GENETIC ANALYSES USING FAMILY-BASED SURVEY DATA

Yan Li
Joint Program for Survey Methodology
University of Maryland at College Park

yli6@umd.edu

4th Baltic-Nordic Conference on Survey Statistics
Aug 25, 2015
National Genetic Household Surveys (NGHS)

- Conducted in various countries
  - e.g. Health 2000 Survey from Finland (Heistaro, 2008);
    Canadian Health Measures Survey (Tremblay et al., 2007);
    U.S. Health and Retirement Study;
    National Health and Nutrition Examination Surveys (NHANES)
    - Phenotypic, environmental and behavioral data
    - Various types of genetic data

- Less bias in NGHS comparing to traditional genetic studies
  - NGHS: random samples representing well-defined populations
  - Traditional genetic studies: volunteers or convenience sample
• Correlation among families due to multistage geographical cluster sampling

• Correlation within families because of biological inheritance

• Differential sampling Weights
OUTLINE

PART I: Hardy-Weinberg Equilibrium Tests

PART II: Genetic Association Studies with Complex Designs
PART I: Hardy-Weinberg Equilibrium Tests

Hardy Weinberg Equilibrium (HWE)

In the case of a single locus with two alleles A and a:
Frequencies of allele A and allele a: \( f(A) = p_A; f(a) = p_a \)

Under ideal conditions,
Hardy Weinberg Equilibrium will be reached after one generation of random mating, i.e., the genotype frequencies remain same:
\( f(AA) = p_A^2; f(Aa) = 2p_A p_a; f(aa) = p_a^2 \)
Why Testing HWE is Important?

• Departure from HWE - infer the existence of natural selection, mutation, migration, assertive (non-random) mating, otherwise infer genotyping errors.

• In Genetic Association Studies

  Preliminary step before testing for association between the alleles and disease (Salanti et al., 2005; Zou, 2006; Zou & Donner, 2006)

• HWE is often an assumption in studies testing association of gene-environment interactions with diseases (Chatterjee and Carroll, 2005)
HWE Testing Methods for NGHS

- Y. Li et al. (2009), Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data. *Biometrics* 65, 1096-104.

- Correlation due to multistage cluster sampling
- Differential weighting
How to take account of genetic correlation within families?
METHODS – HWE TESTS

Notations:

\[ H \text{ strata} \]
\[ \downarrow \]
\[ I_h \text{ PSUs sampled in the stratum } h \]
\[ \downarrow \]
\[ J_{hi} \text{ families sampled in PSU-}hi \]
\[ \downarrow \]
\[ K_{hij} \text{ individuals in family-}hij \]
For a locus with M alleles \((X_1, ..., X_m, ..., X_M)\)

- \(p_m = \Pr\{X_m\}\): Frequency of allele \(X_m\)
- \(p_{mm'} = \Pr\{X_mX_{m'}\}\): Frequency of genotype \(X_mX_{m'}\)
- \(G = \frac{(M+1)M}{2}\): the number of possible distinct genotypes

For example, for a locus with 2 alleles A and a,

\[ M = 2 \]

Allele frequencies: \(p_A\) and \(p_a\), \(p_A + p_a = 1\)

Genotype frequencies: \(p_{AA}, p_{Aa}, p_{aa}\), with \(p_{AA} + p_{Aa} + p_{aa} = 1\)

\[ G = \frac{(M+1)M}{2} = \frac{(2+1)2}{2} = 3 \]
• $\mathbf{y}_{hijk} = (y_{hijk,1}, \ldots, y_{hijk,g}, \ldots, y_{hijk,G-1})^T$
  
  genotype indicators for individual $hijk$ with
  
  $y_{hijk,g} = \begin{cases} 
  1 & \text{if the genotype of individual } hijk \text{ is } g \\
  0 & \text{Otherwise}
  \end{cases}$

• $\mathbf{\mu}_{hijk} = (\mu_{hijk,1}, \ldots, \mu_{hijk,g}, \ldots, \mu_{hijk,G-1})^T$, where

  $\mu_{hijk,g} = \begin{cases} 
  (1 - r)p_l^2 + rp_l & \text{if the genotype } g = l/l \\
  2(1 - r)p_l p_{l'} & \text{if the genotype } g = l/l'
  \end{cases}$

• $r$ : Fixation coefficient to characterize the departure from HWE – correlation between two alleles in an individual.

  Under HWE $H_0$: $r = 0$
Pseudo Score Function – Individual-based

\[ S(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \sum_{k=1}^{K_{hij}} \frac{\partial \mu_{hijk}(\theta)}{\partial \theta} W_{hijk} \text{Var}^{-1}(y_{hijk})(y_{hijk} - \mu_{hijk}(\theta)), \]

where

\[ W_{hijk} = \begin{bmatrix} \vdots & \cdots & 0 \\ \vdots & w_{hijk} & \vdots \\ 0 & \cdots & \ddots \end{bmatrix}_{(G-1)(G-1)} \]

Inverse of the selection probability

\[ \text{Var}(y_{hijk}) - \text{covariance matrix of } y_{hijk} \]

Working correlation among members within families – Independent
To take account of genetic correlations within families

Pseudo Score Function – Family-based

\[
S(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \frac{\partial \mu_{hij}(\theta)}{\partial \theta} w_{hij}^{1/2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2} (y_{hij} - \mu_{hij}(\theta)),
\]

where

\[
y_{hij} = (y_{hij1}, ..., y_{hijk}, ..., y_{hijk_{hij}})^T \text{ across selected family members}
\]

\[
\mu_{hij} = E(y_{hij}) = (\mu_{hij1}, ..., \mu_{hijk}, ..., \mu_{hijk_{hij}})^T
\]
Pseudo Estimating Equations

\[
S(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \frac{\partial \mu_{hij}(\theta)}{\partial \theta} \frac{1}{2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2} (y_{hij} - \mu_{hij}(\theta)),
\]

where

\[w_{hij}\] is sample weight matrix for family-hij with diagonal involving sample weight for each selected family member

\[
w_{hij} = \begin{bmatrix}
w_{hij1}I_{G-1} & 0 & 0 \\
0 & \ddots & 0 \\
0 & 0 & w_{hijK_{hij}}I_{G-1}
\end{bmatrix}
\]

\(I_{G-1} = (G-1) \text{ dimensional identity matrix}\)
Pseudo Estimating Equations

$$S(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \frac{\partial \mu_{hij}(\theta)}{\partial \theta} w_{hij}^{1/2} \text{Var}^{-1}(y_{hij}) w_{hij}^{1/2} (y_{hij} - \mu_{hij}(\theta)),$$

where $\text{Var}(y_{hij})$ -- genetic correlation within family-$hij$
For example, consider family-$hij$ with 1 parent (P) and 2 offspring $(O_1, O_2)$ and locus with allele $A$ and allele $a$

$$Var(y_{hij}) = \begin{bmatrix} \Sigma_P & \Sigma_{P,O_1} & \Sigma_{P,O_2} \\ \Sigma_{O_1} & \Sigma_{O_1,O_2} \\ SYS & SYS & \Sigma_{O_2} \end{bmatrix},$$

where

$$\Sigma_P = \Sigma_{O_1} = \Sigma_{O_2}: 2 \text{ by } 2 \text{ covariance matrices between the indicators of genotypes in the same individual}$$

$$\begin{bmatrix} p_A^2(1 - p_A^2) & -p_A^2 \cdot 2p_Ap_a \\ SYS & 2p_Ap_a \cdot (1 - 2p_Ap_a) \end{bmatrix}$$
\[\Sigma_{P,O_1} = \Sigma_{P,O_2}: \text{covariance between parent and offspring}\]

\[
\begin{bmatrix}
    p_A^3 - p_A^4 & p_A^2 p_a - p_A^2 \cdot 2 p_A p_a \\
    SYS & p_A^2 p_a + p_a^2 p_A - (2p_A p_a)^2
\end{bmatrix}
\]

\[\Sigma_{O_1O_2}: \text{covariance between full siblings}\]

\[
\begin{bmatrix}
    \frac{1}{4} p_A^2 + \frac{1}{2} p_A^3 - \frac{3}{4} p_A^4 & \frac{1}{2} p_A^2 p_a - \frac{3}{2} p_A^3 p_a \\
    SYS & p_A p_a - 3p_A^2 p_a
\end{bmatrix}
\]

\(\Sigma\)'s are functions of coefficient of condensed identities (CCI), and depend on the family relationship between the pair of individuals

(Lange, 2002 on page 82)
Pseudo Estimating Equations $S(\theta) = 0$:

- Unknown Parameters: $\theta = (p, r)^T$
- $S(\theta) = (S_p^T, S_r^T)^T$
Quasi-score test statistic:

\[ TS_1 = \hat{S}_r^T (\widetilde{\theta}) \mathbf{Var}^{-1} (\hat{S}_r) \hat{S}_r^T (\widetilde{\theta}), \]

where

\[ \widetilde{\theta} = (\hat{p}^w, r = 0)^T \text{ – The solution to } S_p (\widetilde{\theta}) = 0 \text{ under } H_0 \]

\[ \mathbf{Var} (\hat{S}_r) \text{ – Consistent estimator of } \mathbf{Var} (\hat{S}_r) \]

By Taylor linearization method (Rao et al., 1998)

\[ \mathbf{Var} (\hat{S}_r) = \sum_{h=1}^{H} \frac{I_h}{I_h - 1} \sum_{i=1}^{I_h} (z^{hi} - \bar{z}^h)(z^{hi} - \bar{z}^h)^T, \]
where

$$z^{hi} = \sum_{j=1}^{J_{hi}} \left( \frac{\partial \mu_{hij}}{\partial r} - I_{21}I_{11}^{-1} \frac{\partial \mu_{hij}}{\partial p} \right) w_{hij}^{1/2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2} (y_{hij} - \mu_{hij}),$$

$$I_{21} = \frac{\partial}{\partial p} S_r(\theta) =$$

$$\sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \left\{ - \left( \frac{\partial \mu_{hij}}{\partial r} \right) w_{hij}^{1/2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2} \left( \frac{\partial \mu_{hij}}{\partial p} \right)^T \right\},$$

and

$$I_{11} = \frac{\partial}{\partial p} S_p(\theta) =$$

$$\sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \left\{ - \left( \frac{\partial \mu_{hij}}{\partial p} \right) w_{hij}^{1/2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2} \left( \frac{\partial \mu_{hij}}{\partial p} \right)^T \right\},$$

evaluated at $\theta = \bar{\theta}$, and $\bar{z}^h = \frac{1}{l_h} \sum_{i=1}^{l_h} z^{hi}$.

Under suitable conditions (Rao et al. 1998),

$$TS_1 = \bar{S}_r^T(\bar{\theta}) \text{Var}^{-1}(S_r) \bar{S}_r^T(\bar{\theta}) \sim \chi^2_{(1)}$$
**Simulations** show that the developed HWE test $TS_1$:

- Maintain the nominal level
- Achieve higher power than the test ($TS_2$) that ignores the genetic correlation within families

**Limitations:**

Within-family sampling depends on

\[
\begin{align*}
\begin{cases}
\text{family relationship (e.g. 1P2O, 30, etc)} & \checkmark \\
\text{genotype related factors} & \times
\end{cases}
\end{align*}
\]

\[w_{hij}^{1/2} \text{Var}^{-1}(y_{hij})w_{hij}^{1/2}\]
To fix the problem, we use the Pseudo Score Function based on the pairwise scores

\[ S(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} w_{hij} S_{hij} = 0, \]

with

\[ S_{hij} = \sum_{k=1}^{K_{hij}} \sum_{l=1}^{K_{hij}} \frac{1}{\pi_{kl|hij}} \frac{\partial \mu_{hij}(\theta)}{\partial \theta} \operatorname{Var}^{-1}(y_{hij})(y_{hij} - \mu_{hij}), \]
where

- $\pi_{kl|hij}$ – Joint inclusion probability for pair $(k, l)$ given family hij is sampled
- $\mathbf{y}_{hij} = (\mathbf{y}_{hijk}, \mathbf{y}_{hijl})^T$ – a vector of indicators of genotypes for pair of individuals $(k, l)$ in family hij
- $\mathbf{\mu}_{hij} = (\mathbf{\mu}_{hijk}, \mathbf{\mu}_{hijl})^T = \mathbb{E}(\mathbf{y}_{hij})$

**Quasi-score test statistic (Rao et al. 1998):**

~ derived along the same line as above:

$$TS_p = \hat{S}_r^T (\hat{\theta}) \mathbb{V}ar^{-1}(\hat{S}_r) \hat{S}_r^T (\hat{\theta}) \sim \chi^2_1$$
Simulations Studies

Population Generation

• 10,000 PSUs with each PSU composed of 40 families

• Generate genotype
  - Consider a biallelic locus (A, a)
  - \( p_A = p_a = 0.5; \ r = 0, 0.1, 0.15, 0.2 \)
    - Parents: multinomial distribution with specified genotype frequencies
      \[ p(AA) = (1 - r)p_A^2 + rp_A; \ p(Aa) = 2(1 - r)p_ap_a; \ p(aa) = (1 - r)p_a^2 + rp_a \]
    - Offspring: randomly generated according to Mendelian law

• Population clustering
  Sort all families by \#(aa). The 10,000 PSUs are then formed by grouping every 40 families sequentially.
**Sampling Designs**

- **Stage 1:** sample 100 PSUs
  - Simple random sampling (srs)
  - Proportional to population size sampling (pps)
    - The measure of size related to genotypes, psu’s with more #aa is oversampled

- **Stage 2:** Sample family members - stratified SRS (SSRS) with stratum defined by
  - Family relationship – SSRS(F)
  - Family relationship & genotype – SSRS(GF)
    - Oversample genotype aa
**Test statistics**

- **TS$_1$**
  - Based on quasi scores at the family level.
  - Considers genetic correlation within families.

- **TS$_2$**
  - Based on quasi scores at the family level.
  - Does NOT consider genetic correlation within families.

- **TS$_p$**
  - Based on quasi pairwise scores within families.
Evaluation Criteria

• RelBias of $\hat{p}_A$
  \[
  \text{RelBias} (\hat{p}_A) = \left[ \text{mean} (\hat{p}_A) - p_A \right] / p_A \times 100\%
  \]

• Variance ratios
  o Analytical variance = Mean of 1,000 estimates of $\bar{\text{Var}} L S_r (\tilde{\theta})$
  o Empirical variance = Variance of 1,000 estimates of $\overline{p}_A$

  \[VR = \text{Analytical variance} / \text{Empirical variance}\]

• Rejection Rates at nominal level 5%

  % rejecting $r = 0$ in 1,000 HWE test
  • Under H0 ($r = 0$): test size
  • Under H1 ($r > 0$): power
Figure 1: Results of 3 HWE tests Under SSRS(F) at 5% nominal level.
Figure 2: Results of 3 HWE tests Under SSRS(GF) at 5% nominal level.
IN SUMMARY,

➢ If within family sampling variables ⊥ Genotypes

$TS_1$ produces approx. unbiased estimate of allele frequencies, maintains the nominal level at the null hypothesis and achieves the highest power under alternative hypothesis

➢ If within family sampling variables related Genotypes

$TS_p$ produces approx. unbiased estimate of allele frequencies, maintains the nominal level at the null hypothesis and achieves the highest power under alternative hypothesis
Conclusions of HWE Tests

• Considers both levels of correlations.
• Considers differential sampling weights

When the within-family sampling is independent of genotypes/disease status:


When the within-family sampling is related to genotypes/disease status:

PART II: GENETIC ASSOCIATION STUDIES WITH COMPLEX DESIGN

Genetic Association Studies (GAS) aim to identify genomic variants (e.g., SNPs, haplotypes) that are associated with disease outcomes.
A motivating example—U.S. Kidney Cancer Case-Control Study

- Population-Based Case-Control Study, Detroit, Michigan and Chicago, Illinois
- Cases: identified from the population-based cancer registry in Detroit
- Selection of controls:
  - Stratified Simple Random Sample design
  - Strata defined by the sex, age and black density
- 1,018 cases and 1,038 controls
- Buccal and blood samples were collected as a source of genomic DNA.
- Tobacco use is one of the risk factors of kidney cancer (Brennan et al., 2008)

Analytical Goal 1: Investigate the interaction effect between tobacco use and the SNPs in the APOE promoter region (Moore, et al. 2009) on the risk of kidney cancer

Analytical Goal 2: Investigate the main effect of the haplotypes inferred from 4 SNPs (Karami et al. 2009) on the risk of kidney cancer.
In GAS, SNP and haplotypes – two common forms of genetic variants

**SNP (single-nucleotide-polymorphism)** is the occurrence of two or more alleles at one locus in a DNA sequence among individuals in the same population.

The bases G and A are referred to as **alleles**, alternative forms of a DNA segment at a single **locus**.
Goal 1: Gene-Environment (G-E) Interaction effect on risk of disease

• Standard Logistic Regression Approaches – G-E interaction term included in the regression model (STATA, SUDAAN, R-SURVEY)

However, Poor power due to small numbers of observations in cells cross-classified genetic variants and exposures.

• **Retrospective** methods can be more efficient – exploring various covariate-distributional assumptions (Chatterjee et al. 2005).

Therefore,

## Analyses results from KCS analysis

<table>
<thead>
<tr>
<th></th>
<th>Weighted Logis. Reg.</th>
<th>Pseudo-SPMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>rs8106922</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>Smoking status×rs8106922</td>
<td>-0.06</td>
<td>-0.19</td>
</tr>
<tr>
<td><strong>Standard Errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>rs8106922</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>Smoking status×rs8106922</td>
<td>0.16</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>p-values</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.56</td>
<td>0.06</td>
</tr>
<tr>
<td>rs8106922</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>Smoking status×rs8106922</td>
<td>0.73</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Goal 2: Haplotype effect on the risk of disease

**Haplotype** is a set of closely linked SNPs (combination of SNPs) on the same chromosome within the genomic region of interest.

**Diplotype** is haplotype pairs on homologous chromosomes.

**Genotype** is a combination of the haplotypes/SNPs on homologous chromosomes.

**Phenotype** is the traits or conditions that you can observe or diagnose, like eye color or breast cancer.

For a simple example,
Individual 1

haplotype 1
SNP A   SNP B

haplotype 2
SNP a   SNP b

haplotype 3
SNP A   SNP b

haplotype 4
SNP a   SNP B

diplotype
AB/ab

genotype
AaBb

phenotype
kidney cancer

Individual 2

haplotype 3
SNP A   SNP b

haplotype 4
SNP a   SNP B

diplotype
Ab/aB

genotype
AaBb

phenotype
kidney cancer
Analyzing haplotype data

Advantages

• There is strong evidence that several variants can interact together to have a large effect on the observed phenotype [Schaid, 2004].
• Haplotypes reduce the dimension of association tests and may gain statistical power [Clark, 2004]

Challenges

• Number of haplotypes can be large, and the number is often an unknown priori [Excoffier and Slatkin, 1995].
• **Phase Ambiguity**

  Can genotype data infer which SNPs form the Haplotype?  

  **NO!**

  Phase ambiguity – **MISSING DATA PROBLEM**
Two-step method

Step 1: Estimation of Haplotype Frequencies $\theta$ – assuming HWE

Challenge: Can be heavy computation if $\theta$ is high dimensional!

Weighted EM algorithm

✓ At E-step, the expected number of each haplotype in the population conditional on the genotypes by HWE and
✓ At M-step, the weighted estimates of haplotype frequencies,
✓ Implemented iteratively until convergence is reached.

The estimate denoted by $\hat{\theta}_{WEM}$
Step 2: Estimation of Regression Coefficients – Treating $\hat{\theta}_{WEM}$ as fixed

The regression parameters $\beta$ can be obtained by maximizing

$$L^w_\beta (y, G, E) = \sum_{i=1}^{n} w_i \sum_{j=1}^{c_i} \{ \log Pr_\beta (y_i | E_i, D^j_i) Pr_{\hat{\theta}_{WEM,\beta}} (D^j_i | obs) \}$$

conditional on the observed data $obs=(y, G, E)$,

$$Pr_{\hat{\theta}_{WEM,\beta}} (D^j_i | obs) = \frac{Pr_\beta (y_i | E_i, D^j_i) Pr_{\hat{\theta}_{WEM}} (D^j_i)}{\sum_{j'=1}^{c_i} Pr_\beta (y_i | E_i, D^{j'}_i) Pr_{\hat{\theta}_{WEM}} (D^{j'}_i)}.$$
where

\( y_i \): Binary indicator of presence, \( y=1 \), or absence, \( y=0 \), of a disease

\( E_i \): Environmental covariates associated with the \( i^{th} \) person

\( G_i \): Genotype of the \( i^{th} \) person

\( obs=(y, G, E) \)

\( D_i^j \): The \( j^{th} \) diplotype that is compatible with genotype \( G_i \)

\( c_i \): the total number of diplotypes that is compatible with \( G_i \)

\( Pr_\theta(D) \): the prior probability of diplotype \( D \)

\( Pr_\beta(y|E,D) \): the risk of disease given the exposure \( (E) \) and \( D \)
\[ L^w_\beta (y, G, E) = \sum_{i=1}^{n} w_i \sum_{j=1}^{c_i} \log \text{Pr}_\beta (y_i | E_i, D_i^j) \text{Pr}_{\theta_{\text{WEM}, \beta}} (D_i^j | obs) \] 

\( w_i \): Sampling weights

- Cross-sectional studies – Population Weights (PW)
- Case-control studies with rare disease
  \( \hat{\beta}_{\text{WEM}} \) - Inefficient due to the large variation of the PWs
  \( \rightarrow \) Rescale the PW of controls [Scott and Wild, 2011]

\( \hat{\beta}_{\text{WEM}} \) for all the coefficients apart from intercept is design consistent
• **One-step method**

~ Estimate haplotype frequencies $\theta$ and regression parameters jointly $\beta$

  o Construct the pseudo log-likelihood

$$L_w^y(y, G, E) = \sum_{i=1}^{n} w_i \sum_{j=1}^{c_i} \left\{ \log P_{r_\beta}(y_i | E_i, D_i^j) Pr_\gamma(D_i^j | obs) \right\},$$

  Unknown parameters $\gamma = (\beta, \theta)$

  o Solving $\gamma$ directly are tedious and even numerically infeasible

  o Instead of maximizing $L^w$ directly – **Extended WEM (EWEM)**
• **E-step:** Compute the probability of diplotypes given observed data (genotypes, covariates, and outcomes)

\[
Pr(D_i^j | obs) = \frac{Pr_{\theta}(y_i | E_i, D_i^j)Pr_{\theta}(D_i^j)}{\sum_{j_i=1}^{c_i} Pr_{\theta}(y_i | E_i, D_i^{j_i})Pr_{\theta}(D_i^{j_i})}.
\]

• **M-step:** maximize the conditional expectation of log-likelihood based on the complete data (i.e. diploptides, covariates, and outcomes)

\[
L_w^w(y, G, E) = \sum_{i=1}^{n} w_i \log \left\{ \sum_{j=1}^{c_i} \left\{ Pr_{\theta}(y_i | E_i, D_i^j)Pr(D_i^j | obs) \right\} \right\}
\]

• The iteration is continued until convergence criterion is satisfied. The resulting estimates are denoted by \(\hat{\theta}_{EWEM}\) and \(\hat{\beta}_{EWEM}\).
Variance estimation of the pseudo log-likelihood estimators

The pseudo log-likelihood estimators for haplotype frequencies $\theta$ and $\beta$ are nonlinear functions of the complex sample data.

By Taylor linearization method,

- Variance of one-step estimators $\hat{\beta}_{EWEM}$, automatically accounting for the variance due to estimating the haplotype frequencies $\theta$.
- Variance of two-step estimators $\hat{\beta}_{WEM}$, however, ignoring the variance due to estimating the haplotype frequencies $\theta$. 
Simulation Studies

– Case-Control Design
– Cross-Sectional Design
Summary of simulation results

✔ Under cross-sectional design, the proposed one-step and two-step methods for estimating haplotype frequencies, $\hat{\theta}_{WEM}$ and $\hat{\theta}_{EWEM}$, and regression coefficients, $\hat{\beta}_{WEM}$ and $\hat{\beta}_{EWEM}$, perform equally well. Note the estimated variances of the one-step estimator $\hat{\beta}_{EWEM}$ automatically account for the uncertainty of $\hat{\theta}_{EWEM}$, and therefore are recommended.

✔ Under case-control design with rare diseases, the two-step estimator $\hat{\theta}_{WEM}$ with population weights (PW) and $\hat{\beta}_{WEM}$ with scaled PW are recommended.
<table>
<thead>
<tr>
<th></th>
<th>Two-Step Estimates</th>
<th>Std Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haplotype 1010</td>
<td>-0.733</td>
<td>-0.427</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>-0.128</td>
<td>-0.057</td>
</tr>
<tr>
<td>Smoking Status by 1010</td>
<td>0.075</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haplotype 1010</td>
<td>0.365</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>0.227</td>
</tr>
<tr>
<td>Smoking Status by 1010</td>
<td>0.207</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haplotype 1010</td>
<td>0.045</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>0.573</td>
</tr>
<tr>
<td>Smoking Status by 1010</td>
<td>0.717</td>
</tr>
</tbody>
</table>
Future Work

- Hardy-Weinberg Equilibrium tests
  - $TS_p$ test requires $\geq 2$ members selected within families; $TS_1$ test requires within-family selection perpendicular to genotypes
  - Future work: New HWE test – combining $TS_p$ and $TS_1$

- Genetic Association Studies (GAS)
  - Haplotype-based inference under retrospective framework
  - Genome Wide Association Studies
  - Sequencing Data

- Surveys help improve genetic studies

  Complex sampling designs offer unique advantages in GAS
  - Cost- and time-effective;
  - Obtain representative samples;
  - Avoid biased selection of controls and/or cases
Thank you!