# Stat 536 Homework 11

due: 12/5/08 (by dead week policy)

Notice, this homework is worth 60 pts and is your last assignment.

1. **[20 pts] In this problem, you will learn about a population-based estimate of heritability. In lecture we discussed methods to estimate heritability using the covariation of a quantitative trait between fixed relative pairs, for example parents and offspring. When sampling pairs of related individuals from a population, you often cannot know their true relationship. Instead, we will infer the relationship from marker data that is assumed independent of the loci (QTLs) that control the quantitative trait (height in inches). The following is a derivation, and you are asked to fill in a few details and perform data analysis at the end.**

   (a) **[5 pts] The coefficient of coancestry $\rho$ is the probability that two alleles, one drawn randomly from each of two individuals, are identical by descent, i.e. they are the same because they were both passed down from some ancestor of both. The closer two relatives, the higher their $\rho$. Consider locus $l$ and suppose it has $m$ unique alleles. If two relatives have coefficient of coancestry $\rho$, then *show* the probability that two alleles, one drawn randomly from each relative, will both be allele $j$ is**

   $$s_{lj} = \rho p_{lj} + (1-\rho)p_{lj}^2$$

   **where $p_{lj}$ is the frequency of allele $j$ at locus $l$ in the population. (Here, we have assumed that the shared ancestors are not inbred.) Collect these probabilities in a vector $s_l = (s_{l1}, \ldots, s_{lm}, 1 - \sum_{i=1}^m s_i)$.**

   **There are 4 equally likely ways to choose two alleles, one from each of a pair of individuals. Let $S_{li} \sim$ Multinomial$(1, s_l), i \in \{1, 2, 3, 4\}$ be the outcome of the $i$th way to draw alleles. $S_{li}$ is a multinomial random column vector, like**

   $$S_{li} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

   **Element $S_{lij}, j \leq m$ indicates whether the two randomly drawn alleles are identical and both allele $j$. Notice, these $S_{li}$ are identically distributed *but not independent* random vectors. They have sample mean vector $\bar{S}_l$, with $j$th element $\bar{S}_{lj} = \frac{1}{4}\sum_{i=1}^4 S_{lij}$.**

   Solution:

Let $X_1, X_2$ be the random variables indicating the two alleles drawn from the two individuals, then

$$\begin{aligned}
s_{lj} &= P(X_1 = j, X_2 = j) \\
&= P[IBD(X_1, X_2) = 1, X_1 = j] + P(IBD(X_1, X_2) = 0, X_1 = X_2 = j) \\
&= P[X_1 = j \mid IBD(X_1, X_2) = 1]P[IBD(X_1, X_2) = 1] \\
&\quad + P[X_1 = X_2 = j \mid IBD(X_1, X_2) = 0]P[IBD(X_1, X_2) = 0] \\
&= p_j\rho + P[X_1 = j \mid IBD(X_1, X_2)]P[X_2 = j \mid IBD(X_1, X_2)](1 - \rho) \\
&= p_j\rho + p_j^2(1 - \rho)
\end{aligned}$$

where $IBD(X_1, X_2)$ indicates whether $X_1$ and $X_2$ are IBD. Notice if $IBD(X_1, X_2) = 0$, then the alleles are independent so long as there is no inbreeding in the family, i.e. ancestors that provided the two alleles $X_1$ and $X_2$ are not themselve related.

(b) **[5 pts] Derive the following Method of Moments estimator from the sample mean statistic $\bar{S}_{lj}, j \leq m$.**

$$\hat{\rho}_{lj} = \frac{\bar{S}_{lj} - p_j^2}{p_j(1 - p_j)}$$

**There are in fact $m$ such estimates at a locus, and multiple estimates across $L$ loci. To produce an estimate based on locus $l$, one naive approach is to use the sample mean**

$$\hat{\rho}_l = \frac{1}{m}\sum_{j=1}^{m} \hat{\rho}_{lj}$$

**We also need to produce a combined estimate across loci, and again we naively take the sample mean**

$$\hat{\rho} = \frac{1}{L}\sum_{l=1}^{L} \hat{\rho}_l$$

**(More sophisticated approaches include weighted sums, where the weights are estimated to minimize the variance of the estimator or maximum likelihood estimation.)**

Solution:

Notice

$$E(S_{lij}) = s_{lj} = \rho p_j + (1 - \rho)p_j^2$$

for all $i$. Therefore,

$$E\left(\bar{S}_{lj}\right) = \rho p_j + (1 - \rho)p_j^2$$

For a method of moments estimator, we plug in the observed sample mean $\bar{S}_{lj}$ for the expectation and solve for the parameter $\rho$.

$$\begin{aligned}\bar{S}_{lj} &= \hat{\rho}p_j + (1-\hat{\rho})p_j^2 \\ \bar{S}_{lj} - p_j^2 &= \hat{\rho}(p_j - p_j^2) \\ \hat{\rho} &= \frac{\bar{S}_{lj} - p_j^2}{p_j(1-p_j)}\end{aligned}$$

(c) **[10 pts] We showed in class that the regression slope of offspring values on parent values for a quantitative trait is**

$$b_{OP} = \frac{\mathbf{Cov}(P,O)}{\mathbf{Var}(P)} = \frac{V_A}{2V_P} = \frac{1}{2}h^2$$

**where $V_A$ is additive variance, $V_P$ is total phenotypic variance, and $h^2$ is narrow-sense heritability. If alleles contribute additively to a quantitative trait, then this relationship generalizes for any pair of relatives $X$ and $Y$ who have coefficient of coancestry $\rho$ (for parent and offspring $\rho = 0.25$)**

$$Z := \frac{\mathbf{Cov}(X,Y)}{V_P} = 2\rho h^2$$

**Suppose we collect many pairs of individuals $(X_n, Y_n)$ and observe $Z_n = \frac{(X_n-G)(Y_n-G)}{V_P}$, where $G$ is the mean genotypic value. If we know their true relationship, i.e. we know $\rho_n$, then**

$$Z_n = 2\rho_n h^2 + \epsilon_n$$

**where $\epsilon_n$ accounts for error in measuring $Z_n$.**

**If we don't know the true relationship of $X$ and $Y$, then $\rho_n$ is unknown, but we can use marker data to generate an estimate $\hat{\rho}_n$, as described earlier. Then a regression of $Z_n$ on $\hat{\rho}_n$ can yield an estimate of heritability $h^2$. Derive a formula for that estimate in terms of covariances and variances. Then apply your method to the following data:**

- **marker data: 200 individuals (100 relatives pairs, rows 1 & 2 are a pair, 3 & 4 another, etc.) genotyped at 100 loci (1a, 1b are the alleles at locus 1, etc.)**
- **trait data: Heights of 200 individuals (100 pairs)**
- **allele frequencies: Contains an object `p` of allele frequencies. Object `p` is a list of 100 variable length vectors, each containing the frequency of all alleles at that locus. Read the file into R using `load(file="hw11q1-p.Rdata")` to make the object `p` available to your code.**

**[Note. Normally, independent variables, such as $\rho_n$ in this case, are presumed known without error. In our case, they are estimated. If we use these estimates to compute Var($\rho$) directly, this variance estimate includes sampling variance. We will not bother to correct that problem now, but simply note it could negatively affect our estimate of heritability.]**

Solution:

Consider the regression equation

$$Z_n = \beta\rho_n + \epsilon_n$$

A least squares estimate of the coefficient is

$$\hat{\beta} = \frac{\mathrm{Cov}(Z_n, \rho_n)}{\mathrm{Var}(\rho_n)}$$

For the data provided, the estimate is

$$\hat{\beta} = 0.69379$$

(You can also obtain this estimate from a call to `R`'s `lm` function, with which you will also see the large variance on the estimate.)

The above is easily turned into an estimate of heritability as

$$\beta = 2h^2$$

so

$$\hat{h}^2 = 0.346896$$

2. **[20 pt] In this question you will test for association between affect status and various markers in a case/control study. You will use the genomic control method to account for possible unrecognized population structure. Use the Armitage statistic and report the significant loci before and after correction by genomic control. Do you find evidence that any of the markers is associated with affect status? Also please report the genomic control parameter estimate $\hat{\lambda}$. The dataset contains data for a random sample of 500 individuals. The first column is the individual ID, the next 99 columns are diallelic SNP data, reporting the number of copies of one of the alleles at each of 99 loci. The last two columns are a quantitative trait (height, again, but you are not asked about this trait) and affect status.**

Solution:

The Armitage statistic for the $i$th locus is

$$Y_i = N\frac{[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N - R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

where $N$ is the total sample size, $r_1$ is the number of cases with one copy of the marker allele, $r_2$ is the number of cases with two copies of the marker allele, and $n_1$ and $n_2$ are the total numbers of cases and controls with one or two copies of the marker allele.

The genomic control parameter is $\lambda$ and it can be estimated as

$$\hat{\lambda} = \max\left\{1, \left(\frac{\hat{X}_{\text{median}}}{0.675}\right)^2\right\}$$

where $X_{\text{median}}$ is the median of data $X_1, X_2, \ldots$, where $X_i = \sqrt{Y_i}$.

In our case, $\hat{\lambda} = 5.20$. Before genomic control is applied, the list of loci with significant Armitage statistics are

3 6 9 11 18 19 21 25 30 33 35 36 37 42 49 52 54 58 63 65 66 68 78 82 90

To apply genomic control, the modified Armitage statistics are

$$Y' = \frac{Y}{\hat{\lambda}}$$

and only locus 33 remains significant, which was, in fact, the only linked locus by simulation.

3. **[20 pt] The distance between a pair of sequences is the expected number of mutations occurring in evolution between them. We can count the number of differences between them, but some of the actual changes are not observable (e.g. $A \to G \to A$ is two mutations, but produces no differences), so the distance is greater than the number of observed differences.**

   (a) **[5 pts] Compute the pairwise JC69 and K80 distances between these Hepatitis B Virus (HBV) sequences using the method of moments estimators discussed in lecture.**

   Solution:

   **JC69**

   The observed number of differences is $30 + 34 + 17 + 13 + 44 + 13 = 151$, leading to $\hat{p} = \frac{151}{1161}$, because there are 1161 positions. The inversion formula produces

   $$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4\hat{p}}{3}\right) = 0.143$$

   **K80**

   The observed number of transitions is $34+44 = 78$ and transversion is $151-78 = 73$, leading to $\hat{P} = \frac{78}{1161}$ and $\hat{Q} = \frac{73}{1161}$. The estimated distance is therefore

   $$\hat{d} = -\frac{1}{2}\ln\left(1 - 2\hat{P} - \hat{Q}\right) - \frac{1}{4}\ln\left(1 - 2\hat{Q}\right) = 0.146$$

(b) [10 pts] **For JC69, the transition probabilities are**

$$P_{XY}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\mu t} & X = Y \\ \frac{1}{4} - \frac{1}{4}e^{-4\mu t} & X \neq Y \end{cases}$$

**for nucleotides $X$ and $Y$ (notice $\mu$ and $t$ cannot be separately identified, so you can only estimate the product $\mu t$). For K80, the transition probabilities are**

$$P_{XY}(t) = \begin{cases} \frac{1}{4}\left(1 + e^{-4t} + 2e^{-2(\kappa+1)t}\right) & X = Y \\ \frac{1}{4}\left(1 + e^{-4t} - 2e^{-2(\kappa+1)t}\right) & X \neq Y \textbf{ differ by a transition} \\ \frac{1}{2}\left(1 - e^{-4t}\right) & X \neq Y \textbf{ differ by a transversion} \end{cases}$$

**Use `R` to compute maximum likelihood estimates of the parameters, $\widehat{\mu t}$ for JC69 and $\hat{\kappa}, \hat{t}$ for K80. Under JC69, the expected number of mutations in one unit of time is $3\mu$, so the MLE for pairwise distance is $3\widehat{\mu t}$. Under K80, the expected number of mutations in one unit of time is $\kappa + 2$, so the mle for pairwise distance is $(\hat{\kappa} + 2)\hat{t}$. Do the MLE distances differ from the MOM distances?**

Solution:

If there are $d = 151$ differences between the sequences and $n = 1161$ positions, then the log likelihood under the JC69 model is proportional to a Binomial

$$\ln L_{\text{JC69}}(\mu t; n, d) \propto d \ln\left(\frac{1}{4} - \frac{1}{4}e^{-4\mu t}\right) + (n - d)\ln\left(\frac{1}{4} + \frac{3}{4}e^{-4\mu t}\right)$$

which can be maximized with a call to `optim(`$\mu t$`, `$\ln L_{JC69}$`, n=`$n$`, d=`$d$`)`.

Similarly, if there are $d_i = 78$ transitional differences and $d_v = 73$ transversional differences, then the log likelihood under the K80 model is proportional to a multinomial

$$\begin{aligned} \ln L_{\text{K80}}(\kappa, t; n, d_i, d_v) \;\propto\; & d_i \ln\left(\frac{1}{4}\left(1 + e^{-4t} - 2e^{-2(\kappa+1)t}\right)\right) \\ & + d_v \ln\left(\frac{1}{2}\left(1 - e^{-4t}\right)\right) \\ & + (n - d_i - d_v)\ln\left(\frac{1}{4}\left(1 + e^{-4t} + 2e^{-2(\kappa+1)t}\right)\right) \end{aligned}$$

which can be maximized with a call to `optim(c(`$\kappa$`, `$t$`), `$\ln L_{K80}$`, n=`$n$`, di=`$d_i$`, dv=`$d_v$`)`.

The MLEs for the distances are

$$\begin{aligned} \hat{d}_{\text{JC69}} &= 3\widehat{\mu t} = 0.138 \\ \hat{d}_{\text{K80}} &= (2 + \hat{\kappa})\hat{t} = 0.110 \end{aligned}$$

(c) **[5 pts] Is there evidence of distinct rates of transition and transversion? You may find it easier to work with the statistical summaries of the data in the following table. These are the counts of the number of mutations of each type observed between the two sequences.**

| | **A** | **C** | **G** | **T** |
|---|---|---|---|---|
| **A** | **205** | **30** | **34** | **17** |
| **C** | **0** | **302** | **13** | **44** |
| **G** | **0** | **0** | **225** | **13** |
| **T** | **0** | **0** | **0** | **278** |

Solution:

The likelihoods evaluated at their mles above are

$$\begin{aligned} \ln L_{\text{JC69}}(\widehat{\mu t}; n, d) &= -614.618 \\ \ln L_{\text{K80}}(\hat{\kappa}, \hat{t}; n, d_i, d_v) &= -553.310 \end{aligned}$$

which allows quick evaluation by a log likelihood ratio test.

$$-2(\ln L_{\text{JC69}} - \ln L_{\text{K80}}) = 122.62$$

which is clearly very significant evidence against JC69. Just looking at the count data would hint at this result, because counts of $A/G$ and $C/T$ differences are the two highest. Also, there is evidence in the count data that the four nucleotides are not equally likely, with $C$ being substantially more common than the others.