Statistical Genetics, Spring 2024

Homework exercise 3

Due date: 12 August 2024 before last class

1. **Trying out the EM algorithm of Tang et al. (2005)**
   The implementation I showed in class is available at
   http://tau.ac.il/~saharon/StatGen2024/EM.r.
   In this problem we will play with performance of the algorithm on its own and as ancestry correction in regression.

   (a) Play with the dimensions that are expected to influence accuracy: number of markers, differences in marker distributions between the populations, number of individuals of known ancestry from each population, number of individuals of unknown ancestry. Each time fix all but one of these parameters, and show the effect of varying the last one on accuracy. This can be measured in MSE between actual and estimated proportions, or perhaps more accurately with KL divergence. Comment on the results

   (b) Create one more SNP that has disparity between the two populations, and make it "causative" by drawing a disease phenotype that is affected by this SNP (say, logit(P(disease)) = SNP-1, where $SNP \in \{0, 1, 2\}$ is the genotype). Draw a disease status vector, and check that:

      i. Some of the SNPs in the original data are associated with it due to the population stratification

      ii. Those SNPs become less associated – or not at all – once the estimated ancestry is included as a covariate

      iii. The true SNP remains associated with the disease after including estimated ancestry in the regression

   Repeat this exercise several times to confirm consistency of results.

   (c) Compare power to detect the causative SNP with and without ancestry estimation: how often does the true causative have the highest coefficient or smallest p value without ancestry correction? How often with the correction? Discuss your results.

2. **Comparing power of different approaches for finding causative mutations**
   In this problem, we will use simulated data to compare the following approaches for identifying a causative mutation among several correlated ones (without stratification):

   - P-values derived from univariate logistic regression (i.e., a separate model for each SNP).
   - Log-likelihood of univariate logistic regression model
   - P-values of Chi-square tests

   Use the data generator in the file http://tau.ac.il/~saharon/StatGen2024/datagen.r, which generates a sample of X_1, X_2, Y values of length $n = 1000$, where we assume $X_j \in \{0, 1, 2\}$ are SNP genotypes and $Y \in \{0, 1\}$ is the disease status (you can play with n or any other aspect of the data generation). . The true causative mode is "recessive", i.e., $X_1 = 2$ is the risk mode. As modelers we believe that only one of X_1, X_2 is "causative" for Y, but do not know which one, and our goal is to find out that it's X_1, based on a random sample of n examples of X_1, X_2, Y. You can assume the recessive mode is known.

(a) Repeat at least 10000 times the experiment of generating data, performing the three procedures above with appropriately defined variables, and finding how often X_1 is chosen by each method. Also record how often the methods disagree. Comment on the results.

(b) Try several different settings of the parameters to see if your conclusions change.

(c) (*) Prove that the first two methods are completely identical (i.e., always lead to choosing the same variable) in linear regression. Explain why this is not the case for logistic regression.

3. **Joint versus separate modeling**
Assume we have two random covariates X_1, X_2, and a random response Y. Furthermore assume $E(Y|X_1, X_2) = f(X_1)$ (for example: $Y = \beta X_1 + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ in the standard linear regression model). As modelers we believe that only one of X_1, X_2 is "causative" for Y, but do not know which one, and our goal is to find out that it's X_1, based on a random sample of n examples of X_1, X_2, Y. For this we have two possible strategies:

- We can build two univariate models, say two univariate regressions $Y \sim X_1$, $Y \sim X_2$, and choose between them based on which attains smaller p-value for the covariate or higher log-likelihood
- We can build a single regression model with both covariates and choose the one that has lower p-value or higher coefficient.

Choose one of the strategies in each of the following scenarios, and justify your choice – it can be based on (sound!) theoretical arguments or thoughtful simulations.

(a) X_1, X_2 are independent (e.g., two SNPs on separate chromosomes).

(b) X_1, X_2 are positively correlated (e.g., two neighboring SNPs).