1. Suppose $x_1$ is an exposure variable with two levels (0,1), $x_2$ is 1 if $x_1 = 2$ and zero otherwise, and that we observed the following table of exposure versus disease status:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D+</td>
<td>2</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>D-</td>
<td>61</td>
<td>82</td>
<td>52</td>
</tr>
</tbody>
</table>

(a) For the model $\log\left(\frac{P}{(1-P)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, test the hypothesis $H_0 : \beta_2 = 0$ using the score test.

(b) Test $H_0 : \beta_2 = 0$ using the Wald test.

The file `eso.dat` contains data from a study of esophageal cancer. The columns are:

(a) Age category: 1 = 25-34, 2 = 35-44, 3 = 45-54, 4 = 55-64, 5 = 65-74, 6 = 75+.

(b) Alcohol Consumption: 1 = 0-39 g/day, 2 = 40-79 g/day, 3 = 80-119 g/day, 4 = 120+ g/day.

(c) Tobacco Consumption: 1 = 0-9 g/day, 2 = 10-19 g/day, 3 = 20-29 g/day, 4 = 30+ g/day.

(d) Number of cases of esophageal cancer.

(e) Total number of subjects (cases + controls).

Using any software you like,

(a) Test the null hypotheses of no association between esophageal cancer and each of age, alcohol and tobacco consumption (ignoring the remaining variables) using three tests: the score and likelihood ratio tests versus the saturated model, and the score test versus the model with linear trend (what SAS calls the “Mantel-Haenszel Chi-Square”).

(b) Fit logistic regression models for the effect of each of age, alcohol and tobacco consumption (ignoring the remaining variables) on esophageal cancer. Test two null hypotheses: (a) that the esophageal cancer is unrelated to the exposure variables and (b) that the log-odds of esophageal cancer is linearly related to the exposure variable. For (a) perform the test against both the alternative that log-odds of esophageal cancer is linearly related to the exposure variable, and the alternative that it is related in an unstructured way. Compare the results to what you obtained in 1.

(c) Find the best fitting logistic model associating risk of esophageal cancer, age, alcohol and tobacco consumption. Note that this model may involve interactions between the
three predictor variables. For this problem, consider a model to be the “best fitting” if a test of the null hypothesis that the given model is the correct model versus an alternative that a more complicated model is the correct model is not rejected, but such a test using a less complicated model would be rejected.

Important Notes:

Please hand in the program/commands used to generate the output. Also try to minimize the amount of computer output that you hand in. Only hand in those portions which are most relevant to the problem.