Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#9:(Apr-06-2010)
Genome Wide Association Studies

Class Projects

- Few Ideas for the class projects:
  1. GWAS – WTCCC Study: See the URL: http://www.nature.com/nature/journal/v447/n7145/full/nature05911
  2. Mendelian Diseases: See the URL: http://www.nature.com/nature/journal/v461/n7261/full/nature08250
  3. Indian Population: See the URL: http://www.nature.com/nature/journal/v461/n7263/abs/nature08365
  4. Mutation Rates in Humans: See URL: http://www.pnas.org/content/107/3/961.abstract
  5. Quartet Analysis: See URL: http://www.sciencemag.org/cgi/content/abstract/science.1186802

Outline

Generalized Linear Models

- Fit a multivariate model to either quantitative or discrete/binary traits. Association of traits and genotypes with or without consideration of additional covariates...
- Distinct from classical stratified univariate analysis — one for each stratum: e.g., smoking status.
- **GLM**: generalized Linear Models, given in matrix notation by the following equation:

$$g(E[\mathbf{y}]) = \mathbf{X}\beta,$$

where $E[\mathbf{Y}] = \mu$ denotes the expectation of $\mathbf{Y}$, $g(\cdot)$ is a *link function* (usually identity or logit) and $\mathbf{X}$ is the *design matrix*.

Multivariate Regression

- Simplest model: $g(\cdot)$ is the identity link, y is a quantitative trait and x is a single genotype (e.g., a SNP)

$$g(E[\mathbf{y}]) = E[\mathbf{y}] = \mathbf{X}\beta,$$

or equivalently,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

- Assume that there are n samples; then

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}; \text{ and } \beta = (\beta_0, \beta_1)^T.$$

Scalar Formulation

- Thus

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

$i = 1, \ldots, n$ indicates individuals. We assume the error terms, ϵ_i to be distributed i.i.d (independent and identically distributed) with mean 0.

- The measure of association is given by the parameter β_1 – defined as the amount of change in y that occurs with one unit of change in x

$$\widehat{\beta_1} = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

and

$$\widehat{\beta_0} = \left(\sum_i y_i - \widehat{\beta_1} \sum_i x_i \right) / n.$$

Interpretation of $\widehat{\beta_1}$

- Note that

$$\begin{aligned}\widehat{\beta_1} &= \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \\ &= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]} = r_{xy}\frac{s_y}{s_x},\end{aligned}$$

where r_{xy} is the *correlation coefficient* between x and y; s_x (resp. s_y) is the *standard deviation* of x (resp. y).

Solution to Linear Regression

- Can be expressed in terms of pseudo- (Penrose-Moore) inverse:

$$\begin{aligned} \mathbf{y} - \mathbf{X}\beta &= \epsilon \\ \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\beta &= \mathbf{X}^T\epsilon \\ \widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

- Thus

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i y_i\right).$$

Covariates

- Suppose we have m covariates, given by $z_{i1}, z_{i2}, \ldots, z_{im}$ for the ith individual:

$$y_i = \beta_0 + \beta_1 x_i + \sum_j \alpha_j z_{ij} + \epsilon_i .$$

- The measure of association between the genotype and trait is given by β_1... while taking into account the additional variables in the model.
- The additional variables may explain the variability in the trait better ... *or they may have several confounders*.

Interactions

- We may model the interactions between the genotypes and the covariates... *nature-nurture interactions*
- Example: Interactions between genotypes and the drug exposure and its phramaco-genomic effects on the trait...
- Let genotypes be represented by x and drug exposure by z. Let the quantitative trait be defined by y:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \gamma x_i z_i + \epsilon_i$$

- γ is the interaction effect and represents the additional effect of z for a particular genotype x

Example

- In the previous model, we may have: x is a polymorphism in ApoCIII gene — involved in triglyceride levels; z corresponds to the current exposure to lipid lowering therapy (LLT). y is fasting glyceride level — a quantitative trait.
- The effect of LLT on triglyceride level in β_2 among individuals without ApoCIII polymorphism ($x_i = 0$) and is $\beta_2 + \gamma$ among individuals with ApoCIII polymorphism ($x_i = 1$)

Solution

- The model in the matrix form:

$$E[\mathbf{y}] = \mathbf{X}\beta,$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & z_1 & (x_1 \times z_1) \\ 1 & x_2 & z_2 & (x_2 \times z_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & z_n & (x_n \times z_n) \end{bmatrix}; \text{ and } \beta = (\beta_0, \beta_1, \beta_2, \gamma)^T.$$

- The solution is:

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

Multiplicative Models

- Multiplicative effects can be modeled easily, by using $\ln(\cdot)$ as the link function:

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

or equivalently

$$y_i = e^{\beta_0} e^{\beta_1 x_i} e^{\beta_2 z_i} e^{\epsilon_i}.$$

- Here the effects of x and z are multiplicative on y... A unit change in x results in e^{β_1}-fold increase in y; Similarly, a unit change in z results in e^{β_2}-fold increase in y;

Logistic Regression

- **Application to a binary trait**
- The link function $g(\cdot)$ is the logit$(\cdot)$ function.

$$\text{logit}(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}.$$

- For a random variable **y** from a Bernoulli trial

$$E[\mathbf{y}] = Pr(\mathbf{y} = \mathbf{1}_n) = \pi = (\pi_1, \pi_2, \ldots, \pi_n)^T.$$

- Thus the model is

$$g(E[\mathbf{y}]) = \text{logit}(\pi) = \mathbf{X}\beta,$$

Logistic Regression

- Simplifying the model

$$g(E[\mathbf{y}]) = \text{logit}(\pi) = \mathbf{X}\beta,$$

we get

$$\ln[\pi_i/(1-\pi_i)] = \beta_0 + \beta_1 x_i,$$

or

$$\pi_i = \frac{e_0^{\beta} e^{\beta_1 x_i}}{1 + e_0^{\beta} e^{\beta_1 x_i}}.$$

- The parameter β_1 is interpreted as "*the effect of a unit increase in x on the log-odds of disease y*."

Caveats

- Overfitting: Number of predictors (degrees of freedom) should be small: Limiting the model to include at most one predictor for every five to ten observations for quantitative trait — or — events for binary traits!
- Avoiding correlated predictor variables. Inclusion of all SNPs for analysis within a single model may not be tenable.
- Model Selection: Eliminate confounding variables by testing on SNP at a time; Shrinkage/Truncated Shrinkage; Cross-Validation;
- Correction for Multiple Hypothesis Testing.

Outline

Multiplicity and High Dimensionality

- **Curse of Dimensionality** A term due to Richard Bellman — GWAS with SNPs involve millions of dimensions; while the data (for humans) is bounded by 6 billion!
  1. *Inflation of Error Rates* — Primarily due to Multiple Hypothesis Testing
  2. *Complex and Unknown Relationship among the Genetic Markers*
- **Model Selection**: Degree of Freedom of the Model vs. Sample Size

Error Inflation

- We wish to reject a *null hypothesis*: H_0, if we are sure that the *alternative hypothesis*: H_1 is in fact correct.
- **False Positive**: *Rejecting the null-hypothesis in favor of alternative, when in fact the null is true*... Also called *type-error*
- If we wish to control the type-error rate (fdr: false-discovery rate) below some threshold α, then we must ensure that

$$\text{type-1 error rate} = Pr(\text{Reject } H_0 | H_0 = \text{true}) \leq \alpha.$$

p-value

- p-value is for a given hypothesis is determined based on a sample of data and is defined as the probability of observing something as extreme or more extreme, given the null is true:

$$p\text{-value} = Pr(\text{Data } D | H_0 = \text{true}).$$

- If p-value is less than α (e.g., 0.05), then we may reject the null hypothesis in favor of the alternative.

Multiple Hypotheses Testing

- We wish to test K different null hypotheses:

$$H_{01}, H_{02}, \ldots, H_{0k}, \ldots, H_{0K}, \text{ for } k = 1, \ldots, K.$$

- *Family-Wise Error under the Complete Null* (FWEC) is defined as the probability of rejecting at least on null, when all the nulls are in fact true.

$$\begin{aligned} FWEC &= Pr(\text{Reject at least one } H_{0k} | H_{0k} = \text{ true } \forall k) \\ &= 1 - Pr(\text{Reject no } H_{0k} | H_{0k} = \text{ true } \forall k) \\ &= (1 - (1-\alpha)^K) \approx 1 - e^{-\alpha K}. \end{aligned}$$

- For $\alpha = 0.05$ ... .

K	FWEC
1	0.05
2	0.0975
10	0.401

Multiple Hypothesis Testing

- As the number of hypotheses increases, so does *FWEC* – a phenomenon called *inflation* of the type-1 error rates.
- Inflation is a serious problem in any GWAS that tries to find association between a large number of SNPs and a trait.
- We need to develop methods to control (1) *family-wise error rates* and (2) *false discovery rates*.

Interaction between Genotypes

- Another Challenge: SNPs are likely to interact (through epistasis and linkages) with one another in a manner that is not well-characterized. The genes affected by the SNPs may belong to the same pathway; the SNPs may affect the structure of the protein they code; they may affect a gene's regulation, etc. The SNPs may act differently in the presence of a varying covariate.
- The Model: A sample of n individuals; M measured SNPs — denoted for individual i by $x_{i1}, \ldots, x_{iM}$. x is a binary indicator for the presence of at least one copy of the minor/mutant allele. Assume that SNPs have an additive effect on the trait, but no interaction:

$$y_i = \beta_0 + \sum_{j=1}^{M} \beta_j x_{ij} + \epsilon_i$$

Interactions

- Adding pair-wise interactions to the model:

$$y_i = \beta_0 + \sum_{j=1}^{M} \beta_j x_{ij} + \sum_{k,l,k\neq l} \gamma_{kl} x_{ik} x_{il} \epsilon_i$$

- In the simpler model (without interactions), there are M null-hypotheses $H_{0j} : \beta_j = 0$ ($j = 1, \ldots, M$) saying that jth SNP has no effect on the trait.
- In the more complex models (with interactions), there are now $\binom{M}{2}$ new null hypotheses to account for.
- Thus the complex model makes the possibility of *inflation* or *overfitting* much more likely — with a higher FWEC.

Missingness

- *Missing and Unobservable Data*:
  1. Rare alleles are difficult to genotype. The frequency estimates are incorrect.
  2. Alignment of alleles on a single homologous chromosome is difficult to infer. *Haplotype Phasing Problem*.

Haplotype Phasing Problem

- Two alleles on the same homologous chromosome are said to be *in cis* — Two alleles on opposite sister homologs are said to be *in trans*.
- A particular combination of alleles on a single homologous chromosome is called a *haplotype*.
- With $(k+1)$ biallelic SNPs, the population can have 2^k possible *haplotypes*, though most of them are likely to be missing.

Haplotype Phasing Problem

- Note that the diploid pair of haplotypes is of the order 2^{2k}:

$$\binom{2^k}{2} + 2^k,$$

the first term corresponding to heterozygous haplotypes and the second corresponding to homozygous haplotypes.

- When $k = 2$, there are four haplotypes: (AB, aB, Ab, ab) and ten diplotypes

$$(AB, AB), (aB, aB), (Ab, Ab), (ab, ab),$$

$$(AB, aB), (AB, Ab), (AB, ab), (aB, Ab), (aB, ab), \text{ and } (Ab, ab).$$

Penetrance

- It is possible to infer a likely haplotype from the genotype data, if we know the LD-structure for the population.
- However, this is further confused by two other effects:
  1. **Penetrance**: The presence of a disease alleles does not lead to the disease phenotype.
  2. **Phenocopies**: Individuals exhibiting disease phenotypes do not carry the allele under consideration.

Genetics Models and Models of Association

- We have considered additive and multiplicative models of association.
- **Genetic Models**: They describe the biological interaction between alleles on a homologous chromosome.
  1. **Additive Model**: k ($k = 0, 1, 2$) copies of T allele increases the trait y by an amount $k\beta$:

$$y_i = \alpha + \beta[I_{x_{i,1}=T} + I_{x_{i,2}=T}] + \epsilon_i$$

  2. **Dominant Model**: Having one or more copies of T allele increases the trait y by an amount β:

$$y_i = \alpha + \beta[I_{x_{i,1}=T \vee x_{i,2}=T}] + \epsilon_i$$

  3. **Recessive Model**: Both homologs must have copies of T allele to increase the trait y by an amount β:

$$y_i = \alpha + \beta[I_{x_{i,1}=T \wedge x_{i,2}=T}] + \epsilon_i$$

M-sample test for Quantitative Traits

- **Two Sample t-Test**: Consider two populations: e.g., (1) one with alleles AA and (2) the other with alleles (Aa, aa).
- Test for the null hypothesis that the mean of the traits for the two populations are the same: $H_0 : \mu_1 = \mu_2$.

$$t = \frac{\bar{y_1} - \bar{y_2}}{\sqrt{s_p^2[1/n_1 + 1/n_2]}} \sim T_{n_1+n_2-2},$$

  where $\bar{y_1}$ and $\bar{y_2}$ are the sample means of the quantitative trait for genotype groups (1) and (2); s_p is the pooled estimate of variance, and n_1 and n_2 are the sample sizes.
- This statistic has a T-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Other Tests

- Wilcoxon Rank-Sum Test
- ANOVA (analysis of Variance)
- Kruskal-Wallis (KW) Test

Model Selection

- Goal is to select a small number of SNPs to build a model: These should be causal SNPS or Tag SNPs in LD with causal SNPs.
- **Bayesian Variable Selection**: Start with a General Linear Model for Genotype-Trait Association:

$$y_i = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_r x_{ir}^* + \epsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

where $(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_r^*)$ is a subset of potential indicator variables, $\mathbf{y}$ is a quantitative trait.

Model Selection

- For the coefficients assume that they are either *relevant* or *nuisance* variables, described by a mixture model:

$$\beta_j|\gamma_j \sim (1-\gamma_j)\mathcal{N}(0,\tau_j^2) + \gamma_j\mathcal{N}(0,c_j^2\tau_j^2),$$

where $gamma = (\gamma_1,\ldots,\gamma_p)$ is a latent (unobservable) vector with elements taking values 0 or 1.

$$Pr(\gamma_j = 1) = p_j, \text{ and } Pr(\gamma_j = 0) = 1 - p_j = q_j,$$

- For the variance in the selected coefficients, we can choose:

$$\sigma^2|\gamma \sim \mathcal{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2),$$

given by an inverse gaussian (Wald) distribution $\mathcal{IG}$.

Distributions

- Gaussian/Normal:

$$X \sim \mathcal{N}(\mu, \sigma)$$

then

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}.$$

- Wald:

$$X \sim \mathcal{IG}(\mu, \lambda)$$

then

$$f(x; \mu, \lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}.$$

Putting it all together

- We now have

$$\mathbf{y}|\beta, \sigma^2 \sim \mathcal{MVN}_n(\mathbf{X}\beta, \sigma^2 I),$$

  where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X}_{n \times p} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ and $\beta = (\beta_1, \ldots, \beta_p)^T$.
- The parameters corresponding to the ONLY true underlying predictors $(\mathbf{x}_1^*, \ldots, \mathbf{x}_r^*)$ are non-zero.

Bayesian Formulation

- Putting everything together,

$$\pi(\gamma|\mathbf{Y}) \propto f(\mathbf{Y}|\beta,\sigma^2)f(\beta|\gamma)f(\sigma^2|\gamma)\pi(\gamma).$$

- We can find the best estimator for γ by Gibb's sampling from the marginal posterior densities for β, σ and γ_j.

Bayesian Variable Selection

Algorithm 1: BVS - pseudocode

Input: Traits $\mathbf{Y}$ and SNPs $\mathbf{x}_i$

Output: Subset of prdictive SNPs $\mathbf{x}_i^*$

1. Initialize β, σ and γ — denoted as $\beta^{(0)}$, $\sigma^{(0)}$ and $\gamma^{(0)}$
2. Let $t = t + 1$ and sample
   - $\beta^{(t)}|\mathbf{y} \sim f(\beta|\mathbf{y}, \sigma^{(t-1)}, \gamma^{(t-1)})$
   - $\sigma^{(t)}|\mathbf{y} \sim f(\sigma|\mathbf{y}, \beta^{(t-1)}, \gamma^{(t-1)})$
3. Randomly select an ordering $\gamma_{(1)}, \ldots, \gamma_{(p)}$ and sample
   - $\gamma_{(1)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(2)}^{(t-1)}, \ldots, \gamma_{(p)}^{(t-1)})$
   - $\gamma_{(2)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(1)}^{(t)}, \gamma_{(3)}^{(t-1)}, \ldots, \gamma_{(p)}^{(t-1)})$
   - $\vdots$
   - $\gamma_{(p)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(1)}^{(t)}, \ldots, \gamma_{(p-1)}^{(t)})$
4. Repeat the steps (2) and (3) M times for a large M.
5.

[End of Lecture #9]