The Stata Journal (2008)
8, Number 3, pp. 305–333

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

Yulia V. Marchenko
StataCorp
College Station, TX
ymarchenko@stata.com

Raymond J. Carroll
Department of Statistics
Texas A&M University
College Station, TX
carroll@stat.tamu.edu

Danyu Y. Lin
Department of Biostatistics
University of North Carolina
Chapel Hill, NC
lin@bios.unc.edu

Christopher I. Amos
M. D. Anderson Cancer Research Center
Houston, TX
camos@mdanderson.org

Roberto G. Gutierrez
StataCorp
College Station, TX
rgutierrez@stata.com

Abstract. In the past decade, many statistical methods have been proposed for the analysis of case–control genetic data with an emphasis on haplotype-based disease association studies. Most of the methodology has concentrated on the estimation of genetic (haplotype) main effects. Most methods accounted for environmental and gene–environment interaction effects by using prospective-type analyses that may lead to biased estimates when used with case–control data. Several recent publications addressed the issue of retrospective sampling in the analysis of case–control genetic data in the presence of environmental factors by developing efficient semiparametric statistical methods. This article describes the new Stata command `haplologit`, which implements efficient profile-likelihood semiparametric methods for fitting gene–environment models in the very important special cases of a rare disease, a single candidate gene in Hardy–Weinberg equilibrium, and independence of genetic and environmental factors.

Keywords: st0148, haplologit, haplotype-based analysis, haplotype–environment independence, case–control data, Hardy–Weinberg equilibrium, profile likelihood, retrospective study, single nucleotide polymorphisms (SNPs)

1 Introduction

Case–control data arise from the retrospective sampling design commonly used when conducting studies of rare diseases. In the retrospective (case–control) design, the sampling is conditional on the disease: cases and controls are drawn independently from two populations with and without a disease, respectively, and their observed covariate information is recorded. As such, the likelihood of the case–control data is based on the $\Pr(X \mid Y)$ of observing covariate information X given the disease status Y as opposed to the likelihood (with prospective data) that is based on the $\Pr(Y \mid X)$ of the disease given the covariates. The use of standard (prospective-type) logistic regression to analyze case–control data (ignoring the retrospective sampling scheme) is justified for the case when the covariate distribution is left unspecified (Andersen 1970; Prentice and Pyke 1979; Breslow, Robins, and Wellner 2000) and the covariates are always observed.

In genetic studies, however, it is usually reasonable to impose certain distributional assumptions about genetic covariates. The most common assumption is Hardy–Weinberg equilibrium (HWE), under which the genotype frequencies in a population remain constant from generation to generation. Deviations from HWE can also be modeled according to certain parametric forms (Weir 1996; Niu et al. 2002). Also, in the presence of environmental factors, it is often assumed that genetic factors are distributed independently of the environmental factors. Under these assumptions, standard logistic regression is no longer semiparametric efficient for the analysis of case–control data because additional information is available about the covariate distribution. The use of prospective-type logistic regression is further complicated by the presence of missing genetic markers, a problem often encountered during genotyping of subjects. When genetic data are missing, the conditional distribution of the genetic covariates given other experimental factors must be modeled.

Increased interest in single nucleotide polymorphisms (SNPs) as genetic markers prompted the development of new statistical methods for the analysis of associations between the disease phenotype and multiple SNPs, the so-called haplotype-based association studies. A *haplotype* is a sequence of SNPs on the same chromosome within the genomic region of interest. A subject's genetic information is described by a *diplotype*, the set of two haplotypes humans carry in the pair of homologous chromosomes. The current genotyping techniques do not provide an easy way of obtaining the diplotype data for which the gametic (parental) phase information is known. Instead, the information about a subject's *genotype*, a combination of the haplotypes from a pair of homologous chromosomes, is available. For homozygous subjects (who carry two copies of the same allele at all SNP loci) and heterozygous subjects who carry two copies of the same allele at all but one locus, the diplotype information is uniquely identifiable from the observed genotype data. For heterozygous subjects who carry different alleles at two or more loci, the problem of "phase ambiguity" arises. For example, consider two SNPs. The four possible haplotypes from two SNP loci are $H_1 = (0,0)$, $H_2 = (0,1)$, $H_3 = (1,0)$, and $H_4 = (1,1)$, where 0 and 1 indicate respectively the absence and presence of a mutation or polymorphism. Consider the heterozygous subject with the genotype $G = (1,1)$ recording the number of mutations in the pair of homologous chromosomes at two loci. There are two diplotypes, $\{H_1, H_4\}$ and $\{H_2, H_3\}$, which are consistent with this genotype, i.e., $G = H_1 + H_4 = H_2 + H_3$. Therefore, for this subject the phase is indeterminant. The other 8 possible genotypes—$(0,0)$, $(0,1)$, $(0,2)$, $(1,0)$, $(1,2)$, $(2,0)$, $(2,1)$, and $(2,2)$—correspond to subjects for which the diplotypes can be recovered uniquely. This "phase ambiguity" problem can be viewed as a missing-data problem, i.e., the true diplotype is unobserved for the heterozygous subjects who carry different alleles at two or more loci. For unphased genotype data, complete nonparametric treatment of the genetic covariates may not be possible because of the identifiability issues (Epstein and Satten 2003). The assumption of HWE for the diplotype population is often used to resolve the phase information.

The recently developed methodology for the analysis of case–control genetic data can be divided into two classes: "prospective" (ignoring the retrospective sampling of the data; Lake et al. 2003; Zhao, Li, and Khalid 2003) and "retrospective" methods

(Epstein and Satten 2003; Chatterjee and Carroll 2005; Spinka, Carroll, and Chatterjee 2005; Lin, Zeng, and Millikan 2005; Lin and Zeng 2006; Chen, Chatterjee, and Carroll 2008; Lobach et al. Forthcoming). Overall, the retrospective methods have increased efficiency compared with the prospective methods; see, for example, Satten and Epstein (2004) and Spinka, Carroll, and Chatterjee (2005) for details.

Many complex diseases such as cancers, bipolar disorder, hypertension, diabetes, and schizophrenia are influenced by both genetic and environmental factors. Therefore, it is important to include both types of factors in the analysis and to explore gene–environment interactions. This motivated the development of the retrospective profile-likelihood methods for analyzing case–control haplotype data in the presence of environmental factors (Chatterjee and Carroll 2005; Spinka, Carroll, and Chatterjee 2005; Lin, Zeng, and Millikan 2005; Lin and Zeng 2006; Chen, Chatterjee, and Carroll 2008; Lobach et al. Forthcoming). Here we demonstrate the new Stata command `haplologit`, which implements the modified retrospective semiparametric profile-likelihood method of Spinka, Carroll, and Chatterjee (2005) and Lin and Zeng (2006) for the specific case of a rare disease, a single candidate gene in HWE, and haplotype–environment independence.

Previously, Stata commands were introduced for haplotype-based analysis of quantitative traits (Cleves 2005; Mander 2002) and binary traits with case–control data (Mander 2001). `haplologit` is designed for the efficient haplotype-based analysis of case–control data in the presence of environmental factors. It allows estimating both haplotype and environmental effects, as well as haplotype–environment interactions for case–control data, by using the retrospective profile-likelihood approach in the presence of missing and unphased genotype data.

The structure of the article is as follows: Section 2 introduces two motivating datasets that we will analyze later with `haplologit`. Section 3 presents the methodology underlying the command. Section 4 describes the `haplologit` command. Section 5 presents the analysis of the two datasets introduced in section 2. Section 6 demonstrates a simulation study conducted to investigate the properties of the implemented method. Section 7 contains final remarks. We provide the syntax, options, and saved results of `haplologit` in the appendix.

2 Motivating examples

Our motivating examples involve the data from a case–control study of colorectal adenoma, a precursor of colorectal cancer. We consider two datasets, CASR and NAT2, used, among others, in Peters et al. (2004), Lobach et al. (Forthcoming), Moslehi et al. (2006), and Chen, Chatterjee, and Carroll (2008). The first study was designed to investigate the interactions of dietary calcium intake and genetic variants in the calcium-sensing receptor (*CaSR*) region. The second study was designed to assess whether smoking-related risk of colorectal adenoma may be modified by certain haplotypes in *NAT2*, a gene known to be important in the metabolism of smoking-related carcinogens. From previous studies, the assumptions of a rare disease and genes in HWE are plausible for these data.

2.1 CASR data

The CASR data consist of 772 cases and 778 controls sampled from the screening arm of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute in the United States. The data also contain information on participants' dietary food intake and genotype data from three nonsynonymous SNPs in the *CaSR* region. Some subjects did not have measurements of calcium intake and were eliminated. The resulting dataset consists of 644 cases and 668 controls.

The genotype data consist of three nonsynonymous SNPs in exon 7 of the *CaSR* gene: R990G (rs1042636) with major/minor alleles A/G, Q1011E (rs1801726) with alleles C/G, and A986S (rs1801725) with alleles G/T (Peters et al., 2004). The genotype data are stored in the variables `g_casr_01`, `g_casr_02`, and `g_casr_03`.

One of the main goals of the study was to investigate the interaction of dietary calcium intake (mg/day) and the three common haplotypes ACT, AGG, and GCG (coded as "001", "010", and "100", respectively). Other rare haplotypes are combined with the most common haplotype, ACG ("000"), to form the base (comparison) haplotype category. The variable `Ldtcal` records the values of the log of dietary calcium intake plus one. Adding one is not necessary with our data because, as we can infer from the output, there are no values with zero calcium intake. However, zeros are often observed with other nutrition data, and so we use the conventional transformation for the dietary calcium intake here. We also wish to adjust for such environmental factors as subjects' ages (variable `agerand`), genders (variable `sex`), and races (Caucasian or not, variable `Caucasian`). The disease status is recorded in variable `casecontrol`, with a value of one corresponding to a case. Below we provide descriptions and basic summaries of the variables in the dataset.

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

variable name	storage type	display format	value 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

```
. 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

2.2 NAT2 data

The NAT2 data consist of 628 prevalent advanced adenoma cases and 635 gender-matched controls, selected from the screening arm of the PLCO Cancer Screening Trial (Gohagan et al. 2000; Moslehi et al. 2006). One of the main objectives of this study was to assess whether smoking-related risk of colorectal adenoma may be modified by certain haplotypes in *NAT2*, a gene known to be important in the metabolism of smoking-related carcinogens. Subjects were genotyped at six SNPs (C282T, T341C, C481T, G590A, A803G, and G857A) in exon 2 of the *NAT2* gene. The dataset includes the genotype information—recorded in variables `g1`, `g2`, `g5`, `g6`, `g7`, and `g8`—and the following four environmental factors: age recorded in years (variable `age`); gender (variable `gender`); and two dummy variables, `smk1` and `smk2`, identifying "former" and "current" smokers, respectively. The disease status is recorded in variable `d`.

```
. describe g1 g2 g5 g6 g7 g8 d age smk1 smk2 gender
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
g1              byte   %8.0g                  SNP locus 1
g2              byte   %8.0g                  SNP locus 2
g5              byte   %8.0g                  SNP locus 3
g6              byte   %8.0g                  SNP locus 4
g7              byte   %8.0g                  SNP locus 5
g8              byte   %8.0g                  SNP locus 6
d               byte   %8.0g                  disease indicator
age             float  %9.0g                  age (in years)
smk1            byte   %8.0g                  smoking status: 1 = Former smoker
smk2            byte   %8.0g                  smoking status: 1 = Current
                                                 smoker
gender          byte   %8.0g                  gender: 0 = Male, 1 = Female
. summarize g1 g2 g5 g6 g7 g8 d age smk1 smk2 gender
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          g1 |      1263    1.163895    .7116161          0          2
          g2 |      1263    .0577989    .2334552          0          1
          g5 |      1263    1.347585    .6734789          0          2
          g6 |      1263    .8780681     .720551          0          2
          g7 |      1263     1.16152    .7127156          0          2
-------------+--------------------------------------------------------
          g8 |      1263     .594616    .6640195          0          2
           d |      1263    .4972288    .5001904          0          1
         age |      1263    62.56838    5.269487    55.0637    74.9897
        smk1 |      1263    .3982581    .4897331          0          1
        smk2 |      1263    .2034838    .4027487          0          1
-------------+--------------------------------------------------------
      gender |      1263    .3246239    .4684196          0          1
```

In both of our examples, we are interested in investigating the association between certain haplotypes (and environmental factors) and the risk of having colorectal adenoma. We are also interested in exploring the effect of certain haplotype–environment interactions on this risk. We use `haplologit` to perform these analyses, which are demonstrated in section 5.

3 Methodology

This section describes the methodology underlying the `haplologit` command. If you are not interested in theoretical details, you can skip this section and proceed to the description of the `haplologit` command in section 4.

The `haplologit` command implements the retrospective profile-likelihood methods of Spinka, Carroll, and Chatterjee (2005) and Lin and Zeng (2006), which are equivalent under the assumptions of a rare disease and HWE.

Suppose that subjects are genotyped at M tightly linked biallelic SNPs. Let $H_k = (h_{1k}, \ldots, h_{Mk})$ be a haplotype from the M SNPs and $H^{\text{hap}} = \{H_k,\, k = 1, \ldots, K = 2^M\}$ be the set of all possible haplotypes from M SNP sites. Each component $h_{ik} \in \{0, 1\}$ represents the absence or presence of a mutant (minor) allele at the ith SNP site for $i = 1, \ldots, M$. Let $H^d = (H_k, H_l)$ be a subject's diplotype with the constituent haplotypes H_k and H_l, and let $H^{\text{dip}} = \{(H_k, H_l) \colon H_k, H_l \in H^{\text{hap}}\}$ be the set of all possible haplotype pairs (all possible diplotypes). Let Y be the disease indicator (the case–control status) and $X = (X_1, \ldots, X_p)^\top$ be environmental factors.

The risk of a disease given a subject's genetic and environmental covariates is modeled by logistic regression. Under the assumption of a rare disease,

$$\Pr(Y = y \,|\, H^d, X) \approx \exp[y\{\beta_0 + m(X, H^d; \boldsymbol{\beta}_1)\}], \quad y = 0, 1$$

where β_0 is an unknown intercept, $\boldsymbol{\beta}_1$ is a vector of unknown coefficients, and $m(\cdot)$ is a known function whose general form is given in (1) below.

Consider the set $H^\star = \left\{H^\star_j,\, j \in J^\star \subset \{1, 2, \ldots, K\}\right\}$ of $K^\star < K$ risk haplotypes. Then the general form of the regression function is

$$m(X, H^d; \boldsymbol{\beta}_1) = \boldsymbol{\beta}_X^\top X + \sum_{j \in J^\star} \beta_{H^\star_j} Z(H^d, H^\star_j) + \sum_{j_1 \in J^\star} \sum_{j_2 = 1}^{p} \beta_{H^\star_{j_1} X_{j_2}} Z(H^d, H^\star_{j_1}) X_{j_2} \quad (1)$$

where $Z(H^d, H^\star_j)$ is defined according to one of the following three genetic (haplotype risk) models:

$$Z(H^d, H^\star_j) = \begin{cases} I(H_k = H^\star_j) + I(H_l = H^\star_j) & \text{if additive;} \\ I(H_k = H^\star_j) + I(H_l = H^\star_j) - I(H_k = H_l = H^\star_j) & \text{if dominant;} \\ I(H_k = H_l = H^\star_j) & \text{if recessive.} \end{cases}$$

$I()$ denotes the indicator function—it is 1 if the condition in parentheses is true, and 0 otherwise.

In (1), the first term represents the main effects of environmental factors X, the second term defines the main additive (or dominant, or recessive) effects of risk haplotypes $H^\star$, and the last term defines the interaction effects of risk haplotypes and environmental covariates. The available regression functions can be derived from (1) by setting various coefficients to zero. For example, suppose we have two environmental factors, X_1 and X_2, and two risk haplotypes, $H^\star_1$ and $H^\star_2$. We want to include the main effects

of all factors and two additive interaction effects of $H_1^\star$ and X_1, and $H_2^\star$ and X_2. Then the regression function (1) corresponds to

$$m(X, H^d; \boldsymbol{\beta}_1) = \beta_{X_1} X_1 + \beta_{X_2} X_2 + \sum_{j=1}^{2} (\beta_{H_j^\star} + \beta_{H_j^\star X_j} X_j)\{I(H_k = H_j^\star) + I(H_l = H_j^\star)\}$$

This reduces to $m(X, H^d; \boldsymbol{\beta}_1) = \beta_{X_1} X_1 + \beta_{X_2} X_2 + 2\beta_{H_j^\star} + 2\beta_{H_j^\star} X_j$ for a subject with a diplotype $H^d = (H_j^\star, H_j^\star)$, $j = 1, 2$; to $m(X, H^d; \boldsymbol{\beta}_1) = \beta_{X_1} X_1 + \beta_{X_2} X_2 + \beta_{H_1^\star} + \beta_{H_2^\star} + \beta_{H_1^\star} X_1 + \beta_{H_2^\star} X_2$ for a subject with a diplotype $H^d = (H_1^\star, H_2^\star)$; and so on.

The true diplotype $H^d = (H_k, H_l)$ is not directly observed. Instead, a genotype $G = (g_1, \ldots, g_M) = H_k + H_l$ is observed. Each component $g_i \in \{0, 1, 2\}$ represents the number of mutant alleles in a pair of homologous chromosomes at the ith SNP site for $i = 1, \ldots, M$. Therefore, genotype data provide incomplete information about which combination of alleles appears along each of the individual chromosomes. As such, the genotype G may be consistent with multiple diplotypes H^d. This is the case for heterozygous subjects who carry different alleles at two or more loci. For these subjects, the phase is ambiguous. For homozygous subjects (who carry two copies of the same allele at all loci) and heterozygous subjects who carry two copies of the same allele at all but one locus, the diplotype information is uniquely identifiable. When G has missing components, a subject's genetic information consists of multiple possible genotypes and, therefore, is consistent with multiple diplotypes. Such phase ambiguity can be viewed as a missing-data problem. The algorithm assumes that genotype data are missing at random and accounts for this by integrating the distribution of the missing data out of the joint likelihood for the observed and missing data.

It is assumed that H^d is independent of X. The distribution of H^d is governed by HWE:

$$\begin{aligned} \Pr\{H^d = (H_k, H_l); \boldsymbol{\theta}\} &= \theta_k^2 \qquad \text{if } H_k = H_l \\ &= 2\theta_k \theta_l \quad \text{if } H_k \neq H_l \end{aligned}$$

where θ_k denotes the frequency for haplotype H_k.

Let (Y_i, X_i, G_i), where $i = 1, \ldots, N$, be the observed disease indicator, the set of environmental variables, and the SNP genotype for the ith subject, respectively. Let N_1 and N_0 denote, respectively, the number of cases and controls in the sample, with $N = N_1 + N_0$. The retrospective likelihood $\prod_{i=1}^{N} \Pr(G_i, X_i \mid Y_i)$ involves a possibly infinite-dimensional (if X has continuous components) nuisance distribution of X. Therefore, it is desirable to profile the distribution of X out of the likelihood prior to estimating the parameters of interest, $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}$.

(*Continued on next page*)

Let $H_G^{\rm dip} = \{(H_k, H_l)$: the haplotype pair is consistent with $G\}$ be the set of all possible diplotypes consistent with the observed genotype data G; $\mu = \beta_0 + \log(N_1/N_0) - \log\{\Pr(Y=1)/\Pr(Y=0)\}$; $\Omega = (\mu, \boldsymbol{\beta}_1, \boldsymbol{\theta})$; and $S(Y, X, H^d; \Omega) \approx \Pr(H^d; \boldsymbol{\theta}) \exp[Y\{\mu + m(X, H^d; \boldsymbol{\beta}_1)\}]$ (under the assumption of a rare disease). The profile retrospective log likelihood is then

$$l^\star(\Omega) = \sum_{i=1}^{N} \left[\log \left\{ \sum_{H^d \in H_{G_i}^{\rm dip}} S(Y_i, X_i, H^d; \Omega) \right\} - \log \left\{ \sum_{y=0}^{1} \sum_{H^d \in H^{\rm dip}} S(y, X_i, H^d; \Omega) \right\} \right] \quad (2)$$

The semiparametric maximum-likelihood estimators for μ, $\boldsymbol{\beta}_1$, and $\boldsymbol{\theta}$ can be obtained from maximizing (2) by Newton–Raphson. The intercept β_0 is not estimable unless the probability of a disease in the population $\Pr(Y=1)$ is known. If $\Pr(Y=1)$ is known, then β_0 can be estimated as $\widehat{\beta}_0 = \widehat{\mu} - \log(N_1/N_0) + \log\{\Pr(Y=1)/\Pr(Y=0)\}$, where $\widehat{\mu}$ is the maximum likelihood estimate (MLE) of μ and is labeled as `_cons` in the output. The initial values of μ and $\boldsymbol{\beta}_1$ are set to zero, and the initial values of haplotype frequencies $\boldsymbol{\theta}$ are obtained from the control sample (if `emsample()` is omitted) by using the expectation maximization (EM) algorithm described below. Subjects' diplotypes with initial frequencies of constituent haplotypes less than $\max(2/N, 0.001)$ (or `hfthreshold(#)`) are not used during the maximization for the purpose of numerical stability.

`haplologit` allows the user to fit haplotype-based models in the absence of environmental factors. Here $\Omega = (\boldsymbol{\beta}_H, \boldsymbol{\theta})$; $S(Y, H^d; \Omega) = \Pr(H^d; \boldsymbol{\theta}) \exp\{Y m(H^d; \boldsymbol{\beta}_H)\}$; and the retrospective log likelihood is

$$l(\Omega) = \sum_{i=1}^{N} \left[\log \left\{ \sum_{H^d \in H_{G_i}^{\rm dip}} S(Y_i, H^d; \Omega) \right\} - \log \left\{ \sum_{H^d \in H^{\rm dip}} S(Y_i, H^d; \Omega) \right\} \right]$$

which corresponds to the approach of Epstein and Satten (2003).

Initial estimates of the haplotype frequencies are obtained by using the EM algorithm. Let $B_i = \{(k,l)$: $(H_k, H_l) \in H_{G_i}^{\rm dip}, k, l = 1, \ldots, K\}$ be the set of indices of diplotypes consistent with the observed genotype G_i. The observed data likelihood is $L(\boldsymbol{\theta}; G) = \prod_{i=1}^{N} \sum \sum_{(k,l) \in B_i} \theta_k \theta_l$, and the expected full-data log likelihood given the observed data is

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} \Pr_{\boldsymbol{\theta}}\{H_i^d = (H_k, H_l) \,|\, G_i\} (\log \theta_k + \log \theta_l) \quad (3)$$

During the expectation step of the algorithm, we evaluate the expected log likelihood (3) at the current values of haplotype frequencies, $\widehat{\boldsymbol{\theta}}^{(m-1)}$. During the maximization step, we maximize (3) with respect to $\boldsymbol{\theta}$ to obtain the estimates $\widehat{\theta}_k^{(m)}$. We iterate

between the two steps until the maximum number of iterations is reached or until convergence ($\sum_{k=1}^{K} |\widehat{\theta}_k^{(m)} - \widehat{\theta}_k^{(m-1)}| < \epsilon = 10^{-6}$), whichever comes first. Sometimes, the EM algorithm may converge to a local maximum. As such, it is recommended to try different starting values for the haplotype frequencies. By default, `haplologit` uses equal frequencies as starting values: $\theta_k^{(0)} = 1/K$, where $k = 1, \ldots, K$. This can be changed by specifying the option `eminit()`.

4 The haplologit command

The command `haplologit` estimates haplotype effects and haplotype–environment interactions from case–control genetic (SNP-based) data for one of three types of genetic (haplotype risk) models: additive, dominant, or recessive. It fits haplotype-effects logistic regression by using the retrospective profile-likelihood method in the special case of a rare disease and a single candidate gene in HWE, under the assumption of gene–environment independence. It allows simultaneous estimation of the effects of multiple risk haplotypes and their interactions with environmental factors.

The observed SNP genotype data is supplied to `haplologit` as subjects' genetic information via the required option `snpvars()`. The genotype data must be recorded in the so-called SNP variables (one variable for each SNP locus), containing only values of 0, 1, 2, and missing (`.`). Data must be in the wide form—a single observation per subject.

The effect of a single gene in HWE on the disease is considered. A *risk haplotype* (or causal haplotype) is a target haplotype whose effect on a disease is of interest. The effects of risk haplotypes can be modeled according to one of three genetic models, specified in the option `inheritance()` as the mode of inheritance: additive (the default), dominant, or recessive. Genetic covariates are viewed as functions of subjects' SNP genotype data and risk haplotypes. Specifically, genetic covariates depend on the number of copies of a risk haplotype present in the subject's diplotype. Their functional forms are determined by the selected genetic model. For example, under the additive risk model, having two copies of a risk haplotype in a subject's diplotype doubles the effect of this haplotype on a disease compared with having only one copy of a risk haplotype. In contrast, under the dominant risk model having one or two copies has the same effect on the disease. Under the recessive model, having only two copies of a risk haplotype has an effect on the disease. `haplologit` uses genetic covariates indirectly in the computation via the supplied information about SNP genotypes (option `snpvars()`), the genetic model (option `inheritance()`), and the risk haplotypes.

A risk haplotype can be specified as a string of a sequence of zeros and ones (binary representation) or as a haplotype index (position of a risk haplotype in the ordered sequence of 2^M possible haplotypes at M SNP sites). Risk haplotypes are specified in the options `riskhap1()`, `riskhap2()`, and so on. By default, if no risk haplotypes are specified, `haplologit` uses the most frequent haplotype estimated from the control sample (or the sample specified in `emsample()`) as the risk haplotype. Environmental covariates can be specified following the dependent variable in the syntax of `haplologit`.

The interaction effects of environmental factors and risk haplotypes can be included by using `riskhap#()`'s suboption `interaction()`.

The distributional assumptions on genetic covariates are HWE and independence with environmental covariates. Environmental covariates can be both continuous and discrete, and their distribution is left unspecified.

`haplologit`'s estimation process consists of three stages: data management, initial estimation of haplotype frequencies, and estimation of haplotype and optionally environmental effects. During the data-management stage, `haplologit` performs data manipulations necessary for handling unphased and missing SNP genotypes in the computation. At the second stage, the initial haplotype frequencies are estimated from the sample specified in `emsample()` by using the EM algorithm. Only the haplotypes with the estimated initial frequencies exceeding a default threshold (or an alternate threshold specified in `hfthreshold()`) are retained for further estimation; this is necessary for numerical stability of the algorithm. At the third stage, the coefficients for environmental covariates, risk haplotypes, and their interactions are estimated simultaneously with the haplotype frequencies by Newton–Raphson. The command displays information from and, optionally, progress at each of the three steps.

See the appendix for details of syntax, options, and saved results.

5 Empirical data analysis

We demonstrate the use of `haplologit` for the analysis of two previously described datasets, CASR and NAT2 (see section 2), containing data from a case–control study of colorectal adenoma.

5.1 Analysis of the CASR data

From the EM algorithm applied to the control sample (see output below), we identified the following four haplotypes with frequencies exceeding 0.001: ACG coded as "000" (0.710), ACT coded as "001" (0.150), AGG coded as "010" (0.055), and GCG coded as "100" (0.084). These results agree with those obtained by Peters et al. (2004) and Lobach et al. (Forthcoming). Our goal is to investigate the effects of the three common haplotypes "001", "010", and "100" and their interaction with dietary calcium intake (`Ldtcal`) on the risk of colorectal adenoma (`casecontrol=1`). The most common haplotype, ACG ("000"), and other rare haplotypes are treated as the base haplotype category. We also estimate effects for age, gender, and race (Caucasian or not).

We consider three regression models: the first model (RM1) includes only main effects of the three haplotypes of interest; the second model (RM2) extends the RM1 model by adding main effects of the environmental factors `Ldtcal`, `agerand`, `sex`, and `Caucasian`; and the third model (RM3) adds the interaction effects of `Ldtcal` with haplotypes "010" and "100". We fit all regression models under each of the three modes of inheritance (additive, dominant, and recessive).

Using haplologit

Below we consider the dominant risk-haplotype model (`inheritance(dominant)`). The recessive or additive models can be fit similarly by specifying `inheritance(recessive)` or `inheritance(additive)` (the latter can be omitted for the additive genetic model).

`haplologit` follows the same syntax as most Stata estimation commands in the specification of a dependent variable and independent covariates (environmental factors). The necessary genetic information is supplied in the options `snpvars()`, `inheritance()`, `riskhap1()`, `riskhap2()`, and so on.

First, we fit the model including the main dominant effects of haplotypes (RM1). We specify the required information about subjects' genotypes (SNP variables) in the option `snpvars(g_casr_01 g_casr_02 g_casr_03)` and choose the dominant mode of inheritance. To include the main effects of our haplotypes of interest (risk haplotypes) in the regression model, we specify the haplotypes in `riskhap1("001")`, `riskhap2("010")`, and `riskhap3("100")`.

```
. haplologit casecontrol, snpvars(g_casr_01 g_casr_02 g_casr_03)
> inheritance(dominant) riskhap1("001") riskhap2("010") riskhap3("100")

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

Performing gradient-based optimization:

Iteration 0:   Retrospective log likelihood = -1881.6901
Iteration 1:   Retrospective log likelihood = -1874.7763
Iteration 2:   Retrospective log likelihood =  -1874.757
Iteration 3:   Retrospective log likelihood =  -1874.757

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(3)       =      9.71
Retrospective log likelihood =  -1874.757        Prob > chi2        =    0.0212
```

casecontrol	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_001	-.2915435	.1238471	-2.35	0.019	-.5342793	-.0488077
hap_010	-.3595679	.1860106	-1.93	0.053	-.7241419	.0050062
hap_100	-.2205036	.1515137	-1.46	0.146	-.517465	.0764578

Haplotype frequencies

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_000	.7029724	.0121205	58.00	0.000	.6792165	.7267282
hap_001	.1527179	.0095216	16.04	0.000	.1340559	.1713798
hap_010	.0578313	.0063137	9.16	0.000	.0454566	.0702059
hap_100	.0864785	.0075397	11.47	0.000	.071701	.101256

The first part of the output displays the results of the EM algorithm used to estimate the haplotype frequencies in the control-only data. Only estimated frequencies exceeding the default threshold are displayed (and used in the subsequent analysis). The default haplotype frequency threshold in our example is 0.0015 and is determined by 2/1312. This can be changed by specifying the option `hfthreshold()`.

The second part of the output displays results from the gradient-based optimization. The table header provides general information about the fitted model. `Mode of inheritance` reports the type of genetic model (`dominant`). The assumed distribution of genetic covariates is HWE. The variables containing genotype information (SNP variables) are listed under `Genotype`.

`Number of obs` reports the number of subjects used in the computation. From the output, all 1,312 subjects (644 cases and 668 controls) in the dataset are used in the computation. `haplologit` also reports the numbers of subjects with observed phased genotypes (`Number phased`), with unphased genotypes (`Number unphased`), and with incomplete genotypes in which at least one SNP variable contains a missing value (`Number missing`). According to the output, 1,253 subjects have phased genotypes, and 59 subjects have unphased genotypes.

The first estimation table reports the estimates of haplotype effects. For example, the log odds-ratio of haplotype "001" is -0.292, suggesting a reduction in the risk of adenoma for subjects carrying this haplotype. The second table reports MLE of the haplotype frequencies.

Next we add four environmental factors to the regression model (RM2) as covariates `sex`, `Ldtcal`, `agerand`, and `Caucasian`. We use the abbreviated versions of options `snpvars()` and `inheritance()`, and we suppress the output from the EM stage and the ML iteration log by using the options `noemshow` and `nolog`, respectively.

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

Building consistent haplotype pairs:

Obtaining initial haplotype frequency estimates from the control sample:

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(7)       =     25.42
Retrosp. profile log likelihood = -2775.9764     Prob > chi2        =    0.0006
```

casecontrol	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.1230768	.122593	-1.00	0.315	-.3633547	.117201
Ldtcal	-.1959127	.113829	-1.72	0.085	-.4190134	.0271879
agerand	.037118	.0105986	3.50	0.000	.0163451	.0578908
Caucasian	.1514149	.2515958	0.60	0.547	-.3417038	.6445336
hap_001	-.2915435	.1238471	-2.35	0.019	-.5342793	-.0488077
hap_010	-.3595678	.1860106	-1.93	0.053	-.7241418	.0050063
hap_100	-.2205036	.1515137	-1.46	0.146	-.517465	.0764578
_cons	.1253638	.293798	0.43	0.670	-.4504696	.7011972

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

Haplotype frequencies
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hap_000	.7029724	.0121205	58.00	0.000	.6792165	.7267282
hap_001	.1527179	.0095216	16.04	0.000	.1340559	.1713798
hap_010	.0578313	.0063137	9.16	0.000	.0454566	.0702059
hap_100	.0864785	.0075397	11.47	0.000	.071701	.101256

The estimates of the environmental factors suggest that the risk of colorectal adenoma is higher for older people and for Caucasians, while the risk is lower for females and for people with higher calcium intake. The estimates of the haplotype effects remain unchanged as expected under the haplotype-environment independence assumption. Also, here, the estimates of the environmental effects are the same as those from the conventional logistic model (`logit`).

In the final regression model (RM3), we investigate the interaction effects of risk haplotypes with the covariate `Ldtcal`. We also recenter variables `Ldtcal` and `agerand` at their sample mean values before fitting the model so that the main haplotype effects correspond to the mean values of age and calcium. To include the haplotype–environment interaction effects, we specify environmental covariates in `riskhap#()`'s suboption `interaction()`. In the following input, we specify `inter(Ldtcal)` in options `riskhap2()` and `riskhap3()` to include the interaction effects of `Ldtcal` with haplotypes "010" and "100". We also use the option `happrefix()` to specify a short prefix to fit haplotype labels in the output tables.

```
. 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))
> noemshow nolog happrefix("_")
Building consistent haplotype pairs:
Obtaining initial haplotype frequency estimates from the control sample:
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)}
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

In table 1, we summarize the results (coefficients) from the three regression models for the dominant genetic model. The respective odds ratios can be obtained by exponentiating coefficients; `haplologit` does it automatically if the option `or` is used.

Table 1. Estimates of environmental and dominant haplotype effects (log odds) from three regression models for the CASR study. Standard-error estimates are shown in parentheses. The significance of the 0.05-level Wald test is labeled as follows: $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.005$.

Factors	RM1	RM2	RM3
`sex`	—	−0.1231 (0.1226)	−0.1223 (0.1226)
`Ldtcal`	—	−0.1959 (0.1138)*	−0.0553 (0.1214)
`agerand`	—	0.0371 (0.0106)***	0.0371 (0.0106)***
`Caucasian`	—	0.1514 (0.2516)	0.1579 (0.2518)
`hap_001`	−0.2915 (0.1238)**	−0.2915 (0.1238)**	−0.2915 (0.1239)**
`hap_010`	−0.3596 (0.1860)*	−0.3596 (0.1860)*	−0.4371 (0.1933)**
`hap_100`	−0.2205 (0.1515)	−0.2205 (0.1515)	−0.2507 (0.1536)
`hap_010*Ldtcal`	—	—	−0.7947 (0.2760)***
`hap_100*Ldtcal`	—	—	−0.5047 (0.2206)**

The model with only haplotype effects (RM1) shows a significant dominant effect of the "001" haplotype on the risk of colorectal adenoma; that is, the presence of the "001" haplotype in a subject's diplotype reduces the risk of colorectal adenoma. Other haplotypes are associated with a reduced risk as well but without significant statistical evidence at the 5% level. The addition of the environmental factors to the model demonstrates a statistically significant impact of age on the risk: age is associated with an increased risk of the disease. Other statistically nonsignificant findings include an increased risk of advanced colorectal adenoma for Caucasians, males, and subjects with lower calcium intake. The last model, RM3, demonstrates interesting results. The inclusion of the interaction of the log calcium intake (`Ldtcal`) with the haplotype "010" reveals this haplotype's statistical significance at the 5% level that was not visible from previously fitted models. Both interaction terms, `hap_010*Ldtcal` and `hap_100*Ldtcal`, are statistically significant at the 1% and 5% levels, respectively. Among the subjects having at least one copy of the haplotype "010" (or "100") in their diplotype, those with higher values of dietary calcium intake have lower risk of colorectal adenoma. Peters et al. (2004) conducted a prospective logistic analysis of these data and found no multiplicative interaction effects of total calcium intake with the *CaSR* common variants to be statistically significant after additionally adjusting for other environmental factors (smoking, alcohol intake, energy intake, etc.). Lobach et al. (Forthcoming) considered a model similar to ours without adjusting for age, gender, and ethnic origin and with additive haplotype effects. They also found the interactions of log calcium intake with haplotypes AGG (combined with the rare haplotypes) and GCG to be statistically significant after taking into account possible measurement error in dietary calcium intake.

In table 2, we summarize the Akaike's information criterion (AIC) obtained for nine models (the three regression models each fit assuming additive, dominant, and recessive modes of inheritance). The dominant genetic model has the smallest AIC for all three

regression models. The large differences in values between the RM1 and RM3 (or RM2) models are due to different likelihood functions being used in each. The model containing only main haplotype effects (RM1) is based on the true retrospective likelihood, whereas the models including environmental factors (RM2 and RM3) are based on the retrospective profile likelihood. As such, it is not appropriate to use AIC to compare models with environmental factors with those without environmental factors.

Table 2. AIC for the additive, dominant, and recessive genetic models considered for the CASR data

Genetic model	RM1	RM2	RM3
additive	3766.942	5579.381	5568.302
dominant	**3761.514**	**5573.953**	**5565.199**
recessive	3762.808	5575.247	5575.913

5.2 Analysis of the NAT2 data

We wish to investigate the effects of certain common haplotypes in *NAT2* (a gene important in the metabolism of smoking-related carcinogens) on the risk of colorectal adenoma. More importantly, we want to assess whether smoking-related risk of colorectal adenoma can be modified by certain haplotypes.

From the EM algorithm applied to the control sample, we obtain the following seven haplotypes with frequencies exceeding 0.002: "001100" (0.377), "100011" (0.302), "101010" (0.238), "110010" (0.028), "101100" (0.027), "001110" (0.019), and "001010" (0.007). The four common haplotypes—"100011", "101010", "110010", and "101100" (with estimated frequencies exceeding 1%)—are chosen as risk haplotypes. The other rare haplotypes are combined with the most frequent haplotype, "001100", into the base category.

Similarly to our CASR example, we fit three regression models with haplotype effects only (RM1); haplotype and environmental effects of age, gender, and two smoking indicators (RM2); and interaction effects of the haplotype "101010" with smoking (RM3). The last model allows us to investigate whether the risk associated with smoking is modified by the presence of one of the chosen risk haplotypes "101010". We report results for the RM3 model only, because those for the RM1 and RM2 models were similar to the results for RM3.

Using haplologit

In our call to `haplologit`, we include the main effects of environmental factors `age`, `smk1`, `smk2`, and `gender`; the main effects of haplotypes "100011", "101100", and "110010"; and the main and interaction effects of haplotype "101010" with covariates `smk1` and `smk2` (within the option `riskhap2()`).

```
. haplologit d age smk1 smk2 gender, snpvars(g1 g2 g5 g6 g7 g8) nolog
> happrefix("_") riskhap1("100011") riskhap2("101010", inter(smk1 smk2))
> riskhap3("101100") riskhap4("110010")

Building consistent haplotype pairs:

Obtaining initial haplotype frequency estimates from the control sample:

Haplotype frequency EM estimation

Number of iterations  =         28
Sample log-likelihood =  -1458.197
```

haplotype	frequency*
001010	.006492
001100	.376972
001110	.018898
100011	.302362
101010	.23839
101100	.026965
110010	.028346

```
  * frequencies > .0015835

Performing gradient-based optimization:
Note: removing 2 observations; constituent haplotype frequencies are smaller
      than .0015835

Haplotype-effects logistic regression
Mode of inheritance: additive                    Number of obs      =      1261
Genetic distribution: Hardy-Weinberg equilib.    Number phased      =       450
Genotype: g1 g2 g5 g6 g7 g8                      Number unphased    =       811
                                                 Number missing     =         0
                                                 Wald chi2(10)      =     54.09
Retrosp. profile log likelihood = -3702.5759     Prob > chi2        =    0.0000
```

d	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.051942	.0110732	4.69	0.000	.0302389	.073645
smk1	.0709409	.1494879	0.47	0.635	-.22205	.3639318
smk2	.9690708	.177891	5.45	0.000	.6204108	1.317731
gender	.0023537	.1255334	0.02	0.985	-.2436873	.2483947
_100011	-.1145104	.0955996	-1.20	0.231	-.3018822	.0728613
_101010	-.1186759	.1365992	-0.87	0.385	-.3864054	.1490537
_101100	-.0970695	.2520265	-0.39	0.700	-.5910324	.3968933
_110010	-.0391023	.2420902	-0.16	0.872	-.5135904	.4353859
_101010*smk1	.0761621	.1553604	0.49	0.624	-.2283387	.3806628
_101010*smk2	-.2912431	.1824043	-1.60	0.110	-.6487489	.0662628
_cons	-.0908722	.1483782	-0.61	0.540	-.3816882	.1999437

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

(*Continued on next page*)

Haplotype frequencies

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_001010	.0054919	.0014702	3.74	0.000	.0026104	.0083735
_001100	.3774709	.013276	28.43	0.000	.3514504	.4034914
_001110	.0190909	.002715	7.03	0.000	.0137696	.0244122
_100011	.3033175	.0129196	23.48	0.000	.2779956	.3286395
_101010	.2391732	.0119945	19.94	0.000	.2156644	.2626819
_101100	.0270196	.0045714	5.91	0.000	.0180598	.0359793
_110010	.028436	.0046715	6.09	0.000	.0192801	.0375919

Here the number of observations used in the computation (1,261) is smaller than the total number of subjects (1,263—628 cases and 635 controls) in the dataset. According to the note displayed during the gradient-based optimization, 2 subjects had frequencies of constituent haplotypes less than the default threshold, 0.0015835, and thus were removed from the estimation. Out of 1,261 subjects, 811 have unphased genotypes and all have complete genotype information.

Results

We fit all three types of genetic models—additive, dominant, and recessive. The estimates of the effects of haplotypes "110010" and "101100" had large standard errors under the recessive model, according to which only subjects with exactly two copies of the haplotype contribute to the estimation of the effect. The estimated frequencies of these haplotypes are low in the observed sample with respective homozygous diplotype frequencies of less than 0.09%. Therefore, there were not enough subjects in the sample to reliably estimate the recessive effects of these two haplotypes, which led to the large standard errors. The above output shows results from the additive model. However, the same inference about the haplotype and environmental effects can be made from the dominant model.

All three regression models relate the four target haplotypes to a reduced risk for colorectal adenoma (negative estimates of the log-odds parameters), although without significant statistical evidence. The only statistically significant effects are observed for age and "current" smokers (relative to nonsmokers). Both of them are positively associated with the risk of disease. The estimated log odds for `age` is 0.0519 with a standard error of 0.0111, and for `smk2`, it is 0.9691 with a standard error of 0.1779. None of the interaction terms are statistically significant at 10%. The negative estimate of the `_101010*smk2` effect agrees with results from previous epidemiological studies: among current smokers, the risk for colorectal adenoma is reduced for carriers of the haplotype "101010" compared with noncarriers. In fact, Chen, Chatterjee, and Carroll (2008) reveal the statistical significance of this interaction term after relaxing the gene–environment independence assumption by modeling the conditional distribution of the haplotypes as a function of the smoking status.

6 Simulation

Spinka, Carroll, and Chatterjee (2005); Lin, Zeng, and Millikan (2005); and Lin and Zeng (2006) present simulations for general methods relaxing the assumptions of a rare disease and HWE. Spinka, Carroll, and Chatterjee (2005) demonstrate simulation results without using the rare-disease approximation. Lin, Zeng, and Millikan (2005) and Lin and Zeng (2006) perform simulations for the case when the population is in Hardy–Weinberg disequilibrium (HWD). We conducted a simulation study to investigate the performance of the implemented method for the special case of a rare disease and HWE.

We reproduced one of the simulation scenarios considered in Lin, Zeng, and Millikan (2005) but assumed that the genotype population is in HWE. We used three SNP sites and generated haplotypes H_1, H_2, H_3, and H_4 with the following respective frequencies: 0.62, 0.27, 0.07, and 0.04. The haplotype frequencies of the other four possible haplotypes were set to zero. The diplotypes were generated assuming HWE. We generated the disease status according to the following logistic model:

$$\text{logit}\{\Pr(Y = 1 \mid H^d, X)\} = \alpha + \beta_X X + \beta_{H_2} N_2(H^d) + \beta_{H_2 X} N_2(H^d) X$$

where X is a Bernoulli random variable with 0.3 success probability and $N_2(H^d)$ is the number of copies of the haplotype H_2 in a diplotype H^d (the additive genetic model). The intercept was set to -4.7 to yield an approximately 1% disease rate in the population and to -3.1 to yield an approximately 5% disease rate. To investigate the haplotype effect, we set $\beta_X = \beta_{H_2 X} = 0.3$ and varied β_{H_2} from -0.3 to 0.3. To investigate the effect of the haplotype–environment interaction, we set $\beta_X = \beta_{H_2} = 0.3$ and varied $\beta_{H_2 X}$ from -0.3 to 0.3. We randomly selected 500 cases and 500 controls from a larger random sample of subjects. The simulation results are presented in tables 3 and 4 and are based on 1,000 replications.

(*Continued on next page*)

Table 3. Simulation results with $\Pr(Y = 1) = 1\%$. Bias and SE are the bias and the standard error of the semiparametric MLE for a parameter. SEE is the mean of the standard-error estimator. CP is the coverage probability of a 95% confidence interval. Power is the power of a 0.05-level Wald test of the null hypothesis H_0: $\beta = 0$.

Parameter			Bias	SE	SEE	CP	Power
β_{H_2}	=	−0.3	−0.0010	0.1263	0.1269	0.947	0.657
β_X	=	0.3	0.0074	0.1544	0.1567	0.941	0.516
β_{H_2X}	=	0.3	−0.0026	0.1552	0.1586	0.942	0.493
β_{H_2}	=	0.0	0.0017	0.1204	0.1211	0.952	0.048
β_X	=	0.3	0.0091	0.1603	0.1646	0.947	0.488
β_{H_2X}	=	0.3	−0.0054	0.1442	0.1494	0.943	0.540
β_{H_2}	=	0.3	0.0022	0.1114	0.1115	0.951	0.777
β_X	=	0.3	0.0051	0.1633	0.1668	0.947	0.459
β_{H_2X}	=	−0.3	−0.0026	0.1539	0.1592	0.945	0.481
β_{H_2}	=	0.3	0.0040	0.1134	0.1138	0.953	0.760
β_X	=	0.3	0.0099	0.1652	0.1684	0.940	0.462
β_{H_2X}	=	0.0	−0.0064	0.1432	0.1473	0.940	0.060
β_{H_2}	=	0.3	0.0053	0.1164	0.1160	0.954	0.736
β_X	=	0.3	0.0175	0.1677	0.1702	0.942	0.472
β_{H_2X}	=	0.3	−0.0145	0.1365	0.1367	0.948	0.545

Table 4. Simulation results with $\Pr(Y = 1) = 5\%$. Bias and SE are the bias and the standard error of the semiparametric MLE for a parameter. SEE is the mean of the standard-error estimator. CP is the coverage probability of a 95% confidence interval. Power is the power of a 0.05-level Wald test of the null hypothesis H_0: $\beta = 0$.

Parameter			Bias	SE	SEE	CP	Power
β_{H_2}	=	−0.3	−0.0025	0.1258	0.1244	0.943	0.678
β_X	=	0.3	0.0128	0.1548	0.1563	0.954	0.527
β_{H_2X}	=	0.3	−0.0102	0.1554	0.1534	0.953	0.466
β_{H_2}	=	0.0	0.0016	0.1201	0.1209	0.945	0.055
β_X	=	0.3	0.0175	0.1605	0.1629	0.948	0.517
β_{H_2X}	=	0.3	−0.0226	0.1449	0.1431	0.948	0.500
β_{H_2}	=	0.3	−0.0067	0.1117	0.1154	0.946	0.749
β_X	=	0.3	−0.0017	0.1632	0.1658	0.949	0.468
β_{H_2X}	=	−0.3	0.0111	0.1543	0.1561	0.954	0.449
β_{H_2}	=	0.3	−0.0005	0.1136	0.1171	0.941	0.754
β_X	=	0.3	0.0091	0.1651	0.1691	0.945	0.473
β_{H_2X}	=	0.0	−0.0089	0.1441	0.1450	0.948	0.052
β_{H_2}	=	0.3	0.0097	0.1163	0.1198	0.947	0.758
β_X	=	0.3	0.0298	0.1674	0.1710	0.951	0.498
β_{H_2X}	=	0.3	−0.0429	0.1375	0.1399	0.931	0.476

From tables 3 and 4, we can see that the biases of all parameter estimates are small. The biases increase slightly as the probability of a disease in a population increases from 1% to 5%. The estimates of standard errors agree with the observed variability in the parameter estimates. The 95% confidence intervals have coverage that is very close to nominal. The Wald tests have reasonable powers. The simulation results also agree with those given in Spinka, Carroll, and Chatterjee (2005) and Lin, Zeng, and Millikan (2005) for more general methods allowing a nonrare disease and HWD.

We also investigated the effect of HWD on the implemented method by simulating genotypes according to the following model (results are not shown):

$$\Pr\{H^d = (H_k, H_l); \boldsymbol{\theta}, \rho\} = (1 - \rho)\theta_k\theta_l + I(k = l)\rho\theta_k$$

We noticed increased bias in the parameter and standard-error estimates in the presence of the extreme excess homozygosity or heterozygosity (for $|\rho| > 0.1$). For more moderate values of ρ ($0.05 < |\rho| < 0.1$), we observed only a slight increase in biases and standard errors.

(*Continued on next page*)

7 Discussion

We have described the new Stata command `haplologit`, which implements the retrospective profile-likelihood method for the analysis of case–control genetic data under the additive, dominant, and recessive models in the special case of a rare disease, a single candidate gene in HWE, and independence of genetic and environmental factors. The `haplologit` command was developed under Stata 10.

Our empirical data analyses and simulation results support the necessity for the developed retrospective profile-likelihood methodology to be available to researchers. Also the NAT2 example and simulation results for a population in HWD demonstrate the importance of relaxing the assumptions of HWE and gene–environment independence—matters to be considered for future versions of `haplologit`.

`haplologit` is not designed for genome-wide SNP association analysis in which the goal is to locate a genetic region associated with the disease from a very large number (hundreds or thousands) of SNP loci. `haplologit` is instead intended for smaller regions (containing 5–10 tightly linked SNPs) to investigate further the effects of gene variants (haplotypes) from those regions on the disease. If studies use a large number of markers, then the regions to be studied would need to be divided into smaller groups, otherwise haplotype frequencies will become too small. Also `haplologit` assumes that subjects are unrelated and is thus not appropriate for family studies.

The algorithms used are computationally intensive. Execution time increases significantly with an increased number of SNP loci and haplotype and environmental effects. The presence of many subjects with missing or unphased genotypes increases the execution time as well.

8 Acknowledgment

This work was supported by the NIH Phase I SBIR grant "Statistical Software for Genetic Association Studies" (1 R43 GM079831-01A1) to StataCorp LP.

9 References

Andersen, E. B. 1970. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* 32: 283–301.

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.

Chatterjee, N., and R. J. Carroll. 2005. Semiparametric maximum likelihood estimation exploiting gene–environment independence in case–control studies. *Biometrika* 92: 399–418.

Chen, Y.-H., N. Chatterjee, and R. J. Carroll. 2008. Retrospective analysis of haplotype-based case–control studies under a flexible model for gene–environment association. *Biostatistics* 9: 81–99.

Cleves, M. A. 2005. Exploratory analysis of single nucleotide polymorphism (SNP) for quantitative traits. *Stata Journal* 5: 141–153.

Epstein, M. P., and G. A. Satten. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73: 1316–1329.

Gohagan, J. K., P. C. Prorok, R. B. Hayes, and B. S. Kramer. 2000. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials* 21(6 Suppl): 251S–272S.

Lake, S. L., H. Lyon, K. Tantisira, E. K. Silverman, S. T. Weiss, N. M. Laird, and D. J. Schaid. 2003. Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous. *Human Heredity* 55: 56–65.

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

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. Forthcoming. Haplotype-based regression analysis and inference of case–control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*.

Mander, A. P. 2001. Haplotype analysis in population-based association studies. *Stata Journal* 1: 58–75.

———. 2002. Analysis of quantitative traits using regression and log-linear modeling when phase is unknown. *Stata Journal* 2: 65–70.

Moslehi, R., N. Chatterjee, T. R. Church, J. Chen, M. Yeager, J. Weissfield, D. W. Hein, and R. B. Hayes. 2006. Cigarette smoking, n-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics* 7: 819–829.

Niu, T., Z. S. Qin, X. Xu, and J. S. Liu. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics* 70: 157–169.

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 and Prevention* 13: 2181–2186.

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

Satten, G. A., and M. P. Epstein. 2004. Comparison of prospective and retrospective methods for haplotype inference in case–control studies. *Genetic Epidemiology* 27: 192–201.

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.

Weir, B. S. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates.

Zhao, L. P., S. S. Li, and N. Khalid. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case–control studies. *American Journal of Human Genetics* 72: 1231–1250.

About the authors

Yulia Marchenko is a senior statistician at StataCorp.

Raymond J. Carroll is a distinguished professor of statistics, nutrition, and toxicology at Texas A&M University. He is a leading figure in statistical research in general. He was the editor of two major journals (*Biometrics* and *Journal of the American Statistical Association*) and the chair of the NIH SNEM-5 and BMRD study sections.

Danyu Lin is a Dennis Gillings distinguished professor of biostatistics at the University of North Carolina. He has made fundamental contributions to the designs and analysis of medical studies, including genetic epidemiologic studies. He has served as an associate editor for *Biometrika* and *Statistica Sinica* since 1999 and for numerous NIH study sections.

Christopher I. Amos is a professor of epidemiology at the M. D. Anderson Cancer Research Center. He has made fundamental contributions to genetic epidemiology. He is the statistical core leader for the North American Rheumatoid Consortium, the Genetic Epidemiology of Lung Cancer Consortium, and a P01 project on the genetic basis of childhood cancers. He served as the president of the International Genetic Epidemiology Society in 2002. He previously served as an associate editor for *American Journal of Human Genetics* and is currently on the editorial board of *Genetic Epidemiology*.

Roberto G. Gutierrez is the director of statistics at StataCorp.

Appendix

A.1 Syntax

`haplologit` *depvar* [*indepvars*] [*if*] [*in*], `snpvars(`*varlist*`)` [*options*]

options	description
Model	
* `snpvars(`*varlist*`)`	specify SNP variables
`inheritance(`*inhmode*`)`	specify mode of inheritance
`riskhap(`*riskhap_spec*`)`	specify a single risk haplotype
`riskhap`#`(`*riskhap_spec*`)`	specify #th risk haplotype
`hfthreshold(`#`)`	retain observations with initial haplotype frequencies exceeding `hfthreshold()`; default is $\max(2/N, 0.001)$
`constraints(`*numlist*`)`	apply specified linear constraints on environmental factors *indepvars*
`collinear`	keep collinear variables
`noconstant`	suppress constant term
Reporting	
`level(`#`)`	set confidence level; default is `level(95)`
`or`	report odds ratios
`happrefix(`*string*`)`	use *string* as a prefix when labeling haplotypes in the output; default is `happrefix(hap_)`
`alldots`	show all iterations (except ML) as dots
`nocoef`	suppress coefficients table
`nofreq`	suppress haplotype-frequency table
`noheader`	suppress output header

(*Continued on next page*)

EM	
`emsample(controls|cases|all)`	obtain initial haplotype frequencies from the specified sample; default is `emsample(controls)`
`emiterate(#)`	number of EM iterations; default is `emiterate(500)`
`emtolerance(#)`	EM convergence tolerance; default is 1e−6
`eminit(`*matname*`)`	specify matrix containing starting values of haplotype frequencies for EM estimation
`sort`	sort haplotypes by frequencies in the EM output
`emlog`	show EM iteration log
`emdots`	show EM iterations as dots
`noemshow`	suppress output from EM estimation
`noemtable`	suppress EM haplotype-frequency table
Max options	
maximize_options	control maximization process; see section 8

⋆ `snpvars(`*varlist*`)` is required.

inhmode	description
`additive`	additive mode of inheritance; the default
`dominant`	dominant mode of inheritance
`recessive`	recessive mode of inheritance

riskhap_spec is

riskhap_str | # [, *riskhap_suboptions*]

riskhap_str specifies the binary representation of a risk haplotype enclosed in quotes, or # specifies a risk haplotype index (the position of a risk haplotype in the ordered sequence of 2^M possible haplotypes from M SNP sites).

riskhap_suboptions	description
`interaction(`*varlist*`)`	specify interaction variables
`noconstant`	suppress constant term (seldom used)

A.2 Options

Model options

`snpvars(`*varlist*`)` is required; it specifies SNP variables (variables recording subjects' SNP genotypes). The SNP variables must contain values of 0, 1, 2, or missing (`.`). A missing value (`.`) indicates missing information at a SNP site; other values represent the number of copies of a mutant (minor) allele at a SNP site in a subject's pair of homologous chromosomes.

`inheritance(`*inhmode*`)` specifies a mode of inheritance (a genetic model). The default is the `additive` risk model in which having two copies of a risk haplotype in a pair of homologous chromosomes results in a twofold effect of the risk haplotype on a disease. The `dominant` risk model assumes that having one or two copies of a risk haplotype has the same effect on a disease. The `recessive` model assumes that having only two copies of a risk haplotype has an effect on a disease.

`riskhap(`*riskhap_spec*`)` requests to include effects of the specified risk haplotype in a regression model. `riskhap()` is a synonym for `riskhap1()`.

`riskhap`#`(`*riskhap_spec*`)` requests to include effects of the #th risk haplotype in a regression model. If `interaction(`*varlist*`)` is specified with `riskhap`#`()`, the respective interaction effects of the risk haplotype with the covariates specified in *varlist* are also included in the regression model. If `noconstant` is used, the main haplotype effect is omitted and only haplotype–environment interaction effects are included in the model (seldom used).

`interaction(`*varlist*`)` specifies variables to be interacted with the specified risk haplotype.

`noconstant` requests that the constant term (the main effect of a risk haplotype) is not included in the model (seldom used). This option requires `interaction()`.

`hfthreshold(`#`)` specifies to retain in the computation only diplotypes with initial frequencies of constituent haplotypes exceeding #. The default is $\max(2/N, 0.001)$, where N is the total number of cases and controls.

`constraints(`*numlist*`)` and `collinear`; see [R] **estimation options**. `constraints()` can be used to define linear constraints on only environmental covariates *indepvars*.

`noconstant` suppresses the constant term.

Reporting options

`level(`#`)`; see [R] **estimation options**.

`or` reports the estimated coefficients transformed to odds ratios, i.e., $\exp(b)$ rather than b. Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated. `or` can be specified at estimation or when replaying previously estimated results.

`happrefix(`*string*`)` uses the specified *string* as a prefix when labeling haplotypes in the output (except for the EM output). The default prefix is `hap_`.

`alldots` specifies that iterations from all (possibly time-consuming) computations be shown as dots except for the ML iterations. `alldots` implies `emdots`.

`nocoef` specifies that the coefficient table not be displayed.

`nofreq` specifies that the haplotype-frequency table not be displayed.

`noheader` suppresses the output header, either at estimation or upon replay.

EM options

`emsample(controls | cases | all)` requests that the initial haplotype frequencies be estimated from the control sample, case sample, or combined case–control sample. The default is to use the control sample.

`emiterate(#)` specifies the number of EM iterations to perform. The default is `emiterate(500)`.

`emtolerance(#)` specifies the convergence tolerance for the EM algorithm. The default is 1e−6. The EM algorithm terminates when the maximum relative change in estimated haplotype frequencies between successive iterations is less than #.

`eminit(`*matname*`)` specifies the $1 \times L$ matrix *matname* containing starting values of haplotype frequencies for EM estimation. If M is the number of SNP loci (SNP variables), then $L = 2^M - 1$. By default, all haplotypes are assumed to be equally likely; that is, all haplotype frequencies are set to $1/2^M$.

`sort` requests that haplotypes be displayed in descending order of frequencies in the EM haplotype-frequency table. By default, haplotypes are displayed according to their binary ordering.

`emlog` specifies that the EM iteration log be shown. The EM iteration log is, by default, not displayed.

`emdots` specifies that the EM iterations be shown as dots. This option can be convenient when the EM algorithm requires many iterations to converge.

`noemshow` suppresses the output from the EM estimation.

`noemtable` suppresses the EM haplotype-frequency table.

Max options

maximize_options: `difficult`, `iterate(#)`, `[no]log`, `trace`, `hessian`, `gradient`, `showstep`, `tolerance(#)`, `ltolerance(#)`, `gtolerance(#)`, `nrtolerance(#)`, `nonrtolerance`, `shownrtolerance`; see [R] **maximize**.

By default, convergence is declared when the `nrtolerance()` criterion and either the `tolerance()` or `ltolerance()` criterion have been met. If `nonrtolerance` is specified, then convergence is declared when either the `tolerance()` or `ltolerance()` criterion has been met.

If `gtolerance()` is specified, then the `gtolerance()` criterion must be met in addition to any other required criterion for convergence to be declared. See [R] **maximize** for more information on the different types of `tolerance` options.

A.3 Saved results

`haplologit` saves the following in `e()`:

Scalars

`e(N)`	number of observations (subjects)	`e(df_m)`	model degrees of freedom
`e(N_phased)`	number of phased genotypes	`e(chi2)`	chi-squared
`e(N_unphased)`	number of unphased genotypes	`e(p)`	significance of model test
`e(N_miss)`	number of incomplete genotypes	`e(em_N)`	number of EM observations
`e(ll)`	retrospective (profile) log likelihood	`e(em_ll)`	EM log likelihood
`e(converged)`	`1` if converged, `0` otherwise	`e(cutoff)`	haplotype-frequency threshold
`e(rc)`	return code		

Macros

`e(cmd)`	`haplologit`	`e(inheritance)`	mode of inheritance
`e(cmdline)`	command as typed	`e(genepop)`	genetic distribution
`e(depvar)`	name of dependent variable	`e(emsample)`	a sample used for EM
`e(snpvars)`	names of SNP variables	`e(happrefix)`	haplotype prefix

Matrices

`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix
`e(em_freq)`	initial haplotype-frequency vector		

Functions

`e(sample)`	marks estimation sample

For other saved results, see [R] **maximize**.