

## Guides for *HapReg* Macro

By Yi-Hau Chen

yhchen@stat.sinica.edu.tw

The SAS macro *HapReg* implements the haplotype-based genetic association analysis for case-control studies, using a flexible model for gene-environment association allowing haplotypes to be potentially related with environmental exposures. The novel methodology is proposed by **Chen, Chatterjee, and Carroll** (*Biostatistics*, to appear).

The following provides typical syntax for a SAS code that implements the SAS macro *HapReg* ([macro\\_hapreg.sas](#)). Examples are given at the end of the document.

**Note:** The program can accommodate missing SNP data, which should be coded as . (dot; the usual coding for SAS missing data) or 9.

**Note:** To ensure numerical stability, it is suggested that the environmental covariates be centered at their respective means (subtracted with the means).

```
data a;
```

```
...
```

```
run;
```

```
/* the above codes specify the data to be analyzed */
```

```
data h;
```

```
input snp1-snp6;
```

```
cards;
```

```
0 0 1 1 0 0
```

```
1 0 0 0 1 1
```

```
1 0 1 1 0 0
```

```
...
```

```
;
```

```
/* the above codes specify the SNP haplotypes to be used as covariates in the logistic regression model. Note that the haplotypes used in the model are usually the common haplotypes in the control or whole population, which can be identified using the EM algorithm as suggested in eq.(14) of Chen and Kao (BMC Genetics 2006, 7:43), or
```

using the R package [haplo.core](#) by D. J. Schaid or other available packages for haplotype frequency estimation. In this example each haplotype contains 6 SNPS snp1-snp6, but in general the number of SNPs is not limited to 6. Each SNP must be coded as 0/1, and the first haplotype given in [cards](#) is the one chosen as the baseline haplotype, usually the most frequent haplotype \*/.

```
%include 'dir\macro_hapreg.sas';
```

```
/* include the HapReg macro stored in the file dir\macro_hapreg.sas, where dir is the full path name of the directory where macro_hapreg.sas is stored. */
```

```
%HapReg(data=a, hdata=h, d=1, zsel=2 3 10 11, zgsel=10 11, zzsel=10 11, snp=4 5 6 7 8 9, rare=5 7, hsel=2 3 4 6, hselg=6, mode=a pr1=0);
```

```
/* the arguments in HapReg macro:
```

**data=** a SAS dataset containing the data to be analyzed

**hdata=** a SAS dataset containing the SNP haplotypes to be used as covariates in the disease risk model

**d=i** specify the *i*th variable (from left to right) in dataset specified in **data** as the disease status data, which must be coded as 0/1

**zsel=k1 k2 ...**

specify the *k1,k2,...*th variables (can be more than one) in dataset specified in **data** as the environmental covariates to be included in the disease risk model

**zgsel=g1 g2 ...**

specify the *g1,g2...th* variables (can be more than one) in dataset specified in **data** as the environmental covariates in the disease risk model that have interactions with haplotypes specified in **hselg**

**zzsel=l1 l2 ...**

specify the *l1,l2...th* variables (can be more than one) in dataset specified in **data** as the environmental covariates to be included in the model for the

haplotype distribution in the control population

`snp=s1 s2 ...`

specify the  $s1, s2, \dots$ th variables (can be more than one) in dataset specified in `data` that contain the SNP genotype data to be analyzed

`rare=r1 r2 ...`

specify the  $r1, r2, \dots$ th rows (haplotypes) in dataset specified in `hdata` as the haplotypes that have smaller (e.g.,  $<1\%$ ) frequency and need to be grouped into the baseline haplotype; if no such haplotype it is set to “.” (dot)

`hsel=h1 h2 ...`

specify the  $h1, h2, \dots$ th rows (haplotypes) in dataset specified in `hdata` as the haplotypes to be considered in the disease risk model (not including the reference haplotype and the haplotypes grouped into the reference)

`hselg=t1 t2 ...`

specify the  $t1, t2, \dots$ th rows (haplotypes) in dataset specified in `hdata` as the haplotypes in the disease risk model that have interactions with environmental covariates specified in `zgsel`

`mode=`

the mode specified for the haplotype effects; where “`mode=a`” or “`mode=A`” specifies additive effects, while “`mode=d`” or “`mode=D`” specifies dominant effects

(recessive mode has not been implemented because with this mode more caution should be made to ensure the convergence of the algorithm when the involved haplotype frequency is small)

`pr1=`

specify a quantity between 0 and 1 that corresponds to the true population disease prevalence rate; when `pr1=0` is specified, a rare-disease approximation is performed

\*/

## **Example 1**

In the NAT2 data set described in the paper, the columns of the data are

- (1) Disease status
- (2) Gender
- (3) Age
- (4) SNP1
- (5) SNP2
- (6) SNP3
- (7) SNP4
- (8) SNP5
- (9) SNP6
- (10) SMK1 = former smoker
- (11) SMK2 = current smoker (note that the paper has a type)

After an initial analysis, we identified 7 haplotypes to have probability  $> 0.5\%$ . This led us to the following program coding.

```
data h; /*set the haplotypes to be analysed, which are obtained from an initial haplotype
estimation analysis*/
input snp1-snp6;
cards;
0 0 1 1 0 0
1 0 0 0 1 1
1 0 1 1 0 0
1 1 0 0 1 0
0 0 1 0 1 0
1 0 1 0 1 0
0 0 1 1 1 0
;
```

**run**;

The first haplotype (0 0 1 1 0 0) is the most frequent and is captured as the reference haplotype.

In the following code, the environmental variables all enter as main effects in the risk model, but only the smoking variables have interactions with the haplotypes. Haplotypes 2, 3, 4 and 6 enter the risk model (haplotype 1 is the reference, 5 and 7 are

fairly rare), with haplotype 6 being allowed to interact with the environmental variables. By setting `zzsel=.`, we are assuming that the distribution of the haplotypes does not depend on the environmental variables.

```
%HapReg(data=rayNAT2,hdata=h, d=1, zsel=10 11 2 3, zgsl=10 11, zzsel=., snp=4 5  
6 7 8 9, rare=., hsel=2 3 4 6, hselg=6, mode=a, pr1=0);
```

## **Example 2**

Continuing with the previous example, consider the following code.

```
%HapReg(data=rayNAT2,hdata=h, d=1, zsel=10 11 2 3, zgsel=11, zzsel=11, snp=4 5 6  
7 8 9, rare=5 7, hsel=2 3 4 6, hselg=6, mode=a, pr1=0);
```

In this program, the disease status is variable 1 ( $d=1$ ). The environmental variables that have main effects in the risk model are variables 2, 3, 10 and 11 ( $zsel=10\ 11\ 2\ 3$ ). The SNP data are in columns 4-9 ( $snp=4\ 5\ 6\ 7\ 8\ 9$ ). We count haplotypes 5 and 7 as rare haplotypes ( $rare=5\ 7$ ) that are to be grouped into the baseline haplotype. Haplotypes 2, 3, 4 and 6 are used in the disease risk model ( $hsel=2\ 3\ 4\ 6$ ), while haplotype 6 has an interaction ( $hselg=6$ ) with current smoking status ( $zgsel=11$ ). In addition, the haplotype frequencies are allowed to depend on current smoking status ( $zzsel=11$ ).