Esophageal Cancer [iev]

Assignment

Problem

The dataset contains data on 200 males diagnosed with esophageal cancer in the French département of Ille-et-Vilaine and on 775 controls drawn at random from electoral rolls. Each subject was interviewed as to average daily consumption of tobacco and of alcohol in the form of beer, wine, cider, aperitif (e.g. whiskey) and digestive (mainly the apple brandy known as Calvados so popular in the region).

In this study, outcome-dependent sampling was used. Separate samples were obtained from the population with outcome variable (esophageal cancer) equal to 1 (the cases), and from the population with outcome variable equal to 0 (the controls). In epidemiology, this sampling strategy is called *the case-control design*, while in econometrics, it would be called *choice-based sampling*. As shown by Prentice and Pyke (*Biometrika*, 1978), we may treat the data as if the cancer indicator were the dependent variable in a logistic regression analysis and obtain valid asymptotic inferences about the odds ratio parameters (except the intercept that does not estimate the log odds of the disease in the population).

With these provisions, evaluate the effect of alcohol and tobacco on the odds of esophageal cancer in Ille-et-Vilaine. Consider whether these variables are best related to cancer risk in the original form or after a log transform or even differently parameterized. Explore carefully the effect of age and whether there are any interactions of age with the exposure variables. Test the hypothesis that it is alcohol alone, rather than alcohol in one of it's component forms, that is responsible for the disease risk.

Requirements

Write a report (prepared by LaTeX, LibreOffice, MS Word, ...) summarizing your solution to the problems specified during the exercise classes.

More precise specification of what exactly should be (and also what should not be) included in the report will be provided during the exercise classes related to this assignment (March 14, 21, 28). Pay attention to those instructions!

The report in the pdf format (file named as `Surname_Firstname_2.pdf`) and the related R script (file named as `Surname_Firstname_2.R`) have to be submitted in Moodle by ***Sunday April 2, 2023 [23:59 CEST]***.

Dataset

The dataset can be downloaded from
https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmst412/Problem_2/GLM_2_iev.RData

The dataframe is called `iev`. It contains 975 rows (subjects) and 11 variables.

Variable list: See Table 1.

Table 1: Variable coding table

Variable Name	Variable Label	Variable Coding
`case`	Case-control status	1 = cancer case, 0 = healthy control
`agediag`	Age at diagnosis	years
`agegr`	Grouped age at diagnosis	factor
`tobgr`	Grouped daily tobacco consumption	factor
`tob`	Tobacco consumption corresponding to tobgr	grams/day
`beer`	Daily alcohol consumption from beer	grams/day
`cider`	Daily alcohol consumption from cider	grams/day
`wine`	Daily alcohol consumption from wine	grams/day
`aper`	Daily alcohol consumption from aper	grams/day
`digest`	Daily alcohol consumption from digestives	grams/day
`alctot`	Total daily alcohol consumption	grams/day

Source: N.E. Breslow and N.E. Day: *Statistical Methods in Cancer Research*, Vol. 1. IARC, Lyon, 1980.