CS&SS/STAT/SOC 536: Logistic Regression for Case-Control Studies

Adrian Dobra
adobra@uw.edu

Consider the following description of a case-control study that uses a retrospective ("look in the past") design:

In 20 *hospitals in London, England, patients admitted with lung cancer in the preceding year were queried about their smoking behavior. For each of the* 709 *patients admitted, researchers studied the smoking behavior of a non-cancer patient at the same hospital of the same gender and within the same* 5*-year grouping on age.*

Our desired binary outcome $Y \in \{1 = yes, 0 = no\}$ is the occurrence of lung cancer. Our desired explanatory binary variable $X \in \{1 = yes, 0 = no\}$ is smoking behavior. We would like to use a logistic regression model to assess the effects of smoking on the occurence of lung cancer:

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

The strength of the association between y and x will be quantified by the slope β_1. If you look at the description of the study, you will see that the number of cases (smokers) and the number of controls (non-smokers) has been *fixed in advance*. This means that we have not actually observed the random variation of Y given X. Instead, we have observed the random variation of X given Y, therefore our data is suitable for the regression $P(x|y)$ having x as the outcome and y as the explanatory variable!

We introduce a latent variable

$$Z = \begin{cases} 1, & \text{if a subject was sampled,} \\ 0, & \text{if a subject was not sampled.} \end{cases}$$

There are two unknown quantities:

1. $\rho_1 = P(z = 1|y = 1)$, i.e. the probability of sampling a case.
2. $\rho_0 = P(z = 1|y = 0)$, i.e. the probability of sampling a control.

Remark that $\rho_1 + \rho_0 \neq 1$. Furthermore, we assume that the probability of a subject being sampled does not depend on the explanatory variable x (in this case smoking behavior):

$$P(z = 1|y, x) = P(z = 1|x).$$

A straightforward application of Bayes' theorem shows that

$$P(y = 1|z = 1, x) = \frac{\exp\left[(\beta_0 + \log(\rho_1/\rho_0)) + \beta_1 x\right]}{1 + \exp\left[(\beta_0 + \log(\rho_1/\rho_0)) + \beta_1 x\right]}$$

This means that, by fitting the logistic regression model $P(y = 1|x)$ to the data from a case-control study, we will get an incorrect estimate of the intercept β_0. However, we are not interested in the estimate of the intercept to begin with! We are interested in the estimate of β_1. *We just proved that we will actually get the correct estimate of β_1 for case-control studies, even if the design of the study seems to indicate otherwise!*