txt file, plus figures, and written answers as appropriate).

a. When using a logistic regression model applied independently to each of these N markers, state the null and alternative hypothesis that will be assumed in each case.

b. For each of the N markers, calculate the MLE of the β parameters under the null hypothesis using the IRLS algorithm. For your answer, provide R code that outputs the IRLS calculated estimates for each parameter of each marker AND for one marker of your choice, write down the estimates you obtained (make sure you note which marker you are reporting!).

c. For each of the N markers, calculate the MLE of the β parameters under the alternative hypothesis using the IRLS algorithm. For your answer, provide R code that outputs the IRLS calculated estimates for each parameter AND for the same marker you selected in part 'b', write down the estimates you obtained.

d. Why is it a reasonable assumption that the parameter estimates you obtained in part ‘c’ are not exactly the true $MLE(\hat{\beta})$? Also, using no more than one sentence, explain why this is not a problem in practice?

e. For each of the N markers, calculate the LRT. For your answer, provide R code that outputs the LRT for each marker AND for the same marker you selected in part ‘b’ and ‘c’, write down the value of the LRT you obtained.

f. For each of the N markers, calculate the p-value using your LRT calculated in part ‘e’, when assuming this statistic is distributed as a $\chi^2_{df=2}$ under the null hypothesis. For your answer, provide R code that outputs the p-value for each marker AND write down the p-value for the MOST SIGNIFICANT MARKER (make sure you note which marker that is).

g. Plot a Manhattan plot for these N markers (provide your R code as well as the plot).

h. Plot a QQ plot for these N markers (provide your R code as well as the plot).

i. What is an appropriate study-wide Type I error used to assess whether you have evidence of the location of causal polymorphism? Present your cutoff and justify why you selected this value.

j. In no more than three sentences, describe why your results in parts ‘g-i’ indicate that you have identified the location of a causal polymorphism.

Problem 3 (Difficult)

Prove that for the (genetic) logistic regression model, if the null hypothesis $H_0 : \beta_a \cap \beta_d$ is true, this means that $Cov(Y, X_a) = 0 \cap Cov(Y, X_d) = 0$.