Presenter’s Name: Colin O. Wu

National Heart, Lung, and Blood Institute

National Institutes of Health

Department of Health and Human Services

# A Statistical Method for Adjusting Covariates in Linkage Analysis With Sib Pairs

Colin O. Wu, Gang Zheng, JingPing Lin,
Eric Leifer and Dean Follmann

Office of Biostatistics Research, DECA, NHLBI

# 1. Framingham Heart Study (GAW13)

1.1 First Generation:

- 5209 subjects (2336 men & 2873 women);
- 29 to 62 years old when recruited;
- 1644 spouse pairs;
- Continuously examined every 2 years since 1948
  - medical history
  - physical exams
  - laboratory tests.

## 1.2 Second Generation (full dataset):

- 5124 of the original participants’ adult children & spouses of these adult children;
- 2616 subjects are offspring of original spouse pairs;
- 34 are stepchildren;
- 898 offspring are children with only one parent in the study;
- 1576 are spouses of the offspring;
- Offspring cohort followed every 4 years;
- Interval between Exams 1&2 is 8 years.

## 1.3 Second Generation (sib-pair subset)

- 482 multi-sib families
  – from 330 pedigrees;
- Observed trait:
  systolic blood pressure;
- Covariates:
  1. age (in years),
  2. gender (0=female, 1=male),
  3. drinking
  (average daily alcohol consumption in ml).
- Genotype data:
  398 random markers with an average of 10cM apart.

# 2. Methods for Linkage Analysis

2.1 Methods based on identity by descent (IBD)

- Association in pedigrees between phenotype and IBD sharing at loci linked to trait loci;
- Linkage for qualitative traits
  – IBD sharing conditional on phenotypes:
  e.g. affected sib-pair methods
  (Hauser & Boehnke, 1998).
- Linkage for quantitative trait loci (QTL)
  – phenotypes conditional on IBD sharing,
  e.g. Haseman & Elston (1972), Amos (1994);
  – extremely discordant sib-pairs,
  e.g. Risch & Zhang (1995, 1996).

## 2.2 The Haseman-Elston method

$X_{1j}, X_{2j}$ : observed traits for 1st and 2nd sibs;

$Y_j = \left( X_{1j} - X_{2j} \right)^2$

– squared trait difference in jth pair;

$X_{ij} = \mu + g_{ij} + e_{ij}$,

$\mu$ : the overall mean trait value,

$g_{ij}$ : the genetic effect on the $(i, j)$-th sib,

$e_{ij}$ : the environmental effect on the $(i, j)$-th sib.

# HE Model without Covariate Adjustment:

Assumptions: (i) one locus determines $g_{ij}$, (ii) two alleles, $B$ and $b$, (iii) gene frequencies $p$ and $q$.

Genotypic values:

$$g_{ij} = \begin{cases} a & \text{for a } BB \text{ individual;} \\ d & \text{for a } Bb \text{ individual;} \\ -a & \text{for a } bb \text{ individual.} \end{cases}$$

$$E\left(Y_j \mid \pi_j\right) = \alpha + \beta\pi_j,$$

$\pi_j =$ proportion of genes IBD for the $j$-th pair,

$\alpha,\ \beta:$ unknown parameters.

Linkage:

Negative $\beta \Rightarrow$ potential linkage between QTL & marker locus.

Hypothesis Testing Problem:

$H_0$ : $\beta = 0$ (no linkage)

$H_1$ : $\beta < 0$ (linkage)

Limitations:

- Covariate effects are not included.
- Genetic and environmental effects are additive.
- Method may not have sufficient power.

## 2.3 HE Method with Linear Covariate Adjustment (SAGE SIBPAL)

- Involve families with more than 2 sibs.
- Can use other measures of trait difference
  - e.g. the mean-corrected cross-product.
- Include covariate effects in linear regression
  - e.g. Elston, Buxbaum, Jacobs and Olson (2000) “Haseman and Elston Revisted”.

## Linear generalization:

$Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}$: covariates for $(i, j)$-th sib,

$Z_{ij} = \left(\pi_j, Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right)^T$ : covariate vector,

$Z_j^{(l)}$: covariate for the $j$-th sib pair,

e.g. $Z_j^{(l)} = \left(Z_{1j}^{(l)} - Z_{2j}^{(l)}\right)$ or $\left|Z_{1j}^{(l)} - Z_{2j}^{(l)}\right|$,

$Z_j = \left(Z_j^{(0)},\ldots,Z_j^{(p)}\right)^T$, $Z_j^{(0)} \equiv \pi_j$.

$$E\left(Y_j \mid \pi_j, Z_j^{(1)},\ldots,Z_j^{(p)}\right) = \alpha + \sum_{l=0}^{p}\left(\beta_l Z_j^{(l)}\right).$$

Linkage: $\beta_0 < 0 \Rightarrow$ linkage.

Covariate effects:

$\beta_l \neq 0,\ l = 1, \ldots, p \Rightarrow$ effect of the $l$-th covariate.

Limitation:

Only the information in $Z_j$ is used

$\Leftrightarrow \left( Z_{1j}^{(l)}, Z_{2j}^{(l)} \right)$ is reduced to $Z_j^{(l)}$.

For example, use $Z_j = \left| \text{age}_{1j} - \text{age}_{2j} \right|^2$.

# 3. The Proposed Method

## 3.1. Modeling the covariates

Goal: To generalize the HE regression model that includes the covariates $\left(Z_{1j}^{(l)}, Z_{2j}^{(l)}\right)$.

Assumptions:

(i) The covariates are not affected by the gene and the environment.

(ii) The effects of gene and environment are additive.

- Cross-sectional data:

$$X_{ij} = \mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right) + g_{ij} + e_{ij},$$

$$\mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right) = \text{ mean of } X_{ij} \text{ given } \left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right).$$

Equivalent form:

$$\begin{aligned} X_{ij}^{*} &= X_{ij} - \mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right) \\ &= \text{ covariate adjusted trait} \\ &= g_{ij} + e_{ij}. \end{aligned}$$

## Regression models for covariates:

Linear model:

$$\mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right)=\theta_0+\sum_{l=1}^{p}\left(\theta_l Z_{ij}^{(l)}\right).$$

Equivalently,

$$X_{ij}^{*}=X_{ij}-\left\{\theta_0+\sum_{l=1}^{p}\left(\theta_l Z_{ij}^{(l)}\right)\right\}$$

$$=g_{ij}+e_{ij};$$

$\theta=\left(\theta_0,\ldots,\theta_p\right)^T$ : linear coefficients.

General parametric models

– e.g. nonlinear models :

$$\mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right)=\mu\left\{\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right);\theta\right\}.$$

Nonparametric models $\left(\text{Härdle, 1991}\right)$:

$$\mu\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right)=\text{ smooth function of } Z_{ij}^{(l)}.$$

Semiparametric models (Bickel et al., 2000).

- Longitudinal data:
  (Repeated measurements over time)

For $j$-th sib pair:

$n_{1j} = n_{2j} = n_j =$ number of repeated measurements,

$T_{ijk} =$ time of the $k$-th measurement,

$$k = 1, \ldots, n_{ij}.$$

$$X^*_{ijk} = X_{ijk} - \mu\left(T_{ijk}, Z^{(1)}_{ijk}, \ldots, Z^{(p)}_{ijk}\right)$$

$$= g_{ij} + e_{ij}.$$

Notation: $X_{ijk}$ = observed trait,

$Z_{ijk}^{(1)},\ldots,Z_{ijk}^{(p)}$ = covariates at time $T_{ijk}$,

$\mu\left(T_{ijk}, Z_{ijk}^{(1)},\ldots,Z_{ijk}^{(p)}\right)$ = conditional mean of $X_{ijk}$,

$X_{ijk}^{*}$ = covariate adjusted trait.

Linear model (Verbeke & Molenberghs, 2000):

$$X_{ijk}^{*} = X_{ijk} - \left\{\theta_0 + \theta_{00} T_{ijk} + \sum_{l=1}^{p}\left(\theta_l Z_{ijk}^{(l)}\right)\right\},$$

$\theta = \left(\theta_0, \theta_{00}, \theta_1, \ldots, \theta_p\right)$ : linear coefficients.

## 3.2 Covariate adjusted linkage detection

- General Procedure:

✓ Select a regression model for the covariates.

✓ Estimate the covariate adjusted trait based on the above regression model.

✓ Apply the linkage procedures, such as the HE model or the variance-components model, using the estimated adjusted trait values and genotypic values.

3.3 Cross-sectional data

Covariate adjusted HE model:

$Y_j^* = \left( X_{1j}^* - X_{2j}^* \right)^2$ : adjusted squared trait difference.

Same derivation in HE (1972) $\Rightarrow$

$$E\left( Y_j^* \mid \pi_j \right) = \alpha + \beta \pi_j .$$

$\alpha$, $\beta$ : unknown parameters.

Testing problem:

$$H_0 : \beta = 0 \text{ (no linkage)},$$

$$H_1 : \beta < 0 \text{ (linkage)}.$$

## Estimation of adjusted trait values:

Data from sib pairs are correlated

$\Rightarrow$ Existing estimation methods for independent data can not be directly applied.

Two approaches:

1) Use methods for correlated data, such as GEE – treat each family as a subject, each member as a single observation.
2) Resample independent observations:

i. Randomly sample one member from each family.
ii. Estimate the parameters and adjusted trait values using the re-sampled data and procedures for independent data, such as LSE, MLE, etc.
iii. Repeat the previous steps many times and compute the estimates using the average of the estimates from the re-sampled data.

This leads to consistent estimates when the sample size (number of families) is large (Hoffman et al., 2001).

## Procedure for linear adjustment model:

Step 1: Estimate $\theta$ by $\hat{\theta}$, a consistent estimator.

Step 2: Estimate $X_{ij}^{*}$ and $Y_{j}^{*}$ by

$$\hat{X}_{ij}^{*} = X_{ij} - \mu\left\{\left(Z_{ij}^{(1)},\ldots,Z_{ij}^{(p)}\right);\hat{\theta}\right\},$$

$$\hat{Y}_{j}^{*} = \left(\hat{X}_{1j}^{*} - \hat{X}_{2j}^{*}\right)^{2}.$$

Step 3: Fit the HE model using $\hat{Y}_{j}^{*}$ and test

$\beta$=0 vs. $\beta$<0.

# 3.5 Longitudinal data

$$Y_{jk}^{*} = \left( X_{1jk}^{*} - X_{2jk}^{*} \right)^{2}$$

– adjusted squared difference in $j$-th pair at $k$-th measurement;

$$\bar{Y}_{j}^{*} = \sum_{k=1}^{n_j} \left( Y_{jk}^{*} / n_j \right):$$ mean adjusted difference;

$$E\left( \bar{Y}_{j}^{*} \mid \pi_j \right) = \alpha + \beta \pi_j .$$

Linkage detection:

$\beta = 0$ (no linkage); $\beta < 0$ (linkage).

Two sources of potential correlations in the estimation of adjusted trait values:

i. Correlation within a sib
⍰ intra-subject correlation.

ii. Correlation between sibs within a family
⍰⍰ intra-family correlation.

⍰ Nested longitudinal data.

⍰ Methods for longitudinal estimation can not be directly applied
(Morris, Vannucci, Brown and Carroll, 2003, JASA).

Resampling approach:

i. Randomly select one sib from each family
⍰ Resampled data contain repeated measurements of independent sibs.

ii. Estimate the covariate adjusted trait values from the above resampled data based on longitudinal estimation methods (GEE, MLE, REMLE, etc.).

iii. Repeat the above steps many times and estimate the parameters using the averages of the estimates from the resampled data.

iv. Fit the HE model using existing procedure.

# 4. Framingham Heart Study

Features of the data:

- Clustered data from families;
- Repeated measurements;
- Multi-sib families;
- Continuous and categorical covariates.

Variables:

- Quantitative trait: SBP;
- Covariates: age, gender (0=female, 1=male), drinking (average daily consumption).

## SAGE HE Model:

Use $\bar{Y}_j = \sum_{k=1}^{n_j}\left(X_{1jk} - X_{2jk}\right)^2 / n_j$ in place of $Y_j$;

$Z_j^{(1)} = \left|\ \text{age}_{1j} - \text{age}_{2j}\right|^2$ : age difference between sibs;

$Z_j^{(2)} = \left|\ \text{gender}_{1j} - \text{gender}_{2j}\right|$ : 0 if same sex, 1 if different sex;

$Z_j^{(3)} = \left|\ \text{drinking}_{1j} - \text{drinking}_{2j}\right|^2$;

Fit regression:

$$E\left(\bar{Y}_j \mid \pi_j, Z_j\right) = \alpha + \beta\pi_j + \beta_1 Z_j^{(1)} + \beta_2 Z_j^{(2)} + \beta_3 Z_j^{(3)}.$$

No linkage: $\beta = 0$; Linkage: $\beta < 0$.

New HE Model:

Use linear model with 1000 resampling replications;

$$\mu\left\{\left(\text{age, gender, drinking}\right);\theta\right\} = \theta_0 + \theta_1 \times \text{age}$$
$$+\theta_2 \times \text{gender} + \theta_3 \times \text{drinking};$$

$$X_{ijk}^{*} = X_{ijk} - \mu\left\{\left(\text{age, gender, drinking}\right);\hat{\theta}\right\};$$

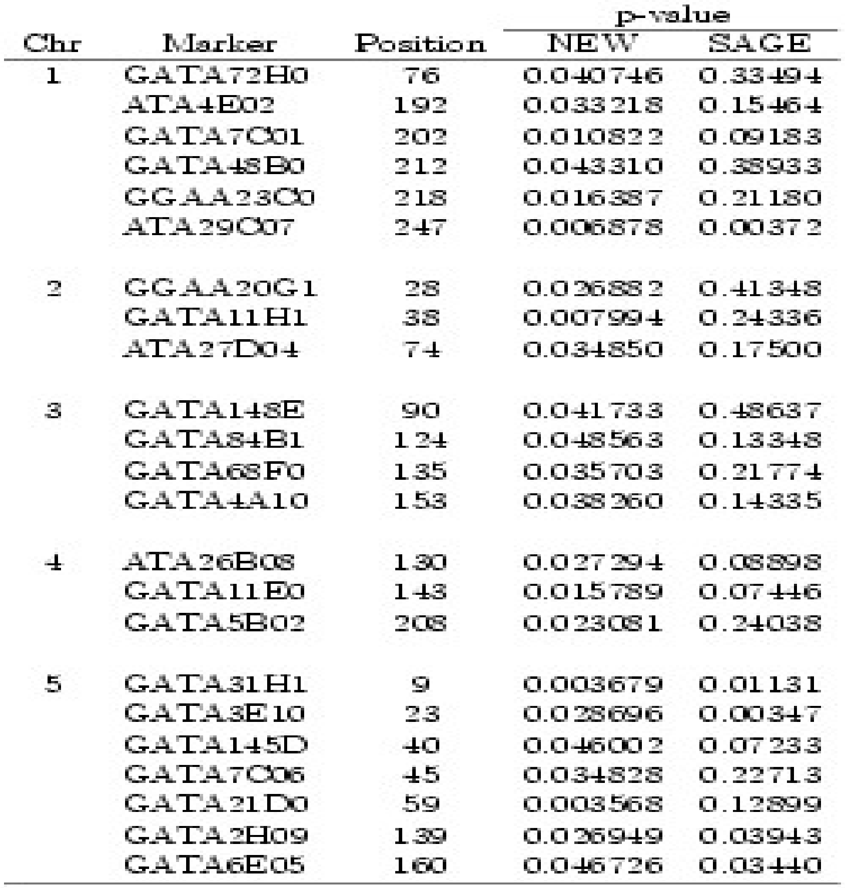
$$\bar{Y}_j^{*} = \sum_{k=1}^{n_j}\left(X_{1jk}^{*} - X_{2jk}^{*}\right)^2 / n_j : \text{ average squared trait difference.}$$

Fit regression:

$$E\left(\bar{Y}_j^{*} \mid \pi_j, Z_j\right) = \alpha + \beta\pi_j.$$

No linkage: $\beta$=0; Linkage: $\beta$<0.

Table 1. Comparison of SAGE HE and NEW HE methods in two-point analysis



| | | | p-value | |
|---|---|---|---|---|
| Chr | Marker | Position | NEW | SAGE |
| 1 | GATA72H0 | 76 | 0.040746 | 0.33494 |
| | ATA4E02 | 192 | 0.033218 | 0.15464 |
| | GATA7C01 | 202 | 0.010822 | 0.09183 |
| | GATA48B0 | 212 | 0.043310 | 0.38933 |
| | GGAA23C0 | 218 | 0.016387 | 0.21180 |
| | ATA29C07 | 247 | 0.006878 | 0.00372 |
| 2 | GGAA20G1 | 28 | 0.026882 | 0.41348 |
| | GATA11H1 | 38 | 0.007994 | 0.24336 |
| | ATA27D04 | 74 | 0.034850 | 0.17500 |
| 3 | GATA148E | 90 | 0.041733 | 0.48637 |
| | GATA84B1 | 124 | 0.048563 | 0.13348 |
| | GATA68F0 | 135 | 0.035703 | 0.21774 |
| | GATA4A10 | 153 | 0.038260 | 0.14335 |
| 4 | ATA26B08 | 130 | 0.027294 | 0.08898 |
| | GATA11E0 | 143 | 0.015789 | 0.07446 |
| | GATA5B02 | 208 | 0.023081 | 0.24038 |
| 5 | GATA31H1 | 9 | 0.003679 | 0.01131 |
| | GATA3E10 | 23 | 0.028696 | 0.00347 |
| | GATA145D | 40 | 0.046002 | 0.07233 |
| | GATA7C06 | 45 | 0.034828 | 0.22713 |
| | GATA21D0 | 59 | 0.003568 | 0.12899 |
| | GATA2H09 | 139 | 0.026949 | 0.03943 |
| | GATA6E05 | 160 | 0.046726 | 0.03440 |

Table 2. Comparison of SAGE HE and NEW HE methods in two-point analysis

| | | | p-value | |
|---|---|---|---|---|
| Chr | Marker | Position | NEW | SAGE |
| 6 | GATA11E0 | 73 | 0.052419 | 0.03843 |
| | GATA31 | 119 | 0.046214 | 0.04866 |
| | GATA32B0 | 138 | 0.022104 | 0.10438 |
| | GATA165G | 155 | 0.003483 | 0.01098 |
| | 242zg5 | 166 | 0.022416 | 0.07021 |
| | GATA81B0 | 173 | 0.005961 | 0.01083 |
| 7 | GATA13G1 | 50 | 0.032951 | 0.12551 |
| | GATA31A1 | 58 | 0.029264 | 0.18204 |
| | GATA24D1 | 70 | 0.001847 | 0.07439 |
| | GATA118G | 79 | 0.004014 | 0.17304 |
| | ATA22G07 | 187 | 0.007621 | 0.02528 |
| 8 | GATA25C1 | 22 | 0.025332 | 0.19608 |
| | GATA23D0 | 26 | 0.002706 | 0.04254 |
| | GATA72C1 | 37 | 0.004479 | 0.01447 |
| | GATA21C1 | 140 | 0.046820 | 0.26550 |
| 9 | GATA62F0 | 14 | 0.09318 | 0.02752 |
| | GATA21A0 | 22 | 0.12223 | 0.04280 |
| | GATA27A1 | 32 | 0.00321 | 0.00901 |
| | GATA7D12 | 66 | 0.03145 | 0.12940 |
| | GATA21F0 | 80 | 0.02700 | 0.06084 |
| | 183xh10 | 92 | 0.02049 | 0.07986 |
| | ATA59H06 | 147 | 0.00930 | 0.09181 |
| 10 | GATA70E1 | 46 | 0.003619 | 0.07436 |
| | GATA87G0 | 94 | 0.046017 | 0.22637 |
| | GGAA2F11 | 117 | 0.006402 | 0.00256 |
| | GATA64A0 | 125 | 0.000170 | 0.00224 |
| | 198zb4 | 171 | 0.039178 | 0.18715 |

Table 3. Comparison of SAGE HE and NEW HE methods in two-point analysis

|  |  |  | p-value |  |
|---|---|---|---|---|
| Chr | Marker | Position | NEW | SAGE |
| 12 | ATA27A06 | 49 | 0.003387 | 0.00083 |
|  | GATA91H0 | 56 | 0.009377 | 0.04915 |
|  | GATA5A09 | 57 | 0.020412 | 0.00958 |
|  | GGAT2G06 | 68 | 0.002970 | 0.00226 |
|  | GATA73H0 | 78 | 0.000139 | 0.00062 |
|  | GATA3F02 | 81 | 0.006071 | 0.06091 |
|  | GATA26D0 | 83 | 0.010513 | 0.19872 |
|  | GATA63D1 | 95 | 0.013884 | 0.01942 |
| 14 | GATA43H0 | 28 | 0.010066 | 0.12741 |
| 15 | GATA151F | 60 | 0.017276 | 0.20761 |
|  | GATA73F0 | 101 | 0.050374 | 0.02575 |
| 16 | GGAA3G05 | 58 | 0.036188 | 0.093515 |
|  | GATA27A0 | 122 | 0.009387 | 0.09375 |
| 17 | GTAT1A05 | 1 | 0.033590 | 0.29007 |
|  | GAAT2C03 | 11 | 0.016140 | 0.27796 |
|  | GATA28D1 | 100 | 0.006351 | 0.04783 |
|  | 044xg3 | 117 | 0.008744 | 0.09982 |
|  | 217yd10 | 126 | 0.002232 | 0.11083 |
| 18 | 321xc9 | 7 | 0.017048 | 0.021697 |
|  | GATA88A1 | 13 | 0.013188 | 0.027941 |
| 19 | GATA29B0 | 88 | 0.021196 | 0.039963 |
| 21 | GATA70B0 | 58 | 0.046708 | 0.35981 |

# 5. Discussion

- Advantages for covariate adjustment:

✓ small variation for the estimates;

✓ better interpretation for the model.

- Directions of further research:

✓ Non-additive models, e.g. covariate-gene and covariate-environment interactions;

✓ Covariate adjustment with other measures of the trait difference;

✓ Methods of model selection;

✓ Models with general pedigrees.