Semiparametric analysis of case–control genetic data in the presence of environmental factors

Yulia Marchenko

Senior Statistician
StataCorp LP

2008 London Stata Users Group Meeting

Outline

1. Haplotype-based disease association studies
   - Genetic markers
   - Data structure
2. Semiparametric logistic analysis of case-control genetic data
   - Retrospective sampling
   - Efficient prospective analysis of case-control data
   - Haplotype-based logistic model
   - Modes of inheritance
   - Characteristics of genetic data
   - Retrospective likelihood of case-control genetic data
3. Stata command haplologit
   - Description
   - Case–control study of colorectal adenoma
   - Analysis of the CASR data
   - Results
4. Future work

Haplotype-based disease association studies

The main goal of genetic disease association studies is to determine the genetic basis for complex diseases. Specifically, the aim is to identify genetic variants which either directly influence complex diseases or are in linkage disequilibrium with such causal variants.

Biallelic SNPs are often used as genetic markers in association studies thanks to availability of high-density SNP maps published by International SNP Map Working Group (2001) and International HapMap Consortium (2003).

Genetic markers - SNPs

Definition

Single nucleotide polymorphism (SNP, pronounced as “snip”) is a single nucleotide (A, T, C, or G) variation of the DNA sequence that occurs in at least 1% of the population.

For example, DNA fragments AAGC**C**TA and AAGC**T**TA from two subjects differ in a single nucleotide. In this example, the bases **C** and **T** are referred to as *alleles*, alternative forms of a DNA segment at a single locus.

A SNP is often coded as 1 if a rare allele is present at a SNP site and 0, otherwise.

Haplotypes and diplotypes

Definition

Haplotype is a sequence of closely linked SNPs on the same chromosome within the genomic region of interest. *Diplotype* is a set of two haplotypes humans carry in the pair of homologous chromosomes.

Using binary coding of SNPs, a haplotype can be represented as a binary sequence and a diplotype can be represented as a pair of binary sequences. Therefore, with M SNP sites (loci), there are 2^M possible haplotypes and 2^{2M} possible diplotypes.

Genotypes

In haplotype-based disease association studies, a subject's genetic information is described by a diplotype. In practice however we observe genotypes instead of diplotypes.

Definition

Genotype is a combination of the haplotypes from a pair of homologous chromosomes.

Mathematically speaking, if H_1 and H_2 are two haplotypes (binary sequences), a genotype $G = H_1 + H_2$ is the sum of these two binary sequences resulting in a sequence of the numbers 0, 1, and 2.

Data structure

Data from a case-control haplotype-based disease association study usually consist of

- a disease (or case-control) status $D = 0, 1$;
- genotype data from M tightly linked SNPs $G = (g_1, \ldots, g_M), g_k \in \{0, 1, 2\}$; and
- subjects' characteristics and environmental exposures $X = (x_1, \ldots, x_p)$.

We consider the case when the data are collected from samples of unrelated individuals using the retrospective sampling scheme.

Case-control (retrospective) sampling

- select people with $D = 1$ and sample from them to obtain values of covariates Z;
- select people with $D = 0$ and sample from them to obtain values of covariates Z;
- samples (covariate values) are obtained conditional on the disease status D.

Retrospective likelihood

Recall that under the case-control sampling design, the likelihood function

$$f_{Z|D}(z|d) = \frac{\Pr(D = d|Z = z) f_Z(z)}{\Pr(D = d)}$$

depends on the probability of disease $\Pr(D = d) = \pi_d$ in the population, the covariate distribution $f_Z()$ in the population, and the likelihood function under the prospective design $\Pr(D|Z)$.

Under the considered logistic model,

$$\Pr(D = 1|Z = z; \alpha_0, \boldsymbol{\beta}) = K(\alpha_0 + \boldsymbol{\beta}^\top z)$$

where $K(a) = \exp(a)\{1 + \exp(a)\}^{-1}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

Link between prospective and retrospective likelihood functions

Prentice and Pyke (1979) showed that one can obtain MLEs of $\boldsymbol{\beta}$ by fitting the standard logistic model (ignoring the retrospective design) to case-control data without any parametric assumptions about the covariate distribution $f_Z()$.

Roeder et al. (1996), besides extending the result to the case of covariates measured with error, provided the explicit relationship between the parameters of the prospective and retrospective likelihood functions:

$$\boldsymbol{\beta}^\star = \boldsymbol{\beta} \text{ and } \alpha_0^\star = \alpha_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$$

where n_1 and n_0 are the respective numbers of cases and controls.

Note

Intercept α_0 is not estimable from a prospective logistic regression with case-control data unless the probability of a disease in the population π_1 is known.

Semiparametric efficiency of the standard logistic model with case-control data

Breslow et al. (2000) showed that the standard logistic regression is semiparametric-efficient (the variance of coefficients attains the lower bound of the underlying semiparametric model) under the case-control sampling design.

Assumption

The semiparametric-efficiency of the prospective-type analysis of case-control data holds under the assumption of an arbitrary covariate distribution.

Risk haplotypes

With SNP genetic data, we want to study the effect of *risk haplotypes* (target haplotypes whose effect on a disease is of interest) on the disease and possibly their interaction with environmental exposures X.

Subjects' genetic information consists of diplotypes, haplotype pairs, $H^{\mathrm{d}} = (H_k, H_l)$ with constituent haplotypes H_k and H_l. As such, the effect of risk haplotypes may be modeled according to various genetic models depending on the number of copies of the risk haplotype present in a subject's diplotype.

Haplotype-based logistic model

Haplotype-effects logistic model with risk haplotypes $H^\star$ and environmental factors X is

$$\Pr(D = 1|X, H^{\mathrm{d}}; \alpha_0, \boldsymbol{\beta}_X, \boldsymbol{\gamma}_{H^\star}) = K\{\alpha_0 + \boldsymbol{\beta}_X X + m(H^{\mathrm{d}}, X; \boldsymbol{\gamma}_{H^\star})\}$$

where function $m(H^{\mathrm{d}}, X; \boldsymbol{\gamma}_{H_\star})$ is linear in risk haplotype parameters $\boldsymbol{\gamma}_{H^\star}$ with coefficients determined by a risk haplotype model (or mode of inheritance).

Modes of inheritance, main effects only

Consider a single risk haplotype $H_1^\star$ and its main effect on the disease. In general (under the codominant model),

$$\begin{aligned} m\{H^{\mathrm{d}} = (H_k, H_l); \gamma_{H_1^\star}\} &= \{I(H_k = H_1^\star) + I(H_l = H_1^\star)\}\beta^a_{H_1^\star} \\ &+ I(H_k = H_l = H_1^\star)\beta^r_{H_1^\star} \end{aligned}$$

where $I()$ denotes the indicator function and $\gamma_{H_1^\star} = (\beta^a_{H_1^\star}, \beta^r_{H_1^\star})^\top$.

- **codominant** – the effects of having two copies of a risk haplotype in a diplotype and a single copy can be different
- **additive** – having two copies of a risk haplotype in a diplotype doubles the effect compared to having only one copy; $\beta^r_{H_1^\star} = 0$
- **recessive** – only having exactly two copies of a risk haplotype has an effect on a disease; $\beta^a_{H_1^\star} = 0$
- **dominant** – having one or two copies of a risk haplotype has the same effect on a disease; $\beta^r_{H_1^\star} = -\beta^a_{H_1^\star}$

Modes of inheritance, interaction

We can add the interaction effect of $H_1^\star$ with an environmental factor X_1. Then, under the codominant model,

$$\begin{aligned} m(H^{\mathrm{d}}, X_1; \gamma_{H_1^\star}) &= \{I(H_k = H_1^\star) + I(H_l = H_1^\star)\}\beta^a_{H_1^\star} \\ &+ I(H_k = H_l = H_1^\star)\beta^r_{H_1^\star} \\ &+ \{I(H_k = H_1^\star) + I(H_l = H_1^\star)\}X_1\beta^a_{H_1^\star X_1} \\ &+ I(H_k = H_l = H_1^\star)X_1\beta^r_{H_1^\star X_1} \end{aligned}$$

where $\gamma_{H_1^\star} = (\beta^a_{H_1^\star}, \beta^r_{H_1^\star}, \beta^a_{H_1^\star X_1}, \beta^r_{H_1^\star X_1})^\top$.
Similarly,

- $\beta^r_{H_1^\star} = \beta^r_{H_1^\star X_1} = 0$ under the additive model;
- $\beta^a_{H_1^\star} = \beta^a_{H_1^\star X_1} = 0$ under the recessive model;
- $\beta^r_{H_1^\star} = -\beta^a_{H_1^\star}$ and $\beta^r_{H_1^\star X_1} = -\beta^a_{H_1^\star X_1}$ under the dominant model.

Analysis of case-control genetic data

A haplotype-effects logistic regression model is

$$\mathrm{Pr}(D = 1|X, H^{\mathrm{d}}; \alpha_0, \boldsymbol{\beta}_X, \boldsymbol{\gamma}_{H^\star}) = K(\alpha_0 + \boldsymbol{\beta}_X^\top X + \boldsymbol{\gamma}_{H^\star}^\top M)$$

where components of a column vector M (genetic covariates) depend on subjects' diplotypes, chosen risk haplotypes $H^\star$, the chosen mode of inheritance, and are evaluated using expressions described earlier for function $m(H^{\mathrm{d}}, X; \boldsymbol{\gamma}_{H^\star})$.

How do we estimate the parameters of the above model?
Let $Z = (X, M)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_X^\top, \boldsymbol{\gamma}_{H^\star}^\top)^\top$ and follow Prentice and Pyke's approach as before.

So, are we done?

Characteristics of genetic data

- We have additional information about the distribution of covariates – can still utilize the prospective approach but the results are no longer semiparametric-efficient.

- In practice, we usually observe genotypes instead of diplotypes – “phase ambiguity” arises (will discuss later).

- Genotype data are often missing – need to deal with the missing-data problem to avoid a possibly significant reduction in the sample size.

Additional information about covariate distribution

- gene-environment independence – $f(X, H^{\mathrm{d}}) = g(X)q(H^{\mathrm{d}}; \boldsymbol{\theta})$
- population in Hardy-Weinberg equilibrium (HWE) – assumption about the distribution of genetic covariates:

$$\begin{array}{rcll} q\{H^{\mathrm{d}} = (H_k, H_l); \boldsymbol{\theta}\} & = & \theta_k^2 & \text{if } H_k = H_l \\ & = & 2\theta_k\theta_l & \text{if } H_k \neq H_l \end{array}$$

  where θ_k denotes the frequency for haplotype H_k.

- population deviates from HWE according to a certain parametric model. For example,

$$\begin{array}{rcll} q\{H^{\mathrm{d}} = (H_k, H_l); \boldsymbol{\theta}\} & = & \theta_k^2 + \rho\theta_k(1-\theta_k) & \text{if } H_k = H_l \\ & = & (1-\rho)\theta_k\theta_l & \text{if } H_k \neq H_l \end{array}$$

  where ρ denotes the inbreeding coefficient.

What is phase ambiguity?

Recall, that we observe subjects' genotypes $G = H_k + H_l$ instead of diplotypes $H^{\mathrm{d}} = (H_k, H_l)$. This creates a problem of "phase ambiguity" for heterozygous subjects who carry different alleles at two or more loci.

Definition

Homozygous subjects carry two copies of the same allele at all SNP loci. *Heterozygous subjects* carry different alleles at at least one locus.

Example

- 2 SNPs – $H_1 = (0,0)$, $H_2 = (0,1)$, $H_3 = (1,0)$, and $H_4 = (1,1)$;
- For a $G = (1,1)$ there are 2 diplotypes $\{H_1, H_4\}$ and $\{H_2, H_3\}$ consistent with it, i.e., $G = H_1 + H_4 = H_2 + H_3$;
- for subjects with such genotype the phase is indeterminant.

Handling missing genotype data

Genotype data G:

- missing at random;
- missing components of G may be any value from $\{0, 1, 2\}$.

Thus, for subjects with missing genotype data we take into account multiple possible genotypes (and consequently multiple possible diplotypes) when computing the likelihood.

Note

We can view the "phase ambiguity" problem as a missing-data problem.

Retrospective likelihood of case-control genetic data

Taking into account the discussed characteristics, the retrospective likelihood function of case-control SNP-based data is

$$\Pr(X = x, H^{\mathrm{d}} \in H_G^{\mathrm{d}}) = \frac{g(x) \sum_{h^{\mathrm{d}} \in H_G^{\mathrm{d}}} \Pr(D = d | X = x, H^{\mathrm{d}} = h^{\mathrm{d}}) q(h^{\mathrm{d}}; \boldsymbol{\theta})}{\Pr(D = d)}$$

where $H_G^{\mathrm{d}} = \{(H_k, H_l)$: the haplotype pair is consistent with $G\}$ is the set of all possible diplotypes consistent with the observed genotype data G.

Semiparametric profile-likelihood method

Idea of the method: profile the possibly infinite-dimensional (if X has continuous components) nuisance distribution of X out of the retrospective likelihood first. Then maximize the resulting profile retrospective log-likelihood with respect to parameters of interest.

For details of the method and formulas, see Spinka et al. (2005), Lin et al. (2005), and Lin and Zeng (2006).

Rare disease

The retrospective sampling design is commonly used when conducting studies of rare diseases. Under the assumption of a rare disease,

$$\Pr(D = d \mid H^{\mathrm{d}}, X; \alpha_0, \boldsymbol{\beta}_X, \boldsymbol{\gamma}_{H^\star}) \approx \exp\{d(\alpha_0 + \boldsymbol{\beta}_X^\top X + \boldsymbol{\gamma}_{H^\star}^\top M)\}, \quad d = 0, 1$$

Lin and Zeng (2006) employed the rare-disease assumption for the development of their algorithm. Spinka et al. (2005) provided a method that does not make the assumption of a rare disease but noted that this assumption results in a simpler and more stable algorithm.

Short description

The Stata command `haplologit` estimates haplotype effects and haplotype-environment interactions from case-control genetic (SNP-based) data in the very important special case of

- a rare disease;
- a single candidate gene in HWE;
- independence of genetic and environmental factors.

`haplologit` accommodates three types of haplotype risk models: additive (the default), dominant, or recessive. It implements the retrospective profile-likelihood methods of Spinka et al. (2005) and Lin and Zeng (2006) which are equivalent under the assumptions of a rare disease and HWE.

For details about the syntax of `haplologit`, underlying algorithms, and examples see Marchenko et al. (2008).

Case–control study of colorectal adenoma

Biological hypothesis

Calcium prevents colorectal cancer, possibly through the calcium-sensing receptor.

Study goal.
To investigate the interactions of dietary calcium intake and genetic variants in the calcium-sensing receptor (*CaSR*) region (Peters et al. 2004; Lobach et al. 2007; Chen et al. 2008).

Data source.
Data come from Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial at the National Cancer Institute, USA and contain information on participants' dietary food intake and genotype data from three nonsynonymous SNPs in the *CaSR* region.

Description of the CASR data

SNP variables

Variables g_casr_01, g_casr_02, and g_casr_03 record genotype data from 3 nonsynonymous SNPs in exon 7 of the CaSR gene.

```
. describe g_casr_01 g_casr_02 g_casr_03 casecontrol Ldtcal sex agerand Caucasian

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
g_casr_01       byte   %8.0g                  first SNP locus
g_casr_02       byte   %8.0g                  second SNP locus
g_casr_03       byte   %8.0g                  third SNP locus
casecontrol     byte   %8.0g                  case-control status
Ldtcal          float  %9.0g                  log(1+dietary calcium from FFQ)
sex             byte   %8.0g                  gender: 1 = Male, 2 = Female
agerand         float  %9.0g                  age (in years)
Caucasian       float  %9.0g                  ethnicity: 0 = Non Caucasian, 1
                                                = Caucasian
```

Description of the CASR data, cont.

Total of 1312 subjects after eliminating subjects with missing calcium information – 644 cases and 668 controls.

```
. summarize g_casr_01 g_casr_02 g_casr_03 casecontrol Ldtcal sex agerand Caucasian

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   g_casr_01 |      1312    .1623476    .3968026          0          2
   g_casr_02 |      1312    .1021341    .3176886          0          2
   g_casr_03 |      1312    .2804878    .4992599          0          2
 casecontrol |      1312    .4908537     .500107          0          1
      Ldtcal |      1312    6.767107     .506731   4.893262   8.544137
-------------+--------------------------------------------------------
         sex |      1312    1.304878    .4605313          1          2
     agerand |      1312    62.53329    5.276247     55.042      74.99
   Caucasian |      1312    .9458841    .2263324          0          1
```

Analysis of the CASR data

Example

We want to investigate the interaction of dietary calcium intake (mg/day) and the three common haplotypes coded as “001”, “010”, and “100”. Other rare haplotypes are combined with the most common haplotype, “000”, to form the base (comparison) haplotype category.

Syntax

```
. haplologit casecontrol sex Ldtcal agerand Caucasian,
> snpvars(g_casr_01 g_casr_02 g_casr_03) inher(d)
> riskhap1("001") riskhap2("010", inter(Ldtcal))
> riskhap3("100", inter(Ldtcal))
> nolog happrefix("_")
```

Output – initial haplotype-frequency estimation

```
Building consistent haplotype pairs:

Obtaining initial haplotype frequency estimates from the control sample:

Haplotype frequency EM estimation

Number of iterations  =         53
Sample log-likelihood = -982.17816

  ┌───────────┬────────────┐
  │ haplotype │ frequency* │
  ├───────────┼────────────┤
  │       000 │     .71033 │
  │       001 │    .150449 │
  │       010 │    .055389 │
  │       100 │    .083832 │
  └───────────┴────────────┘

  * frequencies > .0015244
```

(*Continued on next page*)

Output – haplotype-effects estimation

```
Performing gradient-based optimization:

Haplotype-effects logistic regression
Mode of inheritance: dominant                   Number of obs      =      1312

Genetic distribution: Hardy-Weinberg equilib.   Number phased      =      1253
Genotype: g_casr_01 g_casr_02                   Number unphased    =        59
          g_casr_03                             Number missing     =         0

                                                Wald chi2(9)       =     36.61
Retrosp. profile log likelihood = -2769.5997    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
 casecontrol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.1222521     .12261    -1.00   0.319    -.3625632     .118059
      Ldtcal |  -.0553412   .1213515    -0.46   0.648    -.2931857    .1825032
     agerand |   .0370709   .0105986     3.50   0.000      .016298    .0578439
   Caucasian |   .1579015   .2517616     0.63   0.531    -.3355422    .6513452
        _001 |  -.2915038   .1238556    -2.35   0.019    -.5342563   -.0487513
        _010 |  -.4371039   .1932946    -2.26   0.024    -.8159544   -.0582533
        _100 |  -.2507072   .1535909    -1.63   0.103    -.5517398    .0503254
 _010*Ldtcal |  -.7947331   .2759949    -2.88   0.004    -1.335673    -.253793
 _100*Ldtcal |  -.5047162   .2205877    -2.29   0.022    -.9370601   -.0723723
       _cons |   .1193802   .2939819     0.41   0.685    -.4568137    .6955741
------------------------------------------------------------------------------

Note: _cons = b0 + ln(N1/N0) - ln{Pr(D=1)/Pr(D=0)}
```

(Continued on next page)

Output – haplotype-frequencies estimation

```
Haplotype frequencies

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _000 |   .7029555   .0121212    57.99   0.000     .6791983    .7267127
        _001 |   .1527102   .0095211    16.04   0.000     .1340491    .1713713
        _010 |   .0578413   .0063148     9.16   0.000     .0454646     .070218
        _100 |    .086493   .0075409    11.47   0.000     .0717131    .1012729
------------------------------------------------------------------------------
```

Results

- **haplotype main effects**
  significant dominant effects of "001" and "010" haplotypes
- **environmental factors**
  significant findings – age is associated with an increased risk of the disease;
  nonsignificant findings – an increased risk of advanced colorectal adenoma for Caucasians, males, and subjects with lower calcium intake
- **haplotype-environment interactions**
  Both interaction terms, `hap_010*Ldtcal` and `hap_100*Ldtcal`, are statistically significant at the 1% and 5% levels, respectively. This agrees with results obtained in Lobach et al. (2007).

Things to consider for the future

- allow modeling of departures from HWE according to a number of specific models as described in Lin and Zeng (2006)
- weaken the gene-environment independence assumption (e.g. Cheng et al. 2008)
- complete the list of genetic models by adding a codominant model
- allow multiple candidate genes and gene-gene interactions
- allow population stratification
- handle large number of markers – important in genome-wide association studies

Acknowledgements

Grant.
This work was supported by the NIH Phase I SBIR grant "Statistical Software for Genetic Association Studies" to StataCorp LP.

Consultants.
Raymond J. Carroll is a distinguished professor of statistics, nutrition, and toxicology at Texas A&M University.

Danyu Lin is a Dennis Gillings distinguished professor of biostatistics at the University of North Carolina.

Christopher I. Amos is a professor of epidemiology at the M. D. Anderson Cancer Research Center.

References

Breslow, N. E., J. M. Robins, and J. A.Wellner. 2000. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* 6: 447–455.

Chen, Y.-H., R. J. Carroll, and N. Chatterjee. 2008. Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*, to appear.

International Hapmap Consortium. 2003. The international HapMap project. *Nature* 426: 789–796.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 14.2 million single nucleotide polymorphisms. *Nature* 409: 928–933.

Lin, D. Y. and D. Zeng. 2006. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* 101: 89–118.

References, cont

Lin, D. Y., D. Zeng, and R. Millikan. 2005. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* 29: 299–312.

Lobach, I., R. J. Carroll, C. Spinka, M. H. Gail, and N. Chatterjee. 2007. Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, to appear.

Marchenko, Y. V., R. J. Carroll, D. Y. Lin, C. I. Amos, and R. G. Gutierrez. 2008. Semiparametric analysis of case-control genetic data in the presence of environmental factors. *The Stata Journal* 8(3), to appear.

Peters, U., N. Chatterjee, M. Yeager, S. J. Chanock, R. E. Schoen, K. A. McGlynn, T. R. Church, J. L. Weissfeld, A. Schatzkin, and R. B. Hayes. 2004. Association of genetic variants in the calcium-sensing receptor with risk of colorectal adenoma. *Cancer Epidemiology Biomarkers & Prevention* 13(12): 2181–2186.

References, cont

Prentice, R. L. and R. Pyke. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66: 403–411.

Roeder K., R. J. Carroll, and B. G. Lindsay. 1996. A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* 91:722–732.

Spinka, C., R. J. Carroll, and N. Chatterjee. 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* 29: 108–127.