Testing and Modeling Genotypic Disequilibria

John Maindonald
Centre for Mathematics and Its Applications, Australian National University, Canberra, Australia

Keywords. population genetics, Hardy-Weinberg, genotype, allele, genotypic disequilibrium

1 Introduction

In a diploid, sexually reproducing species, at a locus where there are two alleles A and a, the possible genotypes are AA, Aa and aa. In a population of size N, with p the frequency of the A allele and q the frequency of the a allele, the expected numbers under Hardy-Weinberg equilibrium are $NP_{AA} = Np^2$ for the AA genotype, $NP_{Aa} = 2Npq$ for the Aa genotype, and $NP_{aa} = Nq^2$ for the aa genotype. Writing $m = \log(Np^2)$ and $\log(q/p) = m_a$, the logarithms of the frequencies may be written:

$$\begin{aligned} \log(Np^2) &= m & (1)\\ \log(2Npq) &= m + \log(2) + m_a & (2)\\ \log(Nq^2) &= m + 2m_a & (3) \end{aligned}$$

Thus the model is loglinear, and can be fitted as a generalized linear model with poisson error and offset log(2) for the heterozygote. For example:

```
> obs <- c(AA=147, Aa=78, aa=17)
> oset <- c(0, log(2), 0)
> ma <- c(0,1,2)
> hw.glm <- glm(obs ~ ma, family=poisson, offset=oset)
> summary(hw.glm)

Call:
glm(formula = obs ~ ma, family = poisson, offset = oset)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.96256    0.08137   60.99   <2e-16
ma          -1.20039    0.10778  -11.14   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 149.3527  on 2  degrees of freedom
Residual deviance:   2.0436  on 1  degrees of freedom
AIC: 23.751

Number of Fisher Scoring iterations: 4
```

The function `hwde()` may also be used to fit this model, at the same time introducing a further "disequilibrium" term. The default output is the analysis of deviance table.

```
> hwdat <- data.frame(Observed=c(147,78,17), locus1=c("AA","Aa","aa"))
```

Now call the function.

```
> library(hwde)
> hwde(data=hwdat)

[1] "Analysis of Deviance Table"
    Resid. Df Resid. Dev Df Deviance
1           2    149.353
+a          1      2.044  1  147.309
+aa         0      0.000  1    2.044
```

The disequilibrium term has the form

$$m_{aa} = \log \frac{4P_{AA}P_{aa}}{P_{Aa}^2}$$

Notice that the parameters m_a and m_{aa} have been abbreviated, in the computer output, to a and aa respectively. The parameter m models the reference or baseline level, and is estimated by the intercept term.

To obtain estimates of parameters, including the disequilibrium parameter m_{aa}, do the following:

```
> data.df <- hwde(data = hwdat)$data.df

[1] "Analysis of Deviance Table"
    Resid. Df Resid. Dev Df Deviance
1           2    149.353
+a          1      2.044  1  147.309
+aa         0      0.000  1    2.044

> names(data.df)

[1] "obs"  "data" "a"    "aa"   "oset"

> summary(glm(obs ~ a + aa, offset=oset, family=poisson, data=data.df))$coef

             Estimate Std. Error  z value     Pr(>|z|)
(Intercept) 1.8332133  0.2425356 7.558532 4.076429e-14
a           0.5234955  0.2676640 1.955793 5.048954e-02
aa          1.1102283  0.3419186 3.247055 1.166060e-03
```

Note again that the intercept estimates m, and that aa is the additive version of the disequilibrium parameter.

We leave till later detailed information on the use of `hwde()`, including details on how to obtain fitted values and residuals.

1.1 Several different populations

If there several different populations, there must be a parameter (by default assumed to have the name `Population`), that accounts for different population sizes. In the code, this translates to a main effect `gp` in the log-linear model. Additionally, there may be different values for m_a and m_{aa} in the different populations.

A second locus requires the parameters m_b and m_{bb} for that locus. Additionally, parameters may be required that model quantities that, in the loglinear model, have the role of interactions between the two loci. Huttley and Wilson (2000) introduce the multiplicative versions of the following parameters:

s_{ab}, the "sum of digenic disequilibria for the total sample"

q_{ab}, the "product of digenic disequilibria for the total sample"

m_{aab} and m_{abb}, which are "trigenic disequilibria terms for the total sample"

In the usual case where phase for double heterozygotes is unknown and only nine genotypic classes can be distinguished, no degrees of freedom remain that might be used to estimate a quadrigenic disequilibrium term.

As noted above, the formulae in Huttley and Wilson (2000) give the multiplicative equivalents of these terms, using upper case letters. The additive versions used here (e.g., they have M_A where I have $m_a = \log(M_A)$) use the corresponding lower case letters. Note however that in the second column on p.2131 of Huttley and Wilson, in the equations for $\ln P_{Ab}^{AB}$ and $\ln P_{aB}^{AB}$, $\ln Q_{AB}^2$ should be, in each case, $\ln Q_{AB}$. The equations are given correctly in Weir and Wilson (1986), though with slight changes of notation. See also Weir (1996).

The function allows an arbitrary number of loci. Terms s_{ab}, q_{ab}, m_{abb} and m_{aab} are fitted for every pair of loci. Terms that correspond to second (or, with > 3 loci, higher order) interactions contribute, in the present version of the code, to the residual. Try

```
> hwde(data=mendelABC, loci=c("seedshape","cotylcolor","coatcolor"))
```

2 Details of Use of `hwde()`

First recall the simple example that was described above. The data were entered, from the keyboard, into a data frame `hwdat` that had the form:

```
Observed locus1
147      AA
78       Aa
17       aa
```

The coding used in the column headed `locus1` can be varied; any two characters may be used for the alleles. With the column names that are shown, the corresponding parameter settings for the function `hwde()` can be left at their defaults.

An alternative is to enter the data, exactly as displayed above (though the spacing is immaterial), into a file. If the file is called **hw.txt** and is placed in the working directory, then it can be read in with:

```
> hwdat <- read.table("hw.txt", header=TRUE)
```

If there is a second locus, the default name is `locus2`. The default name for any third locus is `locus3`, etc. Where there is a column that has codes for different populations, the default name is `Population`.

Example – two populations and two loci

With this introduction, we move directly to data, with two populations and two loci, that are suited to fitting all the parameters that the function currently allows, i.e., m_{aa}, m_{bb}, m_{cc}, s_{ab}, s_{ac}, s_{bc}, q_{ab}, q_{ac}, q_{bc}, m_{abb}, m_{acc}, m_{bcc}, m_{aab}, m_{aac}, m_{bbc}.

Data (Mourant et al, 1976) are:

```
Population locus1 locus2 Observed
  Indian     MM     SS       91
  Indian     MM     Ss      147
  Indian     MM     ss       85
  Indian     MN     SS       32
  Indian     MN     Ss       78
```

```
Indian     MN     ss       75
Indian     NN     SS        5
Indian     NN     Ss       17
Indian     NN     ss        7
 Irish     MM     SS      121
 Irish     MM     Ss      248
 Irish     MM     ss      164
 Irish     MN     SS       53
 Irish     MN     Ss      422
 Irish     MN     ss      375
 Irish     NN     SS        9
 Irish     NN     Ss       65
 Irish     NN     ss      241
```

Assuming that this is stored in a file **IndianIrish.txt**, we can read in the data and do the analysis thus:

```
> IndianIrish <- read.table("IndianIrish.txt", header=TRUE)
```

```
> hwde(data=IndianIrish)
```

```
[1] "Analysis of Deviance Table"
              Resid. Df Resid. Dev Df Deviance
1                    17    1724.07
+gp                  16    1090.41  1   633.66
+(a+b)               14     486.72  2   603.69
+(aa+bb)             12     480.31  2     6.41
+sab                 11     463.76  1    16.55
+qab                 10     218.42  1   245.34
+(abb+aab)            8     217.15  2     1.28
+gp:(a+b)             6      37.94  2   179.21
+gp:(aa+bb)           4      35.46  2     2.48
+gp:sab               3      26.29  1     9.16
+gp:qab               2       5.94  1    20.36
+gp:(abb+aab)         0       0.00  2     5.94
```

The above is the compact default output, in which terms that are at the same level of a hierarchy are grouped. For a first pass through the data, this may be the preferred output. A form of output in which each term correspods to a single degree of freedom is available by using the parameter setting `group.terms=FALSE`, i.e.,

```
> hwde(data=IndianIrish, group.terms=FALSE)
```

difference from the last previous Residual Deviance term that is marked with an `r` (= reference) as the first character in the row in which it appears.

The estimates of parameters in the maximal (or, with appropriate modification, any other) model can be extracted thus:

```
> II.hwde <- hwde(data = mendelABC, loci = c("seedshape", "cotylcolor",
+                "coatcolor"), keep.models=T)
```

```
[1] "Analysis of Deviance Table"
                           Resid. Df Resid. Dev Df Deviance
1                                 26    15.3272
+(a+b+c)                          23    14.5604  3   0.7669
+(aa+bb+cc)                       20    13.5445  3   1.0158
+(sab+sac+sbc)                    17     9.7381  3   3.8064
+(qab+qac+qbc)                    14     7.4493  3   2.2889
+(abb+acc+bcc+aab+aac+bbc)         8     5.1362  6   2.3130
```

```
> models <- II.hwde$models
> maxmodel <- models[[length(models)]]
> summary(maxmodel)$coef
```

```
                Estimate Std. Error     z value     Pr(>|z|)
(Intercept)  2.119196819  0.2546858  8.32082671 8.735188e-17
a            0.015218967  0.2645433  0.05752922 9.541236e-01
b            0.245726216  0.2659182  0.92406689 3.554515e-01
c            0.143067467  0.2650647  0.53974548 5.893726e-01
aa           0.021854471  0.4304371  0.05077273 9.595066e-01
bb          -0.177991441  0.4172846 -0.42654686 6.697094e-01
cc          -0.157633137  0.4284390 -0.36792438 7.129296e-01
sab          0.075610484  0.2422532  0.31211346 7.549543e-01
sac         -0.031663422  0.2357458 -0.13431172 8.931561e-01
sbc         -0.247841900  0.2445746 -1.01335891 3.108888e-01
qab         -0.245328249  0.3637890 -0.67436970 5.000763e-01
qac          0.186752307  0.3524097  0.52992949 5.961608e-01
qbc         -0.105419665  0.3673842 -0.28694664 7.741532e-01
abb          0.020721843  0.2250188  0.09208938 9.266270e-01
acc         -0.083962005  0.2255440 -0.37226444 7.096960e-01
bcc         -0.117142586  0.2300915 -0.50911316 6.106729e-01
aab          0.230215504  0.2296021  1.00267155 3.160194e-01
aac         -0.198836824  0.2259632 -0.87995213 3.788852e-01
bbc         -0.008483307  0.2267579 -0.03741130 9.701571e-01
```

3 Obtaining Additional Output

By default, the function returns (invisibly) a list with two elements. The first holds the analysis of variance table. The second holds the data and contrast terms that are required for fitting tbe various models. For example:

```
> hwdat.hw <- hwde(data=hwdat)
```

```
[1] "Analysis of Deviance Table"
    Resid. Df Resid. Dev Df Deviance
1           2    149.353
+a          1      2.044  1  147.309
+aa         0      0.000  1    2.044
```

```
> names(hwdat)
```

```
[1] "Observed" "locus1"
```

```
> hwdat.hw$data.df
```

```
   obs data a aa oset
AA 147   AA 2  1    1
Aa  78   Aa 1  0    2
aa  17   aa 0  0    1
```

The following illustrates the direct use of the information in `hwdat.hw$data.df`, giving the user complete control over the models that are fitted.

```
> data.df <- hwdat.hw$data.df
> model1 <- glm(obs ~ a, family=poisson, data=data.df, offset=log(oset))
> model2 <- glm(obs ~ a+aa, family=poisson, data=data.df, offset=log(oset))
> model1
```

```
Call:  glm(formula = obs ~ a, family = poisson, data = data.df, offset = log(oset))

Coefficients:
(Intercept)            a
      2.562        1.200

Degrees of Freedom: 2 Total (i.e. Null);  1 Residual
Null Deviance:            149.4
Residual Deviance: 2.044         AIC: 23.75
```

Here is the output data frame for the `IndianIrish` data.

```
> II.hw <- hwde(data=IndianIrish, aovtable.print=FALSE)
> dataII.df <- II.hw$data.df
> dataII.df
```

	obs	gp	locus1	locus2	a	b	aa	bb	sab	qab	abb	aab	oset
1	91	Indian	MM	SS	2	2	1	1	0	2	2	2	1
2	147	Indian	MM	Ss	2	1	1	0	0	1	0	1	2
3	85	Indian	MM	ss	2	0	1	0	0	0	0	0	1
4	32	Indian	MN	SS	1	2	0	1	0	1	1	0	2
5	78	Indian	MN	Ss	1	1	0	0	1	0	0	0	4
6	75	Indian	MN	ss	1	0	0	0	0	0	0	0	2
7	5	Indian	NN	SS	0	2	0	1	0	0	0	0	1
8	17	Indian	NN	Ss	0	1	0	0	0	0	0	0	2
9	7	Indian	NN	ss	0	0	0	0	0	0	0	0	1
10	121	Irish	MM	SS	2	2	1	1	0	2	2	2	1
11	248	Irish	MM	Ss	2	1	1	0	0	1	0	1	2
12	164	Irish	MM	ss	2	0	1	0	0	0	0	0	1
13	53	Irish	MN	SS	1	2	0	1	0	1	1	0	2
14	422	Irish	MN	Ss	1	1	0	0	1	0	0	0	4
15	375	Irish	MN	ss	1	0	0	0	0	0	0	0	2
16	9	Irish	NN	SS	0	2	0	1	0	0	0	0	1
17	65	Irish	NN	Ss	0	1	0	0	0	0	0	0	2
18	241	Irish	NN	ss	0	0	0	0	0	0	0	0	1

The user can now fit any sequence of models that may be required. For example, the user may wish to a sequence of models that is different from the sequence fitted by `hwde()`.

Further control is available by supplying values for the parameters `termlist` and `refmodel`. For example, the default action with the data frame `hwdat` is equivalent to:

```
> hwde(termlist=c("+a","+aa"), refmodel=c(1,2), data=hwdat)

[1] "Analysis of Deviance Table"
    Resid. Df Resid. Dev Df Deviance
1           2    149.353
+a          1      2.044  1  147.309
+aa         0      0.000  1    2.044
```

In `refmodel`, 1 refers to the model that has constant term only.

The first six models can be fitted to the data frame `IndianIrish` by setting:

```
> hwde(termlist=c("+gp","+a","+b","+a+b","+aa"), refmodel=c(1,2,2,2,5),
+      data=IndianIrish)
```

```
[1] "Analysis of Deviance Table"
     Resid. Df Resid. Dev Df Deviance
1           17    1724.07
+gp         16    1090.41  1   633.66
+a          15     853.73  1   236.68
+b          15     723.40  1   367.02
+a+b        14     486.72  2   603.69
+aa         13     485.59  1     1.13
```

Extraction of the sequence of fitted models

A further possibility, with the parameter setting `keep.models=TRUE`, is to include the full sequence of models that have been fitted in the list that is returned by the function. For example:

```
> hwdat.hw <- hwde(data=hwdat, keep.models=TRUE)

[1] "Analysis of Deviance Table"
    Resid. Df Resid. Dev Df Deviance
1           2    149.353
+a          1      2.044  1  147.309
+aa         0      0.000  1    2.044

> hwdat.hw$models[[2]]          # The Hardy-Weinberg model

Call:  glm(formula = obs ~ a, family = poisson, data = data.df, offset = log(oset))

Coefficients:
(Intercept)            a
      2.562        1.200

Degrees of Freedom: 2 Total (i.e. Null);  1 Residual
Null Deviance:            149.4
Residual Deviance: 2.044         AIC: 23.75

> fitted(hwdat.hw$models[[2]])

       AA        Aa        aa
142.95868  86.08264  12.95868
```

The function `fitted()` can be replaced by any of the functions (`coef()`, `resid()`, `predict()`, etc.) that are available for use with a `glm` model object. Note that there are several different choices of residuals, with deviance residuals as the default. For the `IndianIrish` data there are, with the parameter setting `group.terms=FALSE`, 24 models from which to choose. Choose carefully!

4 Exact Hardy-Weinberg Test

The function `hwexact()`, supplied by Randall Johnson, does an exact test for Hardy-Weinberg equilibrium, conditional on the observed relative numbers of the two alleles. The only case implemented is for a single population and single locus. The algorithm is described in Wigginton et al (2005).

5 References

Huttley, G.A. and Wilson, S.R. 2000. Testing for concordant equilibria between population samples. Genetics 156: 2127-2135.

Mourant, A.E., Kopec, A.C. and Domaniewska-Sobczak, K. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms.* Oxford University Press.

Weir, B.S. 1996. *Genetic Data Analysis II.* Sinauer.

Weir, B.S. and Wilson, S.R. 1986. Log-linear models for linked loci. Biometrics 42:665-670.

Wigginton, J.E., Cutler, D.J. and Abecasis, G.R. 2000. A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76: 887-893.