BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 4 (version 3) - posted October 3

Assigned October 2; Due 11:59PM October 9

Problem 1 (Easy)

a. For the genetic regression model:

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

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

prove that 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) is entirely a function of the genotype and the β parameters.

b. Write down parameter values for the genetic regression model that would produce a case of 'pure' dominance with Mendelian genetic inheritence (=a single locus where genotype exactly predicts phenotypic value).

Problem 2 (Medium)

For the following question, you are going to simulate and then analyze GWAS data, where these data will include one normally distributed phenotype and $N = 200$ diploid genotypes measured for each of $n = 100$ samples. For many parts of this question, the answer will be your R code presented 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, name your files appropriately, send a zip file, and do not copy someone else's code!! etc.).

a. Write R code to create a 100 x 400 matrix of sample genotypes (i.e. simulate 200 genotypes for each of the 100 individuals, where each genotype is represented by two columns). For each genotype, there should be two alleles (you must represent these with two of 'a', 'c', 't', 'g' OR as 'A1' and 'A2') where all three genotype states (two homozygotes and a heterozygote) should be present for each genotype. The minor allele frequency for each genotype is up to you but should not be < 0.2 for any of the genotypes. Also, each genotype should not be highly correlated with other genotypes, (e.g. don't just repeat the same genotype 200 times!).

b. Write R code to convert your genotype matrix into two new matrices, the first a 100 x 200 matrix where each genotype is converted to the appropriate X_a value and the second a 100 x 200 matrix where each genotype is converted to the appropriate X_d value.

c. Write R code to simulate a 100 x 1 vector of Y values for the 100 individuals where we are assuming there are only two causal genotypes (genotypes number 25 and 150 of your simulated set). Use the following equation to simulate the phenotypes:

$$y_i = \beta_\mu + x_{i,a,25}\beta_{a,1} + x_{i,d,25}\beta_{d,1} + x_{i,a,150}\beta_{a,2} + x_{i,d,150}\beta_{d,2} + \epsilon_i \quad (3)$$

$$\epsilon \sim N(0, \sigma^2_\epsilon) \quad (4)$$

where the $x_{i,a,25}$ and $x_{i,d,25}$ are the random variable codings for the 25th genotype of individual i and the $x_{i,a,150}$ and $x_{i,d,150}$ are the random variable codings for the 150th genotype. Assume the following true parameter values (in real situations, these will be unknown to you!!): $\beta_\mu = 1, \beta_{a,1} = 1.0, \beta_{d,1} = 0, \beta_{a,2} = 0.5, \beta_{d,2} = 0.5, \sigma^2_\epsilon = 1$.

d. Using one sentence, explain why can we simulate the phenotypes in c. by considering only the two causal genotypes and ignoring all others?

e. For each genotype individually, write R code to calculate the $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$. Plot three x-y graphs with the genotypes in order on the x-axis and the parameter estimates on the y-axis = one graph for each of the three parameters.

f. For each genotype individually, write R code to calculate an F-statistic. Plot a histogram of the F-statistic values you obtained.

g. For each genotype individually, calculate a p-value based on your F-statistic and produce a Manhattan plot.

h. Your Manhattan plot will not look like the empirical Manhattan plots we have discussed in class. Using no more than two sentences, describe what is different about your Manhattan plot and what explains why it looks different?

i. Unless you got incredibly unlucky... the two causal genotypes (25 and 150) resulted in the lowest p-values overall (=the highest -log p-values on your Manhattan plot). If you had used a type 1 error $\alpha = 0.05$ to decide whether these were significant or not, would you have rejected the null for these two genotypes? How many other genotypes would you have rejected the null in your analysis? Using no more than two sentences, explain why it is not that surprising that you had additional (incorrect!) cases where you rejected the null at this type 1 error level?

j. For your two causal SNPs (25 and 150) produce two x-y plots for each (=four plots total) with the X_a values on the x-axis and the phenotype Y on the y-axis for the first, and X_d values on the x-axis and the phenotype Y on the y-axis for the second.

Problem 3 (Difficult)

When using a genetic linear regression to model the family of probability distributions that could describe the relationship between phenotypes and genotypes $Pr(Y, X_a, X_d)$, the following are three

equivalent statements of the null and alternative hypotheses when assessing whether a genotype is causal:

$$H_0 : Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d) \tag{5}$$

$$H_A : Pr(Y, X_a, X_d) \neq Pr(Y)Pr(X_a, X_d) \tag{6}$$

$$H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \tag{7}$$

$$H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \tag{8}$$

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \tag{9}$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \tag{10}$$

Prove that these are equivalent. Note that to get full credit, you must show the steps of each derivation and provide a justification for each step (i.e. what rule(s) you are using).

The following may be helpful (!!): note that for arbitrary random variables X_1, X_2, and X_3 that $Cov(X_1, X_2 + X_3) = Cov(X_1, X_2) + Cov(X_1, X_3)$, that for a linear regression, the random variable ϵ is assumed to be independent of the variable(s) X such that $Pr(X_a, \epsilon) = Pr(X_a)Pr(\epsilon)$ and $Pr(X_d, \epsilon) = Pr(X_d)Pr(\epsilon)$, that for the genetic linear regression $Pr(Y|X) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d$, $\sigma^2_\epsilon)$, and that for $Pr(X_a, X_d)$, only three combinations of X_a and X_d have non-zero probabilities.