Solutions of session 7

a. What distribution does variable `N1` have and what is the variable's meaning (i.e., what does it measure)?

`N1` measures, for each covariate pattern j (j=1,…,k) the number of women using contraception out of `N1+N2`. The distribution of `N1` is binomial with probability π_j and total number of subjects `N1+N2`.

b. Why does the deviance statistic above have a chi-square distribution with 3 degrees of freedom?

The deviance (and Pearson chi-square) statistics have k-(p+1) degrees of freedom. k is the number of covariate patterns (groups; 8 in our example) and p is the number of parameters, p=4 (the number of design variables associated with "age" and "more" in the example).

c. The predicted probabilities `prob` from the model that includes `more` and `age` as well as `more×age` interaction are equal to the observed probabilities of the data. Why?

This is because, for eight groups in the data, the observed probabilities ($\hat{\pi}_j = \mathrm{N1}_j/(\mathrm{N1}_j + \mathrm{N2}_j)$), j=1,…,k can be reproduced exactly by a saturated model with k-1 degrees of freedom. Thus, the model with `more`, `age` and `more×age` interaction has seven degrees of freedom (it's the saturated model).

d. A residual larger than 2.0 should be inspected more carefully. Why?

This is because **IF** the data can be grouped into k categories with k substantially smaller than n, then we are talking about binomially distributed counts of the members of each category. Eventually, the approximate distribution is standard normal, so that a residual above 2.0 (or below –2.0) represents the extreme 5% of the distribution.