2^{nd} assignment
Logistic Regression

Consider the “OESOPHAGEAL CANCER” data set (grouped.dta). Stata dataset has 6 variables all of which have labels (and value labels where necessary) attached. Data are collected within a case-control study where cases are 200 males diagnosed with oesophageal cancer and controls are 775 healthy males (population based controls).

The main interest in this study is to investigate the effect of alcohol and tobacco consumption while age is a known risk factor.

1) Perform a preliminary EDA (Exploratory Data Analysis) in order to describe the data and present the results in tabular form. If you think that you need to treat the qualitative variables as quantitative ones, at this or a later stage of the analysis, assign the values (20, 60, 100, 150) to the four levels of alcohol consumption, the values (5, 15, 25, 40) to the four levels of tobacco consumption and the values (30, 40, 50, 60, 70, 80) to the six age groups.

2) Try to find the best multivariate logistic regression model in terms of fit and parsimony. Include only main effects (not interactions). You can treat all covariates as qualitative or quantitative but you must explain and justify your choices.

3) Try to improve the fit of the previous multivariate model by including appropriate interaction terms between age and alcohol consumption (use the categorical version of these variables; consider regrouping if necessary to avoid combinations with 100% “success” or “failure” rates).

4) Check your final model and go back to steps 2 and 3 if necessary in order to find a better alternative.

5) Report the results of your final model and interpret the main findings.