Practical exercise on (recorded) population stratification

David Clayton and Joanna Howson

The data for this exercise concern the population–based case–control study of the association between juvenile rheumatoid arthritis (JRA) and the NRAMP1 gene (and a closely linked marker, D2S1471) which we used in the preceding exercise. A feature of these data which we have not used so far is that they were collected from two ethnic groups. However, ethnic origin was recorded in the study.

As before, if you wish to use the menu system to access `genassoc` commands, you will need to activate it:

```
. gamenu
```

1. Read in the data either using the **Read Stata dataset** option from the **Data management** sub menu, or the command line

   ```
   . use jranramp
   ```

2. If you use `describe` you will see that there is a variable named `popgrp`. This identifies whether subjects are of Russian or Latvian ethnic origin. To see how this variable is coded,

   ```
   . codebook popgrp
   ```

   and to see how cases and controls are distributed between ethnic groups:

   ```
   . table caco popgrp
   ```

   Is this a *matched* case–control study? If it were unobserved, would this ethnic stratification within the study present problems for its interpretation?

1 Analysis at chromosome level

1. As before, we start by analysing at the chromosome level. We must reshape the data file so that each line refers to a chromosome rather than a person. We can do this either using the **Genasoc** $\rightarrow$ **Data Management** $\rightarrow$ **Reshape data file** menu, or the command line

   ```
   .  greshape, id(id) gen(chr)
   ```

   You may also wish to derive the grouped version of the tandem repeat marker:

```
. egen d2simp = cut(d2s1471), at(1,8,20)
. grprare d2s1471, gen(g_d2s1471)
```

2. First you should look at allele frequencies within ethnic groups:

   ```
   . table caco nramp popgrp
   ```

   Is there obvious confounding? You should also look at the tandem repeat marker, but this is rather harder:

   ```
   . table caco d2s1471 popgrp
   . table caco d2simp popgrp
   ```

3. As before, we will find it useful to set up indicator variables for each allele of `d2s1471`:

   ```
   . tab d2s1471, gen(Allele_)
   ```

   Then, to look at allele 1,in Latvians for example,

   ```
   . tab caco Allele_1 if popgrp==1, chi2
   ```

4. Controlling for ethnic stratification can be carried out by the Mantel–Haenszel method, using the `mhodds` commands. The command

   ```
   . mhodds caco nramp, co(2,3) by(popgrp)
   ```

   calculates odds ratios separately in each ethnic group, a test for "effect modification" (variation of the odds ratio between ethnic groups), and a "pooled" estimate under the assumption that the odds ratio is constant. A shorter form is

   ```
   . mhodds caco nramp popgrp, co(2,3)
   ```

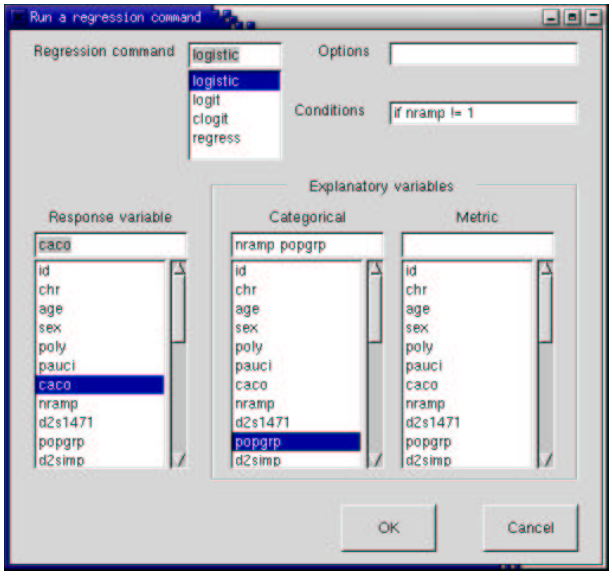
   which only calculates the pooled odds ratio estimate. You might like to compare the odds ratio with and without control for ethnic stratification.

5. Association between disease and the D2S1471 marker could also be controlled for stratification in the same manner, but the Mantel–Haenszel method is limited to pair-wise comparisons of alleles and there are a very large number of possibilities. You could test for an effect of allele 1 with

   ```
   . mhodds caco Allele_1  popgrp
   ```

   You might like to try some of the other (more common) alleles.

6. We can also control for ethnic group by using logistic regression. We relate the case:control odds to two categorical variables: *allele* of the genetic locus and *ethic group*. In the NRAMP case, this can be done within the **Fit** menu option by completing the fields as follows:



The appropriate command line to do this (omitting the two cases with the rare 1 allele) is,

```
. xi:logistic caco i.nramp i.popgrp if nramp!=1
```

You will find that the odds ratio for `nramp` is almost the same as that obtained using the Mantel–Haenszel method. As before, the `xi:` and i. machinery in the above command is there to signal the fact that `nramp` and `popgrp` are to be treated as *categorical* variables.

7. In the case of D2S1471 marker, we have already generated the indicator variables and we could carry out the stratified test by once again selecting **Fit** from the menu options. Choose `logistic` as the regression command, `caco` as the response variable, `popgrp` as the *categorical* variable and `Allele_*` as the *Metric* variable. Having clicked OK, select **Test(drop)** from the menus, with `Allele_*` as the coefficients, and `Wald` as the type of test. Which is equivalent to the command lines,

```
. xi:logistic caco Allele_* i.popgrp
. testparm Allele_*
```

2 Analysis at person level

Analysis at the chromosome level depends on the assumption of Hardy–Weinberg equilibrium in the population. As before, we can avoid this assumption by appropriate analysis at the person level. We will start by going back to the original data. So clear the dataset we have been using with,

```
. clear
```

Read in the data again, either using the **Read Stata dataset** option from the **Data management** sub menu, or the command line

```
. use jranramp
```

1. Use `gtab` to generate indicate variables which count the number of occurrences of each allele at the NRAMP locus. This can be done from the **GenAssoc** → **Tabulate** → **Allele frequencies** menu or with the command line:

   ```
   . gtab nramp*, gen(NR_)
   ```

   Compare the distribution of NRAMP genotypes in cases and controls, and compute a chi–squared test on 2 df, separately for the two ethnic groups: as follows:

   ```
   . tab caco NR_2 if popgrp==1, chi2
   . tab caco NR_2 if popgrp==2, chi2
   ```

   What three groups of genotypes are compared in these tests?

2. Each ethnic group alone is a small study. We can use Mantel–Haenszel methods to combine the evidence across ethnic groups. To calculate odds ratios between genotypes:

   ```
   . mhodds caco NR_2, co(0,1) by(popgrp)
   . mhodds caco NR_2, co(1,2) by(popgrp)
   ```

3. The 1 df test for association which is derived from the model in which the odds ratio of 1 vs 0 copies of an allele is the same as that for 2 vs 1 copy is known as the Cochran-Armitage test. The stratified version is sometimes called the *Mantel extension* test:

   ```
   . mhodds caco NR_2 popgrp
   ```

4. To see the effect of controlling for population stratification compare the above results with the Cochran-Armitage test which ignores population stratification:

   ```
   . mhodds caco NR_2
   ```

   Controlling for stratification makes little difference here.

5. You can achieve the same results as above by using logistic regression. For the 1 df analysis, complete the **Fit** menu from the **Regression** sub menu. Click `logistic` as the 'Regression command', `caco` as the 'Response variable', and `NR_2` as the 'Metric' explanatory variable. Or use the command,

   ```
   . logistic caco NR_2
   ```

   and, controlling for stratification,

   ```
   . xi:logistic caco NR_2 i.popgrp
   ```

To control for stratification within the menu system, complete the **Fit** menu with, `logistic` as the 'Regression command', `caco` as the 'Response variable', `popgrp` as the 'Categorical', and `NR_2` as the `Metric` explanatory variables. The odds ratio represents the estimated change in the ratio of cases to controls for each copy of allele 2.

6. For the 2 df analysis we declare `NR_2` as categorical, with `caco` as the response variable and use `logistic` as the regression command.

```
. xi:logistic caco i.NR_2
```

To control for stratification, include `popgrp` in the regression model also:

```
. xi:logistic caco i.NR_2 i.popgrp
```

We have now estimated separate odds ratios for homozygous and heterozygous genotypes.

7. The multiple degree of freedom test based upon all the alleles of D2S1471 can also be carried out by generating indicator variables and using logistic regression. You should first generate the indicator variables (whose names we will assume to start with the prefix `Allele_`) and carry out the regression of case-control indicator on the allele indicators. Contol for population stratification is achieved by simply adding `popgrp` into the regression (as a categorical variable). The command lines to do this are

```
. gtab d2s1471_1 d2s1471_2, gen(Allele_)
. xi:logistic caco Allele_* i.popgrp
. testparm Allele_*
```

The commands may also be launched from the **GenAssoc** menus.

8. The logistic regression analyses can be extended to look for "effect modification" by age or sex, or for differential effects in different subtypes of disease. In the case where several polymorphisms are candidates for *functional variants*, we can also use logistic regression to discriminate between direct associations between a locus and disease and indirect effects caused by linkage disequilibrium with another locus. For example, to see if the effect of D2S1471 may be explainable as an indirect effect of NRAMP, we carry out a logistic analysis. Choose `logistic` as the regression command, and `caco` as the response variable, `Allele_*` as the *Metric* and `NR_2` as the *Categorical* explanatory variables. Again select the **Test (drop)** menu and fill in the window as before. One could achieve the same result with,

```
. xi: logistic caco Allele_* i.NR_2
. testparm Allele_*
```

How would you interpret this result?