Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Odds and Odds Ratios : Keogh And Cox

The Population Model, the Sampling Model and the Inverse Model

Consider two events, A and B, their marginal probabilities, their conditional probabilities and their joint probabilities :

$$\begin{array}{ccccc} & & P(\bar{A}) & P(A) & \\ P(B) & & P(\bar{A}B) & P(AB) & P(A|B) \\ P(\bar{B}) & & P(\bar{A}\bar{B}) & P(A\bar{B}) & P(A|\bar{B}) \\ & & P(B|\bar{A}) & P(B|A) & \end{array}$$

We then have $\text{odds}(B|A)=\frac{P(B|A)}{P(\bar{B}|A)}$ and the three other odds.

We have the odds ratio $OR_B = \frac{\text{odds}(B|A)}{\text{odds}(B|\bar{A})}$ which the odds of B in the presence of A divided by the odds of B in the absence of A.

We have the odds ratio $OR_A = \frac{\text{odds}(A|B)}{\text{odds}(A|\bar{B})}$ which the odds of A in the presence of B divided by the odds of A in the absence of B.

But, it turns out that $OR_B = OR_A = \frac{P(AB)P(\bar{A}\bar{B})}{P(A\bar{B})P(\bar{A}B)}$.

Now, if we have disease D and exposure E and if all of the probabilities make sense, we see that :
$OR_D = OR_E$ so the odds ratio can be expressed in terms of either disease status or in terms of exposure status. A cross sectional study would, in principle, have all the probabilities.

Now consider a case-control study where we determine individuals with disease; the cases. and then we determine individuals without disease; the controls. There is no probability distribution for case-control status. Most authors write $P(E|D)$ for the probability of E for those selected individuals with disease and $P(E|\bar{D})$ for the probability of E for those selected individuals without disease. With this understanding, we have OR_E which is the odds of exposure for those with disease divided by the odds of exposure for those without disease.

Keogh & Cox (2014) 'Case Control Studies' provides the clearest presentation of this material and much more.
They consider three different models:
The population model : based on the joint probabilities.
The sampling model : based on the probability for exposure those with disease and the probability for exposure for those without disease.
...and a third model which they call the inverse model or the formal interpretative model : now 'formally' based on the probability of disease for those exposed and the probability of disease for those not exposed.
They make the clear statement that the inverse model is "not the distribution generating the data".

They, then, present the sampling model and the inverse model explicitly, first for this simple example:

If, for the sampling model, $p = P(E)$, then $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 D$

If, for the inverse model, $q = P(D)$, then $\log(\frac{q}{1-q}) = \alpha_0 + \alpha_1 E$.

Then, Keogh & Cox show that the MLE for α_1 is the same as the MLE for β_1 . In addition, they point out that the 'asymptotic' standard errors for the MLE's are the same and that the exact inferences based on the hypergeometric are the same.

Later in their book, Keogh & Cox, expand this material to more 'realistic' settings using logistic regression with many explanatory variables. They include considerably more elaborate contexts as well.

There is a wide class of models where the MLE's for the appropriate parameters will be identical. There are, though, many settings where the two models [sampling and inverse] will not yield the same MLE's but, as developed in Keogh & Cox, there is, nevertheless, a theoretical justification for the inferences from the inverse model. In other words, the 'classic' theory for Maximum Likelihood is shown to apply to these inverse models.

The sampling model may not be based on logistic regression. If, for example, the exposure is a measured 'continuous' variable. One could argue that the sampling model would based on linear regression and the study of the expected exposure. The inverse model would still be a logistic regression studying the log odds of disease. In this situation, there are no directly comparable regression coefficients.

Also, the sampling model and the inverse model would determine different diagnostics and assessments like the AUC, outlier detection and goodness of fit, for examples.

Selection Bias

Now consider the study of the prevalence of disease D. Suppose the probability of selection is conditional on disease status. So we have : $P(S|D)$ and $P(S|\bar{D})$ and then we have:

OR_S which is the odds of selection for those with disease divided by the odds of selection for those without disease.

Now if we are prepared to assume an analog of the Population Model with the four joint probabilities, then we have that $OR_S = OR_D$ which is the odds of disease for those selected divided by the odds of disease for those not selected.

Now if, for example, $P(S|D)$ is greater than $P(S|\bar{D})$, then $OR_S = OR_D$ is greater than one . We can then see that $P(D|S)$ is greater than $P(D)$ indicating selection bias.