Applied Logistic Regression
Exercises 1

1. In the Doll and Hill (1950) study 649 male lung cancer cases were compared to 649 controls. Among the cases 647 were smokers, and among the controls 622 were smokers. (The controls were individually matched to cases, but we ignore this for now.) (a) Form a 2×2 table of the data, and compute the odds ratio of smoking among the cases relative to the non-cases. Does this suggest that smoking increases the risk of cancer? (b) Test whether or not there is an association between smoking and lung cancer using the ordinary Pearson's χ^2 test statistic (no continuity correction here, or in the sequel!). Does the P-value obtained support your conclusion in (a)?

2. Consider the following data (where the first row is for the exposed and the second for the non-exposed, and the first column corresponds to those ill and the second to those not ill):

57	118	175
87	185	272
144	303	447

Compute the odds ratio and carry out Pearson's χ^2 test to see if there seems to be evidence that exposure increases risk of illness. (Here and in the sequel give the values of the relevant test statistics and the corresponding P-values.)

3. Suppose then that the population was, in fact, stratified, and when classified according to the third factor that takes two values, the counts in the table were sums of the two tables below:

$$\begin{array}{|cc|} \hline 16 & 80 \\ 50 & 165 \\ \hline \end{array} \quad + \quad \begin{array}{|cc|} \hline 41 & 38 \\ 37 & 20 \\ \hline \end{array}$$

Compute the odds ratios and carry out the tests of association between rows and columns for the two tables separately. Do they confirm the results obtained in the previous exercise?

4. Compute the Mantel-Haenszel estimator for the common odds ratio to the stratified data of the previous section.

5. Carry out the Cochran-Mantel-Haenszel test for the association between the exposure and the illness for the same stratified data under the assumption of homogeneous odds ratios (using the appropriate R function). What is your overall conclusion now? In particular, how does this analysis differ from the separate analyses of the two tables?

6. Consider a logistic regression model with one explanatory variable and $n = 30$, and $x_i = (i - 15.5)/14.5, i = 1, \ldots, n$. Make one plot that displays the values p_i with the following coefficients: (i) $\beta_0 = logit(0.3), \beta_1 = 2$, (ii) $\beta_0 = logit(0.3), \beta_1 = -2$, and (i) $\beta_0 = logit(0.7), \beta_1 = 2$.

7. (a) If the log-odds for becoming ill are $= -2$, what is the probability of becoming ill? (b) If the probability of becoming ill is 0.119203, what are the log-odds?