

HAPPY: A SAS macro for estimating haplotype and haplotype-environment interaction odds ratios

Peter Kraft, Rong Chen
Last updated 27 July 2004

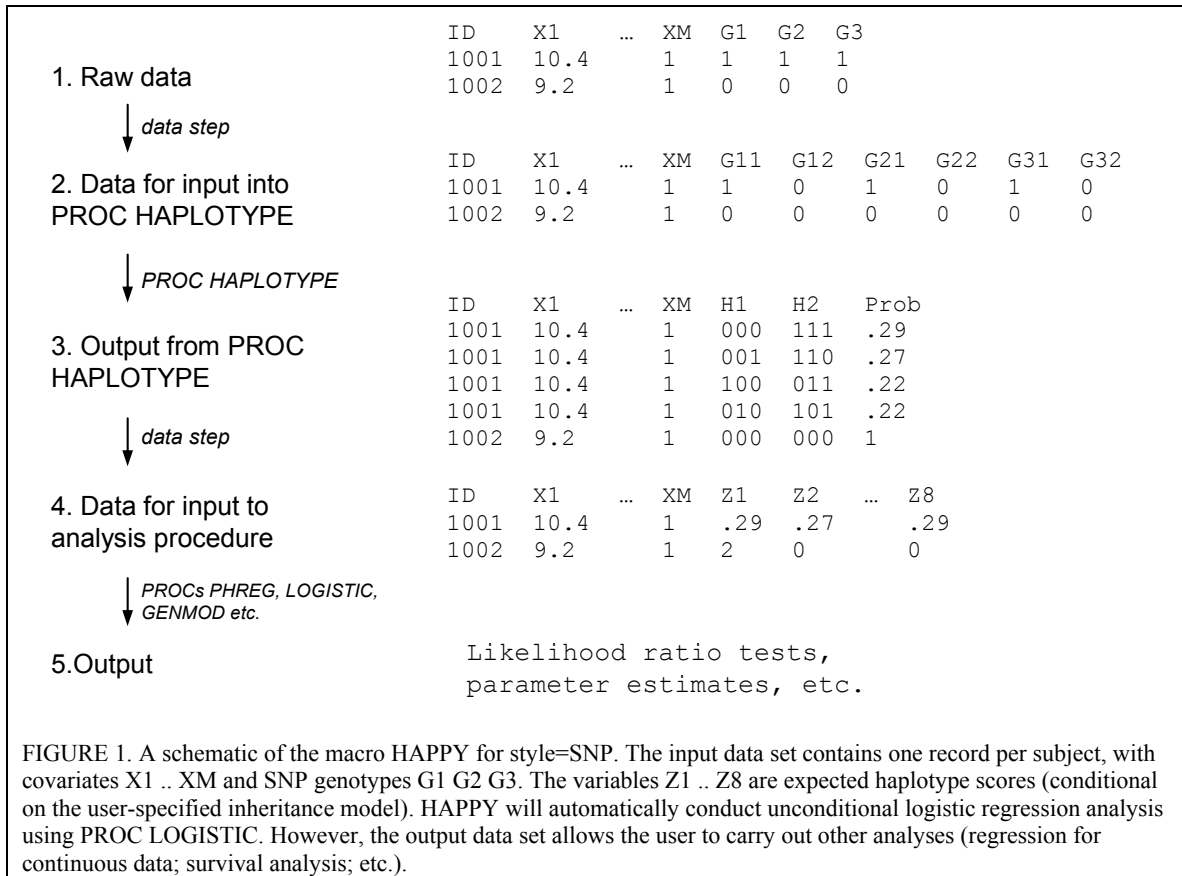
DESCRIPTION

HAPPY estimates haplotype-specific odds ratios from genotype data on unrelated cases and controls using unconditional logistic regression. It can adjust for the main effects of relevant covariates and estimate stratum-specific haplotype effects. Aside from confidence intervals around individual odds ratio estimates, HAPPY calculates omnibus tests of haplotype association and haplotype-environment interactions. HAPPY uses the "expectation substitution" approach [1,2], which treats expected haplotype scores (calculated under a user-specified inheritance model) as observed covariates in a standard unconditional logistic analysis. The macro outputs these expected scores to an auxiliary data set; the scores can be then be used in customized analyses.

SYNTAX

```
%happy(indsn=in,  
       keep=,  
       caco=,  
       style=SNP,  
       outdadd= addscore,  
       outreces=null,  
       outdomnant=null,  
       outcodomnant=null,  
       thresh=.0001,  
       range=.01,  
       dir=,  
       covar=,  
       stratum=,  
       fmt=,  
       covar1=,  
       cutoff=.05,  
       outdsn1=happlotb,  
       outdsn2=origfreq)
```

indsn	Name of input data set
keep	List of input SNPs, coded as 0, 1 or 2 = number of minor alleles: REQUIRED
caco	Dichotomous outcome variable; if absent, only expected haplotype scores are output
style	Genotype input style; SNP is counts of alleles, MAR is individuals alleles
outdadd	Name of output data set containing expected haplotype scores under <i>additive</i> model
outreces	Name of output data set containing expected haplotype scores under <i>recessive</i> model
outdomnant	Name of output data set containing expected haplotype scores under <i>dominant</i> model
outcodomnant	Name of output data set containing expected haplotype scores under <i>codominant</i> model
thresh	Haplotypes with estimated frequencies below this threshold are discarded
range	Haplotypes below this threshold are pooled for score calculations
dir	directory where results are to be saved
covar	Covariates in unstratified analysis: entered as they would appear in <code>model</code> statement
stratum	Name of variable to be used for stratum-specific haplotype odds ratio estimates
fmt	Format for stratum to be used in output tables
covar1	Covariates in stratified analysis: entered as they would appear in <code>model</code> statement (<i>sans</i> stratum)
cutoff	Haplotypes with frequencies below this frequency are excluded from logistic regression
outdsn1	Name of data set containing raw per-subject haplotype counts (before pooling rare haps)
outdsn2	Name of output data set containing haplotypes and their frequencies



INPUT

Genotypes can be input in two ways: first, as counts of alleles (style=SNP), in which case there will be one input variable per SNP; second, as pairs of alleles (style=MAR), in which case there will be two input variables per SNP. If style=MAR, the data step from 1 to 2 shown in Figure 1 is bypassed, as the data are already in the format required by PROC HAPLOTYPE. Missing genotypes/alleles should be coded using the appropriate SAS convention: . for numeric variables, " " (space) for character.

Note that happy isn't smart enough to pass numbered variable lists, e.g. G1-G4. You'll have to list the variables explicitly: G1 G2 G3 G4.

STATISTICAL DETAILS

If `caco` (a case/control indicator) is specified, HAPPY fits a logistic model for the probability of failure as a function of observed covariates \mathbf{X} and (unobserved) haplotypes:

$$\text{logit } \Pr(D|\mathbf{X},\mathbf{H}) = \alpha + \beta'\mathbf{Z}(\mathbf{H}) + \gamma'\mathbf{X}, \quad (1)$$

where $\mathbf{Z}(\mathbf{H})$ is some coding of the haplotype pair \mathbf{H} corresponding to an assumed multiallelic inheritance model [3]. The default model is an additive model:

$$\mathbf{Z}(\mathbf{H}) = (\#H_2, \dots, \#H_K)',$$

where $H_1 \dots H_{K-1}$ are the $K-1$ haplotypes in the population with frequency above `range`, ordered from most to least frequent, and H_K is a category combining all haplotypes with frequency less than `range`. The number of H_1 haplotypes is omitted from the vector $\mathbf{Z}(\mathbf{H})$ for identifiability purposes. Thus the parameter β_j should be interpreted as the log odds ratio for carrying haplotype H_j relative to haplotype H_1 . One important advantage of the additive model is that omnibus likelihood ratio tests of haplotype effect—tests comparing model (1) to the model with $\boldsymbol{\beta}=\mathbf{0}$ —do not depend on the choice of reference.

The other models available in HAPPY include recessive:

$$\mathbf{Z}(\mathbf{H}) = (\mathbf{I}(\#H_2=2), \dots, \mathbf{I}(\#H_K=2))'$$

dominant:

$$\mathbf{Z}(\mathbf{H}) = (\mathbf{I}(\#H_2>0), \dots, \mathbf{I}(\#H_K>0))'$$

and co-dominant:

$$\mathbf{Z}(\mathbf{H}) = (\mathbf{I}(\#H_2=1), \mathbf{I}(\#H_2=2), \dots, \mathbf{I}(\#H_K>0))'$$

where $\mathbf{I}(\cdot)$ is the indicator function. (The co-dominant model is not recommended as there will be few homozygote carriers for most haplotypes. Currently HAPPY calculates the expected scores for these models but only fits the logistic regression for the additive model.)

To account for ambiguous haplotypes, HAPPY uses `PROC HAPLOTYPE` to calculate expected haplotype scores conditional on observed genotypes: $\mathbf{Z}^*(\mathbf{G}) = E_{\mathbf{H}|\mathbf{G}} \mathbf{Z}(\mathbf{H})$. It then fits (1) with $\mathbf{Z}^*(\mathbf{G})$ replacing $\mathbf{Z}(\mathbf{H})$. Parameter estimates retain their interpretation as haplotype (-pair) specific log odds ratios [1,2,4]. The expected scores \mathbf{Z}^* for each individual are output to the data sets `outadd`, etc.

Omnibus tests of haplotype effect are calculated using a likelihood ratio test, comparing model (1) to the model without \mathbf{Z} : $\text{logit Pr}(D|\mathbf{X},\mathbf{H}) = \alpha + \boldsymbol{\gamma}'\mathbf{X}$. Parameter-specific confidence intervals are calculated and reported in the output and in `main*.html`.

If `stratum` is specified, model (1) is fit by `stratum`, yielding stratum-specific haplotype odds ratios. Haplotype \times stratum interactions are tested by comparing the model

$$\text{logit Pr}(D|\mathbf{X},\mathbf{H},\mathbf{S}) = \alpha + \boldsymbol{\beta}'\mathbf{Z}(\mathbf{H}) + \boldsymbol{\beta}_2'\mathbf{Z}(\mathbf{H}) \mathbf{I}(\mathbf{S}=\mathbf{S}_2) + \dots + \boldsymbol{\beta}_L'\mathbf{Z}(\mathbf{H}) \mathbf{I}(\mathbf{S}=\mathbf{S}_L) + \boldsymbol{\gamma}'\mathbf{X}$$

to model (1). Results are reported in the output and in `strata*.html`.

If `caco` is not specified, the expected scores are calculated and output; the logistic regression is not run.

EXAMPLE

The sample data consist of 216 cancer cases and 657 controls, genotyped at four SNPs in a candidate gene. Measurements of smoking packyears (`packyrs`, a covariate) and body mass index less or greater than 30 (`bmistrat`, a 0/1 stratifying variable) are also available for each subject. The output presented below was generated by the following code:

```
%happy(caco=cacotype,
        keep=SNP1 SNP2 SNP3 SNP4,
        outadd=addscore,
        covar=packyr bmistrat,
        covarl=packyr,
        stratum=bmistrat,
        cutoff=.05);
```

<i>Descriptive Statistics and main effect of haplotypes in additive model</i>						
	case		control		Pool	
	frequency	percent	frequency	percent	frequency	percent
z1 1-0-0-1	152.06	35.19%	446.92	34.01%	598.98	34.30%
z2 1-0-0-0	74.28	17.19%	242.48	18.45%	316.76	18.14%
z3 0-1-1-1	71.74	16.60%	227.42	17.30%	299.17	17.13%
z4 0-0-0-1	49.70	11.50%	130.08	9.89%	179.78	10.29%
z5 1-1-0-1	34.35	7.95%	91.31	6.94%	125.66	7.19%
z6 1-1-0-0	20.36	4.71%	86.67	6.59%	107.03	6.13%
z7 0-0-1-1	26.11	6.04%	77.41	5.89%	103.52	5.92%
z8 <0.01	3.39	0.78%	11.70	0.89%	15.09	0.86%

*The frequency of z8 <0.01
<= .05, it will not be included in future analysis*

FIGURE2. Output from HAPPY to file main*.html. This table lists the haplotypes and tabulates their frequencies in cases and controls. Some counts are fractional because of ambiguous genotypes. Note that because the cumulative frequency of rare haplotypes (Z8, those with freq <.01) is less than the `cutoff` of 5%, Z8 is excluded from subsequent analyses.

Descriptive Statistics and main effect of haplotypes in additive model

Haplotype	Odds Ratio	Lower	Upper
bmistrat	2.055	1.427	2.961
packyr	0.990	0.982	0.998
z2	0.934	0.638	1.367
z3	0.924	0.654	1.304
z4	1.158	0.785	1.708
z5	1.330	0.771	2.296
z6	0.575	0.305	1.083
z7	1.072	0.630	1.823

Model with haplotypes : -2 logL= 951.08
Model without haplotypes : -2 logL= 957.06
Degree of freedom is : 6
P of LRT test is : 0.4258

Unconditinal logistical model adjusted for packyr bmistrat

FIGURE3. More output to main*.html: haplotype-specific odds ratios and the omnibus test for haplotype association.

bmistrat 0							
	case		control		Pool		
	frequency	percent	frequency	percent	frequency	percent	
z1 1-0-0-1	105.39	33.99%	374.74	34.19%	480.13	34.14%	
z2 1-0-0-0	54.57	17.60%	201.06	18.34%	255.63	18.18%	
z3 0-1-1-1	52.88	17.05%	184.81	16.86%	237.69	16.90%	
z4 0-0-0-1	38.20	12.32%	113.06	10.31%	151.26	10.75%	
z5 1-1-0-1	22.93	7.39%	78.48	7.16%	101.41	7.21%	
z6 1-1-0-0	14.21	4.58%	68.37	6.23%	82.57	5.87%	
z7 0-0-1-1	21.02	6.78%	64.80	5.91%	85.82	6.10%	
z8 <0.01	0.80	0.25%	10.68	0.97%	11.48	0.81%	
subtotal	310.00	100.00%	1095.99	100.00%	1405.99	100.00%	

bmistrat 1							
	case		control		Pool		
	frequency	percent	frequency	percent	frequency	percent	
z1 1-0-0-1	46.67	38.25%	72.18	33.10%	118.85	34.95%	
z2 1-0-0-0	19.71	16.15%	41.43	19.00%	61.13	17.98%	
z3 0-1-1-1	18.86	15.46%	42.62	19.54%	61.48	18.08%	
z4 0-0-0-1	11.50	9.42%	17.02	7.80%	28.52	8.38%	
z5 1-1-0-1	11.42	9.36%	12.83	5.88%	24.25	7.13%	
z6 1-1-0-0	6.15	5.04%	18.30	8.39%	24.46	7.19%	
z7 0-0-1-1	5.09	4.17%	12.61	5.78%	17.69	5.20%	
z8 <0.01	2.60	2.12%	1.02	0.46%	3.61	1.06%	
subtotal	122.00	100.00%	218.00	100.00%	340.00	100.00%	

FIGURE 4. Descriptive statistics for stratified analysis, output to strat*.html.

Covariate	Sub Group	Haplotype	Odds Ratio	Lower	Upper	
bmistrat	0	z2	1.030	0.668	1.590	
		z3	1.028	0.689	1.535	
		z4	1.208	0.775	1.883	
		z5	1.213	0.646	2.277	
		z6	0.706	0.340	1.468	
		z7	1.262	0.699	2.279	
		packyr	0.987	0.978	0.997	
		z2	0.704	0.311	1.593	
		z3	0.712	0.357	1.421	
		z4	1.019	0.446	2.331	
1		z5	1.689	0.504	5.655	
		z6	0.355	0.103	1.220	
		z7	0.613	0.183	2.060	
		packyr	0.995	0.979	1.010	
		Model without interaction: -2 log L= 951.08 Model with interaction : -2 log L= 946.88 Degree of freedom is : 6 P for interaction is : 0.65				

Unconditional logistical model adjust by packyr

FIGURE 5. Stratum-specific haplotype odds ratios and omnibus test of haplotype × stratum interaction.

id	case	packyr	bmistrat	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
6661	control	53.66	0	0.6172	0.3828	0.0000	0.0000	0.3828	0.6172	0.0000	0.0000
6662	case	47.17	0	0.0306	0.9694	0.0000	0.9694	0.0000	0.0000	0.0000	0.0306
6663	case	0.00	0	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
6664	control	0.00	0	0.9852	0.0000	0.0000	0.0148	0.0000	0.0000	0.9852	0.0148
6665	control	32.30	0	0.0000	0.0000	0.9999	0.0000	0.5398	0.4601	0.0000	0.0000
6666	control	50.70	0	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
6667	control	15.60	0	0.0000	0.0000	0.9888	0.9888	0.0000	0.0000	0.0112	0.0112
6668	control	0.00	0	0.0306	0.9694	0.0000	0.9694	0.0000	0.0000	0.0000	0.0306
6669	control	34.80	0	0.6542	0.3459	0.0000	0.9836	0.0000	0.0000	0.0000	0.0165
6670	case	0.00	0	0.6542	0.3459	0.0000	0.9836	0.0000	0.0000	0.0000	0.0165
6671	control	23.57	0	0.9317	0.0000	0.9317	0.0005	0.0676	0.0000	0.0676	0.0009
6672	control	22.80	0	0.9317	0.0000	0.9317	0.0005	0.0676	0.0000	0.0676	0.0009
6673	control	0.00	0	0.0001	0.8951	0.8951	0.0000	0.0000	0.1047	0.1047	0.0001
6674	case	0.00	0	0.0000	0.0000	0.9888	0.9888	0.0000	0.0000	0.0112	0.0112
6675	control	0.00	0	0.0000	0.0000	0.9888	0.9888	0.0000	0.0000	0.0112	0.0112
6676	control	0.00	0	0.0000	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6677	control	54.10	0	0.0000	0.0000	0.9999	0.0000	0.0001	0.9999	0.0000	0.0001
6678	control	67.77	0	0.9317	0.0000	0.9317	0.0005	0.0676	0.0000	0.0676	0.0009

FIGURE 6. Partial listing of the addscore data set, containing the expected haplotype scores (additive model).

REFERENCES

1. Zaykin et al. (2002) *Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals*. Hum Hered 53:79-91.
2. Stram et al. (2003) *Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals*. Hum Hered 55:179-190.
3. Schaid (1996) *General score tests for associations of genetic markers with disease using cases and their parents*. Genet Epidemiol 13:423-449.
4. Kraft (submitted) *Accounting for haplotype uncertainty in association studies: A comparison of simple and flexible techniques*.