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

Key to Homework 5 (Version 2 posted March 11) (Version 3 posted March 14)

Assigned March 8; Due 11:59PM March 14

Problem 1 (Easy)

a. Provide a rigorous formula that defines a causal polymorphism and explain intuitively what a causal polymorphism is in terms of biology.

Among the acceptable answers: A causal polymorphism is a location (locus) in the genome where there are at least two alleles, where experimental switching one allele for the other changes the phenotype.

Additional components to the answer that are not necessary for full credit, the formulas:

$$A_1 \rightarrow A_2 \Rightarrow \Delta\bar{Y} \tag{1}$$

$$Cov(X, Y) \neq 0 \tag{2}$$

and that the impact on the phenotype is on average and / or under certain conditions.

b. We have noted that we will use the family of probability models defined by a regression model to describe the possible relationships between random variables X and Y (where one particular parameterization of the regression model is the true = correct model!!). We have noted that this family of probability models has the following structure $Pr(X, Y) = Pr(Y|X)$. Explain what this implies about the uncertainty of X for this family of probability models.

Among the possible answers: By conditioning on X we are assuming that while there is a probability distribution associated with X, our genotypes (represented by X) are being measured without error and once we know the genotype of an individual, we have complete information about the probability of the phenotype Y.

Problem 2 (Medium)

For the following question, you are going to simulate (very unrealistic!) GWAS data and then analyze these GWAS data. These data will include measurements on one normally distributed phenotype and $N = 250$ diploid genotypes (i.e., there will be 250 total polymorphic sites) measured

for each of $n = 200$ 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.).

Note that written answers are provided below and coding answers are in the accompanying Key files provided by Jin.

a. Write R code to create a 200 x 250 matrix of sample genotypes (i.e., simulate 250 genotypes for each of the 200 individuals, where each genotype is represented by a column). For each of the entries for each of the 250 polymorphic sites in your sample (i.e., for each entry of each column), the genotype should be represented by a character 'A1A1', 'A1A2', or 'A2A2'. Simulate each genotype in each column randomly such that the EXPECTED number of 'A1A1', 'A1A2', and 'A2A2' in each column will be $n/4$, $n/2$, $n/4$, respectively (i.e., each column does not have to have these exact genotype frequencies but you should simulate them using a strategy that these are the expected frequencies).

b. Write R code to convert your genotype matrix into two new matrices, the first a 200 x 250 matrix where each genotype is converted to the appropriate X_a value and the second a 200 x 250 matrix where each genotype is converted to the appropriate X_d value.

c. Write R code to simulate a 200 x 1 vector of phenotypes (y values) for the 200 individuals using the following equation to simulate the phenotype of each individual i:

$$y_i = \beta_\mu + x_{i,a,25}\beta_{a,25} + x_{i,d,25}\beta_{d,25} + x_{i,a,100}\beta_{a,100} + x_{i,d,100}\beta_{d,100}$$

$$x_{i,a,175}\beta_{a,175} + x_{i,d,175}\beta_{d,175} + x_{i,a,225}\beta_{a,225} + x_{i,d,225}\beta_{d,225} + \epsilon_i \quad (3)$$

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

where the $x_{i,a,25}$ and $x_{i,d,25}$ are the random variable codings for the 25th genotype of individual i and similarly for the other genotypes. Assume the following true parameter values (in real situations, these will be unknown to you!!): $\beta_\mu = 1, \beta_{a,1,25} = -0.75, \beta_{d,1,25} = 0, \beta_{a,2,100} = 0, \beta_{d,2,100} = -0.75, \beta_{a,2,175} = 0.75, \beta_{d,2,175} = 0.75, \beta_{a,2,225} = 0, \beta_{d,2,225} = 0, \sigma^2_\epsilon = 1.$

d. Plot a histogram of your phenotypes (label your axis!). What probability distribution does your phenotype data resemble (at least approximately)? Explain intuitively why this makes sense. Since the error term is normally distributed, it makes sense that the distribution of phenotypes looks approximately normal.

Not necessary for full credit: the genetic effects are also small enough that we do not expect them to change the shape of the distribution and / or since there are several uncorrelated genetic effects, these will tend to approach a normal distribution (central limit theorem).

e. For each of the four sites 25, 100, 175 and 225 produce two x-y plots for each (=eight 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.

f. Write a 'for loop' to calculate $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$ for each polymorphic site in your simulated dataset. Plot three histograms, one for each of the $\hat{\beta}_\mu, \hat{\beta}_a$, and $\hat{\beta}_d$ parameter estimates across all $N = 250$ genotypes for your entire sample (label your plots!).

g. Write a 'for loop' to calculate an F-statistic for each polymorphic site in your simulated dataset. You should make use of your MLE estimates from part 'e' to calculate your F-statistic using the ratio of MSM and MSE.

h. Use pf(F-statistic, 2, 197, lower.tail = FALSE) to calculate a p-value for each genotype based on your F-statistic calculated in part 'f'. Produce a Manhattan plot (i.e., genotypes in order on the x-axis and -log(p-values) on the y-axis. 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? Why would this be a problematic case?

The Manhattan plot does not have contiguous sets of significant p-values since the genotypes simulated are not correlated (i.e., genotypes near each other are not correlated). This would be problematic in real cases if the causal genotype were not measured, since we depend on testing non-causal genotypes (that are correlated with causal genotypes) to detect the location of causal genotypes.

i. Using a Type I error of 0.05 for your your simulated GWAS data set, report which of your $N = 250$ polymorphic sites were significant. Were the sites 25, 100, 175, 225 among these cases? Explain which of these four sites you expected to find among your significant $N = 250$ sites and explain your reasoning.

Our expectation is that sites 25, 100, and 175 should be among you significant cases but 225 should not and this should have been the result of our analysis (unless you got unlucky).

j. Using a Type I error of 0.05 / 250 for your your simulated GWAS data set, report which of your $N = 250$ polymorphic sites were significant. Were more or less sites significant than for a Type I error of 0.05? Again, were the sites 25, 100, 175, 225 among these cases?

Less total sites were found to be significant than for a higher Type I error of 0.05, where we again expect that sites 25, 100, and 175 should still be among the significant cases but 225 should not (although there is a higher probability that 25, 100, 175 are not significant, since the power of this analysis will be lower).

Problem 3 (Difficult)

In quantitative genomics, the null hypothesis of interest can be stated in the general form $H_0 : Cov(X, Y) = 0$, where if we have random variable X_a and X_d, the null hypothesis is really $H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0$ and when assuming a probability distribution described by a linear regression, the null hypothesis can be expressed as $H_0 : \beta_a = 0 \cap \beta_d = 0$. To see the connection, demonstrate that $Cov(X_a, Y) = 0$ and $Cov(X_d, Y) = 0$ when $\beta_a = 0$ and $\beta_d = 0$. 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 $Pr(X_a, \epsilon) = Pr(X_a)Pr(\epsilon)$ and $Pr(X_d, \epsilon) = Pr(X_d)Pr(\epsilon)$. Show the steps of the derivation and explain the rules you use where appropriate.

Shown for X_a where the same approach is used for X_d:

$$Cov(X_a, Y) = Cov(X_a, \beta_{mu} + \epsilon) \tag{5}$$

since $\beta_a = 0$ and $\beta_d = 0$

$$Cov(X_a, Y) = Cov(X_a, \beta_{mu}) + Cov(X_a, \epsilon) \tag{6}$$

$$Cov(X_a, Y) = Cov(X_a, \beta_{mu}) \tag{7}$$

since covariance of uncorrelated random variables is zero. By the definition of covariance:

$$Cov(X_a, Y) = \mathrm{E}X_a\beta_{mu} - \mathrm{E}X_a\mathrm{E}\beta_{mu} \tag{8}$$

$$Cov(X_a, Y) = \beta_{mu}\mathrm{E}X_a - \beta_{mu}\mathrm{E}X_a = 0 \tag{9}$$

since $\mathrm{E}cX = c\mathrm{E}X$ the expectation of a constant is the constant.