



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Haplotype analysis in population-based association studies

A. P. Mander
MRC Biostatistics Unit, Cambridge, UK
adrian.mander@mrc-bsu.cam.ac.uk

Abstract.

This paper describes how to use the command `hapipf` and introduces the command `profhap` written for Stata that analyzes population-based genetic data. For these studies, association can be linkage disequilibrium within a set of loci or allelic/haplotype association with disease status. Confidence intervals for odds ratios are calculated with or without adjustment for possible factors that are confounding the relationship. Additionally, this command allows the specification of many models of association that are not widely implemented.

Keywords: st0003, haplotype analysis, association tests, profile likelihood, odds ratio

1 Introduction

Many genetic analysis programs are written as stand-alone programs. This may be more efficient in terms of computer resources but not in terms of ease of use, flexibility, and availability. This paper describes advanced use of the command `hapipf` introduced in Mander (2001) and covers the issues adjusting for confounders, missing data, effect modification, calculating odds ratios with confidence intervals, and grouping haplotypes. The code for `hapipf` has been updated for Stata 7 and now includes a Windows interface to help users specify the correct syntax.

The command discussed here permits analysis of population-based association studies and is similar to packages such as EH described in Terwilliger and Ott (1994). This command extends the methods by allowing models that are not the saturated model to be fitted. Analysis using haplotypes requires an EM algorithm to resolve phase uncertainty, see Long, Williams, and Urbanek (1995); Fallin and Schork (2000); Hawley and Kidd (1995); Sham (1998); and Mander (2001). Here, a log-linear model is embedded within the EM algorithm to estimate the expected haplotype/allele frequencies rather than a counting algorithm, see Chiano and Clayton (1998), and allows a range of intermediate models. In genetic analysis, the dimensionality of the model is often large (due to the number of haplotypes/alleles) and maximum likelihood estimation of the log-linear model is performed using the iterative proportional fitting algorithm, see Agresti (1992). This algorithm always converges when the maximum likelihood estimates exist even when the likelihood is badly behaved.

In association studies, interest lies in testing for association between a set of loci and disease status. Odds ratios are calculated to show the strength and direction of this association. Confidence intervals for the odds ratio can be obtained by using bootstrap methods, see Efron and Tibshirani (1993), or can be constructed from the profile likelihood using a constrained log-linear model.

A significant allelic/haplotype association may be due to population admixture or ethnic stratification. In order to test for this unseen stratification, a set of unlinked markers can be used to investigate whether there are any associations between them, see Pritchard and Rosenberg (1999). If these data are unavailable, one approach is to obtain some information about this stratification using surrogate measures such as the origin of relatives. These data are included as part of a stratified analysis using log-linear modeling either to adjust for confounders or to investigate possible effect modifiers.

2 Syntax

```
hapipf varlist [using exp] [, ldim(varlist) display ipf(string) start known
    phase(varname) acc(#) ipfacc(#) nolog model(#) lrtest(,#)
    convars(string) confile(string) menu mv mvdel ]

profhap [if exp] , or(string) ipf(string) [ by(varlist) acc(#) level(#)
    hapacc(#) savegraph ]
```

2.1 Options for hapipf

`ldim(varlist)` specifies the variables that determine the dimension of the contingency table. By default, the variables contained in the `ipf` option define the dimension.

`display` specifies whether the expected and imputed haplotype frequencies are shown on the screen.

`ipf(string)` specifies the log-linear model. It requires special syntax of the form `11*12+13`. `11*12` allows all the interactions between the first two loci, and locus 3 is independent of them. This syntax is used in most books on log-linear modeling.

`start` specifies that the starting posterior weights of the EM algorithm are chosen at random.

`known` specifies that phase is known.

`phase(varname)` specifies a variable that contains ones where phase is known and zeros, where phase is unknown.

`acc(#)` specifies the convergence threshold of the change of the full log likelihood.

`ipfacc(#)` specifies the convergence threshold of the change in the log likelihood of the log-linear model.

`nolog` specifies whether the log likelihood is displayed at each iteration.

`model(#)` specifies a label for the log-linear model being fitted. This label is used in the `lrtest` option.

`lrtest(,#, #)` performs a likelihood-ratio test using two models that have been labeled in the `model` option.

`convars(string)` specifies a list of variables in the constraints file.

`confile(string)` specifies the name of the constraints file.

`menu` specifies that the syntax is specified using a window interface.

`mv` specifies that the algorithm should replace missing data (“.”) with a copy of each of the possible alleles at this locus. This is performed at the same stage as the handling of the missing phase when the dataset is expanded into all possible observations. If this option is not specified but some of the alleles do contain missing data, the algorithm sees the symbol “.” as another allele.

`mvdel` specifies that people with missing alleles are deleted.

2.2 Options for `profhap`

`or(string)` specifies first the case-control variable and then two haplotypes/alleles. The first haplotype/allele indicates that the unexposed group and the other haplotype/allele is the exposed group; for example, `or(D 1 2)` specifies that the case-control variable is D and that the allele 1 represents the unexposed group and allele 2 the exposed.

`ipf(string)` is the same as for `hapipf`.

`by(varlist)` specifies the stratifying variable.

`acc(#)` specifies the accuracy of the estimated upper and lower bounds of the confidence interval.

`level(#)` specifies the significance level of the confidence interval.

`hapacc(#)` specifies the convergence threshold of both `hapipf` and `ipf`.

`savegraph` specifies that the profile graph is saved to file `profile.gph`.

3 Methods

3.1 Resolving phase using the EM algorithm

For illustration, take two diallelic loci with alleles *a* and/or *A* at the first locus and *b* and/or *B* at the second locus. The haplotype is the set of alleles that occur on the same chromosome, for example, the *ab* haplotype. When phase is unknown, the parent origin

of the alleles cannot be determined and hence the haplotype cannot be constructed. Using the two diallelic loci, the unphased genotype for the double heterozygote may be represented as (a, A, b, B) and is henceforth referred to as the phenotype.

For subject i , when phase is unknown, the genotype frequencies follow a mixture distribution

$$\sum_{\tilde{g}_i \in G_i} \pi_{\tilde{g}_i}$$

where G_i is the set of all possible phases conditional on the phenotype, and $\pi_{\tilde{g}_i}$ is the probability of genotype \tilde{g}_i . Not all subjects will have phase ambiguity and this information is used to resolve phase in the EM algorithm. However, in order to determine phase in subjects that are heterozygotes, the Hardy–Weinberg assumption is needed. Under this assumption, the genotype probability is the product of the two haplotype probabilities, $\pi_{g_i} = p_{h_{1i}} p_{h_{2i}}$, and hence the EM algorithm is used to estimate haplotype probabilities.

The algorithm consists of the “E-step”, which calculates the posterior probability of each phase. The j th estimate of a particular phase g_i^* for subject i is $\hat{z}_{g_i^*}^{(j)}$. This phase probability is given below. Note that the expression is conditional on the previous estimates of the haplotype probabilities.

$$\hat{z}_{g_i^*}^{(j)} = \frac{\hat{p}_{h_{1i}}^{(j-1)} \hat{p}_{h_{2i}}^{(j-1)}}{\sum_{G_i} \hat{p}_{h_{1i}}^{(j-1)} \hat{p}_{h_{2i}}^{(j-1)}}$$

The “M-step” maximizes the full likelihood, given below, conditional on the phase probabilities.

$$\prod_{i=1}^n \prod_{g_i^* \in G_i} \{p_{h_{1i}} p_{h_{2i}}\}^{\hat{z}_{g_i^*}^{(j)}}$$

Traditionally this is performed by a counting algorithm but this command uses a log-linear model with the phase probabilities as weights. The most efficient algorithm for estimating these expected frequencies when there are numerous cells is iterative proportional fitting (IPF).

3.2 Iterative proportional fitting

This algorithm calculates the expected frequencies of a contingency table, see Agresti (1992). For a $2 \times 2 \times 2$ table, let the observed frequencies be n_{ijk} and the expected frequencies m_{ijk} . The algorithm consists of the following steps:

- Set $\hat{m}_{ijk}^{(0)}$ to exhibit less structure than the model being fitted. In other words, the initial expected frequencies come from a model that is nested in the model being fitted.

- Successively adjust $\hat{m}_{ijk}^{(0)}$ using appropriate scaling factors so they match each marginal table in the set of minimal sufficient statistics.
- Continue until changes in the likelihood are small.

The log-linear model is specified by the syntax first introduced by Wilkinson and Rogers (1973), where a “+” symbol indicates independence and the “*” indicates interaction. Given three factors X , Y and Z and the model $X * Y + X * Z + Y * Z$ (this is a model with all interactions except the three-way interaction), the minimal sufficient statistics are $n_{ij.}$, $n_{i.k}$ and $n_{.jk}$, where the “.” represents summation over that margin. The sufficient statistics can be identified by the marginal models, terms separated by the + symbol. In this example, there are three marginal models and the algorithm consists of the repetition of the following three steps:

$$\hat{m}_{ijk}^{(1)} = \hat{m}_{ijk}^{(0)} \left(\frac{n_{ij.}}{\hat{m}_{ij.}^{(0)}} \right) \quad (1)$$

$$\hat{m}_{ijk}^{(2)} = \hat{m}_{ijk}^{(1)} \left(\frac{n_{i.k}}{\hat{m}_{i.k}^{(1)}} \right) \quad (2)$$

$$\hat{m}_{ijk}^{(3)} = \hat{m}_{ijk}^{(2)} \left(\frac{n_{.jk}}{\hat{m}_{.jk}^{(2)}} \right) \quad (3)$$

For the first step it can be seen that $\hat{m}_{ij.}^{(1)} = n_{ij.}$, and the $X - Y$ margin is matched. For the other equalities the margins $X - Z$ and $Y - Z$ have expected frequencies equal to the observed frequencies, respectively.

This algorithm can be adapted to estimate models that have constrained parameters using initial expected frequencies from a model that is not nested in the fitted model. It can be shown that this initial structure remains unchanged during the steps. Let the factors in the above example all have two levels (0 or 1) and the model being fitted be $X + Y + Z$. If $m_{111}^{(0)} = 2$ and 1 for the other cells, then from the initial expected frequencies the odds ratio (in the 2×2 $X - Y$ sub-table) when $Z = 1$ is 2 (when $Z = 0$ the odds ratio is 1). This relationship does not change during the iterations. The model actually fitted is the *base model* $X + Y + Z$, with the $X.Y$ interaction parameter being fixed by the initial values.

4 Allelic association in case–control data

To show that a disease is associated with a particular marker allele, frequencies are compared between cases and controls. The algorithm ignores subjects and analyses the chromosome data under the assumption of Hardy–Weinberg Equilibrium (HWE). Deviations from HWE have an impact on the significance of the association test as the phase information changes, see Fallin and Schork (2000). Cases and controls are assumed to be representative of the affected and unaffected underlying population, the

study base. When the locus is diallelic (e.g., a SNP locus), a simple chi-squared test can be performed on the resulting 2×2 table to test for the association. The same test can be obtained using the log-linear model approach by constructing a variable, D , say, with two levels, where 1/0 represent case/control status respectively. Let the variable L_1 represent the diallelic locus with levels 1 and 2, which represent the alleles. When there is no association, the locus is independent of disease status and the expected allele frequencies are the same in cases and controls, this model is $L_1 + D$ and the odds ratio is 1. The alternative model, where the allele frequencies depend upon disease status, is $L_1 * D$ and the odds ratio is the maximum likelihood estimate. The test for allelic association is the likelihood-ratio test comparing the model $L_1 + D$ to $L_1 * D$, which is equivalent to testing whether the maximum likelihood estimate of the odds ratio is not equal to 1. This test has 1 degree of freedom (df), but for polymorphic markers, the degrees of freedom will increase, losing power to detect an association. High dimensional contingency tables occur when investigating highly polymorphic marker loci and algorithms that require derivatives of the likelihood will be slower than IPF.

4.1 Stata output and commands

The command here requires a variable list consisting of *paired* alleles and disease status, where the rows are the individuals. For illustration, data are taken from Sham (1998). The format of the data is shown below for the first six lines of the dataset (identification number is contained in `id`). The two columns `a1` and `a2` are the alleles for one autosomal locus and D is the variable of case/control data. As phase is not a problem for one allele, the first pair of columns are actually the genotype.

id	a1	a2	D
1	1	2	1
2	2	2	1
3	1	2	0
4	2	2	0
5	1	1	1
6	1	2	1

The association command `hapipf` requires a variable list of paired alleles. The log-linear model is specified using the `ipf()` option and takes the notation discussed above with loci labeled $l1, l2, l3, \dots$, and so on, in order of the list of paired variables (for a single locus this is $l1$). The likelihood and df can be saved internally using the `model()` option to label the models, and then the likelihood-ratio test can be automatically calculated using the two labeled models in the `lrtest()` option—the model with the fewer parameters is entered last.

```
. hapipf a1 a2, ipf(l1*D) model(0)
(output omitted)
```

```
. hapipf a1 a2, ipf(l1+D) model(1) lrtest(0,1)
      (output omitted)
Likelihood Ratio Test Comparing Model l1+D to l1*D
```

```
      llhd2 (df2)          = -395.60522 1
      llhd1 (df1)          = -391.78348 0
-2*(llhd2-llhd1)          = 7.6434815
Change in df              = 1
p-value                   = .00569778
```

There is a significant association ($p = 0.006$) between the locus and the disease status. This does not explain which of the two alleles increase the risk of disease or by how much.

The expected frequencies can be obtained by using the `display` option as seen in the first command below.

```
. hapipf a1 a2, ipf(l1*D) display
. hapipf a1 a2 using filename, ipf(l1*D)
```

Alternatively, the expected frequencies can be stored as a Stata datafile by using the `using` option. In the example above, the file `filename.dta` contains the frequencies.

4.2 Odds ratios and profile likelihood

To describe the relationship between the alleles and disease status, the odds ratio is calculated for the 2×2 table. The saturated model $L_1 * D$ from the last analysis gave estimates of the expected allele frequencies for cases and controls, and these are displayed below.

locus	D	efreq	eprob
1	0	32	.10738255
1	1	87	.29194631
2	0	76	.25503356
2	1	103	.34563758

To calculate the appropriate odds ratio, the analyst needs to identify which of the alleles is the exposed group and which is the unexposed group. In this case, interest is in the effect of the allele 1 on the probability of being a case and the odds ratio is $87 \times 76 / (103 \times 32) = 2.006$. Approximate confidence intervals can be obtained using the standard error of a log-odds ratio, see Clayton and Hills (1993), assuming normality. Alternatively, approximate confidence intervals can be obtained from a profile likelihood approach, see McCullagh and Nelder (1989). Let the parameter of interest in the 2×2 table be the odds ratio, θ , say, and let the $100(1-\alpha)\%$ confidence interval be (θ_l, θ_u) . The maximum likelihood estimate of the odds ratio is $\hat{\theta}$ and the log likelihood calculated at this point is $l(\hat{\theta})$. Thus, the $100(1-\alpha)\%$ confidence interval is the set of values $\{\theta : 2l(\hat{\theta}) - 2l(\theta) \leq \chi^2_{1,1-\alpha}\}$. To find θ_l and θ_u requires an algorithm to calculate the likelihood for various values of the odds ratio, a constrained log-linear model.

To fit a model that has an odds ratio value of x , say, requires a specific set of initial starting values in the IPF algorithm, see Section 3.2. The model $L_1 + D$ has the odds

ratio fixed at 1, and the following initial values will constrain the odds ratio to x . The base model should contain all the parameters that are not fitted by the initial values. In this example, the base model is $L_1 + D$ and the $L_1.D$ interaction is fitted by the initial values.

l1	D	Ifreq
1	0	1
1	1	1
2	0	1
2	1	x

If x was the maximum-likelihood estimate of the odds ratio, then the model being fitted is $L_1 * D$. The following constraints file, with a missing value instead of x , allows estimation of the maximum likelihood estimate of the odds ratio,

l1	D	Ifreq
1	0	1
1	1	1
2	0	1
2	1	.

where the “.” corresponds to a missing data point. For the data used in the previous section, the likelihood is calculated for a set of values for x in order to obtain the profile likelihood and confidence intervals. In the example dataset, the profile likelihood 95% confidence interval is (1.222, 3.343) and the profile is plotted in Figure 1 using a cubic spline to join the points. There are numerous points close to the edge of the confidence interval as the algorithm searches for the upper and lower bounds to a specified accuracy. The graph can sometimes be used to check that the model is specified correctly. The maximum likelihood estimate of the odds ratio is estimated without the constraint files, and if the base model is misspecified the maximum likelihood estimate will not lie on the correct profile. The profile likelihood can be used for two parameters of interest to deliver profile contours of a bivariate distribution of parameters but this is not implemented here.

(Graph on next page)

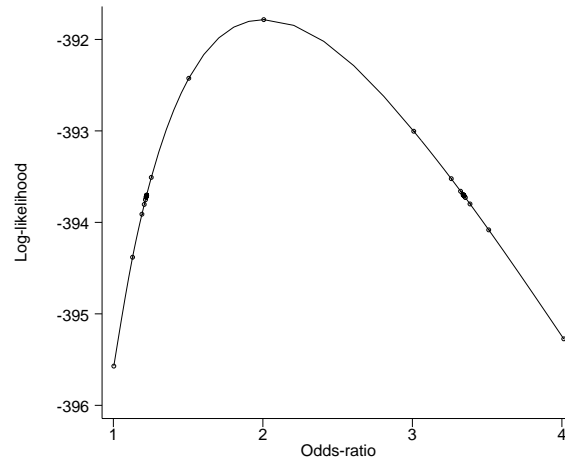


Figure 1: The profile likelihood for the odds ratio parameter in a diallelic locus association with disease.

4.3 Stata commands and output

The command `profhap` is used to obtain the profile likelihood estimates of the confidence intervals as it requires estimation of the likelihood using `hapipf` multiple times. `profhap` constructs the values of the odds ratio that are passed to `hapipf`, and it uses very similar syntax to `hapipf`. The command needs the set of loci as paired variable list and additionally requires the `or()` option that contains, in order, the case-control variable, what is the exposed category, and which is the unexposed.

```
. profhap a1 a2 , ipf(l1+D) or(D 2 1) hapacc(1e-11) acc(1e-4)
(output omitted)
      Case-control table
      +-----+
      Alleles    Cases    Controls
      2          103.0000    76.0000
      1           87.0000    32.0000
OR = 2.0061 with 95% CI interval (1.2221 ,3.3429 )
```

Note that here the odds ratio is slightly different from the one in the saturated model. This is because the constrained IPF algorithm does not converge in one step and is subject to the convergence criteria on the log likelihood. The convergence criteria is controlled by the option `hapacc()`. The option `acc()` specifies the accuracy of the estimated bounds of the confidence interval.

5 Haplotype analysis

The previous models and descriptions have been on contingency tables with four cells and therefore only one odds ratio is calculated. As the dimension of the tables increase,

there are many more models and odds ratios. When analyzing haplotypes, the phase is usually unknown and resolving the phase requires an EM algorithm. When phase is unknown, the haplotypes are not observed and the table of frequencies can not be constructed.

The EM algorithm expands the dataset over all possible haplotypes per phase and subject; this is a high-dimensional contingency table of imputed frequencies. As part of the expectation step (E-step), these frequencies are scaled by their posterior probability, assuming HWE and current estimates of the haplotype frequencies. The log-linear model is fitted to the imputed frequencies as part of the maximization step (M-step). For the saturated model, the imputed frequencies and the expected frequencies are equal.

5.1 Linkage disequilibrium (LD) in two loci

Association between marker loci is described as linkage disequilibrium, the simplest dataset has two diallelic marker loci. For illustration, the alleles are labeled 1 or 2 at each loci and the possible haplotypes can be seen in Table 1.

Table 1: The possible haplotypes for a two loci system.

		Loci 2	
		1	2
Loci 1	1	n_{11} $p_1 p_2 + \delta$	n_{12} $p_1(1 - p_2) - \delta$
	2	n_{21} $(1 - p_1)p_2 - \delta$	n_{22} $(1 - p_1)(1 - p_2) + \delta$

In Table 1, n_{ij} is the number of haplotypes that have allele i at locus 1 and allele j at locus 2. Below the counts are the probabilities that a random individual from the same population has that haplotype (p_i is the probability of allele i), and δ is the coefficient of linkage disequilibrium, see [Terwilliger and Ott \(1994\)](#). If loci are in linkage equilibrium, then the haplotype frequencies are the product of the corresponding allele frequencies. This occurs when $\delta = 0$ and corresponds to a nonsignificant χ^2 test of association between loci. A significant test would suggest that $\delta \neq 0$ and linkage disequilibrium is present between the loci.

In terms of log-linear models, the count data in the table is the dependent variable and for locus i , L_i is a factor variable with two levels 1 and 2. When there is linkage equilibrium the two factors are independent and when LD is present, there will be an interaction between the factors. The model of linkage equilibrium (independence) is $L_1 + L_2$. The extra term δ is included in the model by an interaction between the loci, $L_1.L_2$, in other words, the model $L_1 * L_2$. The estimated coefficient of the term $L_1.L_2$ is not a direct estimate of δ as the model has been reparameterized.

The likelihood ratio comparing the models $L_1 * L_2$ and $L_1 + L_2$ is the test for linkage disequilibrium. These tests can be extended to polymorphic markers; however, the de-

degrees of freedom will increase, and the power to detect significant linkage disequilibrium will therefore decrease.

5.2 Stata commands

Data are taken from [Terwilliger and Ott \(1994\)](#), and the form of the data is shown below. For the first 6 lines of data, the first two columns represent the alleles for one autosomal locus and the last pair of variables are the alleles for the second locus.

a1	a2	b1	b2
1	1	1	1
1	2	1	1
2	2	1	1
1	1	1	2
1	2	1	2
2	2	1	2

The syntax for linkage disequilibrium testing is nearly exactly the same as for allelic association. The main differences are that there are four variables and that the D term of the `ipf()` option is replaced by the second locus term $l2$.

```
. hapipf a1 a2 b1 b2, ipf(l1*l2) model(0)
(output omitted)
. hapipf a1 a2 b1 b2, ipf(l1+l2) model(1) lrtest(0,1)
(output omitted)
Likelihood Ratio Test Comparing Model l1+l2 to l1*l2
```

```
llhd2 (df2)          = -303.28056 1
llhd1 (df1)          = -298.83431 0
-2*(llhd2-llhd1)     = 8.8925028
Change in df         = 1
p-value              = .00286344
```

As seen from the output, the p -value is significant, and the removal of the $L_1.L_2$ term leads to a large drop in the log likelihood indicating that it should not be removed. In terms of the genetic hypothesis, this confirms there is strong evidence in support of LD between the two loci.

5.3 Testing LD in a region with more than 2 loci

Section 5.1 showed the test for LD using two loci, and now these tests will be applied to $k(> 2)$ loci of interest. The loci factor variables are L_1, L_2, \dots, L_k , with the alleles as levels. The test for linkage disequilibrium is similar to the two loci case where comparison, by the likelihood-ratio test, is between the saturated model, $L_1 * L_2 * \dots * L_k$, and the model of independence between the set of loci, $L_1 + L_2 + \dots + L_k$. The second model assumes that the haplotype frequencies can be obtained by multiplying the corresponding alleles frequencies. In other words, the second model assumes linkage equilibrium. If the loci are all diallelic, then the degrees of freedom of the test is $2^k - 1$. If every locus has n alleles, there are $n^k - 1$ degrees of freedom.

We take data from Bitti et al. (2000), in particular three loci in the HLA region, A, B, and DR. Interest lies in the one haplotype A30*B18*DR3 in the HLA region. Each of the loci are coded with two levels, i.e., whether the allele is present or not. The test statistic comparing models $L_1 * L_2 * L_3$ with $L_1 + L_2 + L_3$ is 331.1 on 4 degrees of freedom, a hugely significant result, and confirms that there is a huge amount of LD in the HLA region in this population.

Intermediate models

In the last section each haplotype is considered distinct. A parsimonious approach would be to group “similar” haplotypes. While this raises the question of the definition of “similar”, this approach increases the power to detect an association since there are fewer parameters. For example, the structural relationship between loci along the chromosome may be a first-order Markov process. In log-linear terms, a first order process is represented as $L_1 * L_2 + L_2 * L_3 + \dots + L_{k-1} * L_k$, see Chiano and Clayton (1998).

5.4 Case-control data

The likelihood-ratio test for an association between the set of haplotypes and disease compares the model $L_1 * L_2 * \dots * L_k * D$ to $L_1 * L_2 * \dots * L_k + D$. This test may be less informative than a conditional independence approach. If the loci are ordered along the chromosome from 1 to k , then the end loci, 1 and k , are tested for an association with disease conditional on the alleles at the other loci. For example, the model $D * L_2 * \dots * L_k + L_1 * \dots * L_k$ allows L_1 to be independent of disease status given the other loci. This is compared to the saturated model to test whether L_1 is conditionally independent of disease. A nonsignificant likelihood-ratio test indicates that locus 1 could be removed from the analysis. With enough power, and if the disease is affected by a single locus, this method should be able to reduce the set of loci to the locus that is closest to the disease locus. However, for polymorphic markers it still may not have enough power.

5.5 Stata command and output

For the HLA data, there are three loci A, B, and DR (in chromosomal order). From a univariate analysis, it is suggested that the allelic association was predominantly from the B locus. Because the DR locus was further away from B than A, this locus was tested for conditional independence.

```
. hapipf a1 a2 b1 b2 dr1 dr2, ipf(l1*l2*l3*D) model(0)
(output omitted)
```

```
. hapipf a1 a2 b1 b2 dr1 dr2, ipf(l2*l1*D+l1*l2*l3) model(1) lrtest(0,1)
(output omitted)
Likelihood Ratio Test Comparing Model l1*l2*D+l1*l2*l3 to l1*l2*l3*D
```

llhd2 (df2)	=	-1661.3706	4
llhd1 (df1)	=	-1660.9831	0
-2*(llhd2-llhd1)	=	.77505058	
Change in df	=	4	
p-value	=	.94176124	

For this analysis, the likelihood-ratio test is not significant. This suggests that the DR locus appears to be conditionally independent of disease given the A and B loci. In Bitti et al. (2000), the analysis continued on the table of frequencies collapsed over the DR margin.

6 Confounding and effect modification

When there is population admixture or ethnic stratification, any association may be confounded. The usual approach to this problem is to stratify the analysis by variables that may be confounding the association.

To demonstrate this, data are taken from a case-control study with one diallelic locus. The association between locus and disease can be summarized by the odds ratio. If there is a stratifying variable S with r levels, then the odds ratio comparing alleles with disease can be calculated for each level of S . Let these be $\theta_1, \theta_2, \dots, \theta_r$. These odds ratios are estimated from the saturated model $L_1 * D * S$. The classic epidemiology approach to control for the stratifying variable is by assuming a constant odds-ratio model. This means that the θ_i 's are all equal to a common value θ .

The model $L_1 * S + D * S$ is when L_1 and D are conditionally independent given S , and all the θ_i 's are 1. If S is a binary variable (two strata), the model to obtain the maximum likelihood estimate of the odds ratios θ_1 and θ_2 must include the terms $L_1.D$ and $L_1.D.S$ (this is the saturated model). Dropping the $L_1.D.S$ interaction constrains the odds ratios to be equal; $\theta_1 = \theta_2$. The model $L_1 * S + L_1 * D + D * S$ is therefore the model for a common odds ratio, and θ_1 is the adjusted odds ratio controlling for S . Inclusion of the $L_1.D.S$ term allows the stratifying variable to be an effect modifier. Comparison of the common odds-ratio model to the saturated model in this example allows for the test of effect modification or an interaction test.

The common odds-ratio model, $L_1 * S + L_1 * D + D * S$, can also be fitted using constraint files.

The base model is $L_1 * S + D * S$ and the other parameters can be fitted using constraints. The Stata data file (`strata1.dta`) is needed to fit the $L_1.D$ term and is given below.

	l1	D	Ifreq
1.	1	0	1
2.	1	1	1
3.	2	0	1
4.	2	1	.

There is only one odds ratio in the $L * D$ margin, and the missing value in the constraint file specifies that the command will produce the maximum likelihood estimate. For this file, the marginal model, $L_1 * D$, is defined by the variables included in the file, L_1 and D . In conjunction with the base model, this is the common odds-ratio model.

To fit the additional $L_1.D.S$ term requires the following file (**strata2.dta**):

	l1	D	Ifreq	S
1.	1	0	1	0
2.	1	0	1	1
3.	1	1	1	0
4.	1	1	1	1
5.	2	0	1	0
6.	2	0	1	1
7.	2	1	.	0
8.	2	1	.	1

The two missing values specify that the two odds ratios of the $L * D * S$ margin are the maximum likelihood estimates and are not constrained. Similarly, the three variables L_1 , D and S in this data file specify the marginal model to be $L_1 * D * S$, allowing two separate odds ratios to be fit.

6.1 Stata commands and output

The `confile()` option must be used to specify the Stata constraint datafiles (the file extension `.dta` is not needed). The first two commands use the constraint files and second two fit the same models, respectively, without the constraints.

```
. hapipf a1 a2, ipf(S*D+l1*S) confile(strata2) convars(l1 D S)
(output omitted)

. hapipf a1 a2, ipf(S*D+l1*S) confile(strata1) convars(l1 D)
(output omitted)

. hapipf a1 a2, ipf(S*D*l1) model(0)
(output omitted)

. hapipf a1 a2, ipf(S*D+l1*S+D*l1) model(1) lrtest(0,1)
(output omitted)
Likelihood Ratio Test Comparing Model S*D+l1*S+D*l1 to S*D*l1
```

```
llhd2 (df2)          = -1567.1607 1
llhd1 (df1)          = -1565.9427 0

-2*(llhd2-llhd1)     = 2.435933
Change in df         = 1
p-value              = .11858334
```

From the p -value, it can be seen that there is no evidence that the stratifying variable is an effect modifier.

Probably of more interest is the estimate of the common odds ratio and its confidence interval. This is estimated using profile likelihood, and by using the first file, the dot can be replaced by a series of specific odds ratios to obtain the profile. The use of the **profhap** command is very similar to the last example, except that a **by()** option is required to specify the stratifying variable. The default is to calculate the odds ratio in a randomly chosen strata, and hence the summary tables can differ from run to run, but as each strata shares a common odds-ratio model, the command will estimate the same odds-ratio and confidence interval.

```
. profhap a1 a2, ipf(S*D+I1*S) or(D 1 2) by(S) acc(1e-5)
(output omitted)
```

Case-control table			
		Cases	Controls
Alleles	1	36.8863	186.1137
	2	15.1138	43.8863

```
OR = 1.7376 with 95% CI interval (1.2212 ,2.4555 )
```

7 Confounding and effect modification using haplotypes

For the two diallelic loci example with a binary stratifying variable, there will be three odds ratios per stratum. There are four possible haplotypes labeled H_1, H_2, H_3, H_4 . Let the odds ratio comparing haplotype H_i to H_1 in stratum j be $\theta_j^{(i)}$. The model $L_1 * L_2 * D * S$ estimates the maximum likelihood values for the odds ratios $\theta_j^{(i)}$. Removing the $L_1 * L_2 * D * S$ term constrains the odds ratios according to the rule $\theta_0^{(i)} = \theta_1^{(i)}$ for all i ; the model is $L_1 * L_2 * S + S * D + L_1 * L_2 * D$. In this model, each haplotype is considered as a separate parameter, but could easily be the Markov model when dealing with a large number of loci. The test for effect modification of the stratifying variable is the likelihood-ratio test between this model and the saturated model (a 3 df test).

7.1 Grouping haplotypes

For two diallelic loci, there are 4 possible haplotypes. If there is some a priori reason that the association is due to only one of the haplotypes, then the effect modification test discussed previously will have lower power than one that groups the other 3 haplotypes as the comparison group. This would result in only one odds ratio per stratum and a 1 df test. As the phase is unknown, the raw data cannot be grouped and the desired model can only be achieved by constraining the odds ratios $\theta_j^{(i)} = 1$ when $i \neq 3$, thus giving the odds ratios of interest as $\theta_j^{(3)}$ for $j = 1, 0$.

Constraint files can specify the relationship between $\theta_0^{(3)}$ and $\theta_1^{(3)}$. The common odds-ratio model is when $\theta_0^{(3)} = \theta_1^{(3)}$, and this is compared to the model where $\theta_0^{(3)} \neq \theta_1^{(3)}$ for the test of effect modification. The base model will be $L_1 * L_2 * S + S * D$ with the $L_1 * L_2 * D$ margin being fit using the constraint files. The file **strata3.dta** below is the constraint file for the common odds model. Note that only one odds ratio is freely

estimated, and all the cells in the $L_1 * L_2 * D$ margin are specified and the exposure haplotype is 2.2.

	l1	l2	D	Efreqold
1.	1	1	0	1
2.	1	1	1	1
3.	1	2	0	1
4.	1	2	1	1
5.	2	1	0	1
6.	2	1	1	1
7.	2	2	0	1
8.	2	2	1	.

The file (`strata4.dta`) below is the constraint file for the effect modification for one specific haplotype. All the cells in the $L_1 * L_2 * D * S$ margin are specified and two odds ratios are allowed, and the base model is exactly the same.

	l1	l2	D	Efreqold	S
1.	1	1	0	1	0
2.	1	1	0	1	1
3.	1	1	1	1	0
4.	1	1	1	1	1
5.	1	2	0	1	0
6.	1	2	0	1	1
7.	1	2	1	1	0
8.	1	2	1	1	1
9.	2	1	0	1	0
10.	2	1	0	1	1
11.	2	1	1	1	0
12.	2	1	1	1	1
13.	2	2	0	1	0
14.	2	2	0	1	1
15.	2	2	1	.	0
16.	2	2	1	.	1

It may be that it is unreasonable that only one haplotype is associated with disease. Alternatively, there may be a “dose–response” relationship. In this case, three odds ratios, $\theta_0^{(i)}$, may have a functional relationship to the odds ratios in the other strata, e.g., $\alpha\theta_0^{(i)}$. Then the test of effect modification is the test of whether $\alpha = 1$, i.e., a 1 df interaction test. These types of model are not handled by this command.

7.2 Stata commands

The grouping of haplotypes requires the use of the constraint files. The following commands obtain the likelihood for both models:

```
. hapipf a1 a2 b1 b2, ipf(S*D+l1*l2*S) confile(strata3) convars(l1 l2 D)
. hapipf a1 a2 b1 b2, ipf(S*D+l1*l2*S) confile(strata4) convars(l1 l2 D S)
```

From the output the likelihood-ratio test statistic is .82741365 on 1 df, which is not significant at the 5% level.

It is also possible to perform the 3 df test of effect modification using the commands

```
. hapipf a1 a2 b1 b2, ipf(S*D*l1*l2)
. hapipf a1 a2 b1 b2, ipf(S*D+l1*l2*S+l1*l2*D)
```

The 3 df test statistic is 6.0321425, which is not significant as well.

8 Conclusions

This paper has discussed two commands: one for testing (`hapipf`) and one for calculating confidence intervals (`profhap`). Embedding the log-linear modeling within the EM algorithm has allowed a lot of flexibility in modeling. Although most of the discussion has been focused on case-control studies, the `hapipf` command allows the dependent variable to have many levels.

Missing-marker information is also handled by the algorithm assuming it is “missing at random”, see Little and Rubin (1987). The missing marker is assumed to be one of the observed alleles, and this is implemented when creating the imputed frequencies in the EM algorithm. The Stata option is `mv`.

Application of this command is ideal for SNP association studies as the degrees of freedom are low. The command can also be applied to polymorphic loci, but the power to detect an association may be low. Due to the sparse nature of these data the degrees of freedom may be overstated. Another possibility is in discovering relationships between loci that can lower the number of parameters in the saturated model.

9 References

- Agresti, A. 1992. Modeling patterns of agreement and disagreement. *Statistical Methods in Medical Research* 1: 201–18.
- Bitti, P., B. Murgia, A. Ticca, R. Ferrai, L. Musu, M. Piras, E. Puledda, S. Campo, S. Durando, C. Montomoli, D. Clayton, A. Mander, and L. Bernardinelli. 2000. Association between the ancestral haplotype HLA A30*B18*DR3 and multiple sclerosis in central sardini. *Genetic Epidemiology* 20: 271–83.
- Chiano, M. and D. Clayton. 1998. Fine genetic mapping using haplotype analysis and the missing data problem. *Annals of Human Genetics* 62: 55–60.
- Clayton, D. and M. Hills. 1993. *Statistical Models in Epidemiology*. London: Oxford University Press.
- Efron, B. and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Fallin, D. and N. Schork. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics* 67: 947–59.

- Hawley, M. and K. Kidd. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity* 86: 409–411.
- Little, R. and D. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- Long, J., R. Williams, and M. Urbanek. 1995. An EM algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics* 56: 799–810.
- Mander, A. P. 2001. sbe38: Haplotype frequency estimation using an EM algorithm and log-linear modeling. *Stata Technical Bulletin* 57: 5–7. In *Stata Technical Bulletin Reprints*, vol. 10, 104–107. College Station, TX: Stata Press.
- McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*, 2d ed. London: Chapman & Hall.
- Pritchard, J. and N. Rosenberg. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 65: 220–228.
- Sham, P. 1998. *Statistics in Human Genetics*. London: Edward Arnold.
- Terwilliger, J. and J. Ott. 1994. *Handbook of Human Genetic Linkage*. Baltimore: Johns Hopkins University Press.
- Wilkinson, G. and C. Rogers. 1973. Symbolic description of factorial models for analysis of variance. *Applied Statistics* 22: 392–399.

About the Author

Adrian Mander works at the MRC Biostatistics Unit in the field of genetic epidemiology and missing data.