Discrete Response Model Lecture 3

datascience@berkeley

Odds Ratio in the Context of Categorical Explanatory Variables

Odds ratios are useful for interpreting a categorical explanatory variable in a model; however, they are easily misinterpreted.

For example, suppose we want to interpret $OR = e^{\beta_1}$ from the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

in the example with the categorical explanatory variable having 4 levels. A common **mistake** is to interpret this odds ratio as

The odds of a success are e^{β_1} times as large as for level B than *all of the other levels*

	Indicator variables		
Levels	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

To see why this is wrong, note that the odds of a success at level B are

Levels	Indicator variables x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

$$e^{\beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 0} = e^{\beta_0 + \beta_1}$$

In order to have a resulting OR = e^{β_1}, we need the denominator of the odds ratio to be e^{β_0}. Thus, $x_1 = x_2 = x_3 = 0$ for this second odds, which corresponds to level A of the categorical variable. The correct interpretation of the odds ratio is then

The odds of a success are e^{β_1} times as large as for level B than for level A

Similar interpretations are found for e^{β_2} (compare level C to A) and for e^{β_3} (compare level D to A).

What if you would like to compare level B to level C? You need to find the ratio of two odds:

$$OR = \frac{Odds_{x_1=1, x_2=0, x_3=0}}{Odds_{x_1=0, x_2=1, x_3=0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0+\beta_2}} = e^{\beta_1-\beta_2}$$

Similarly, the odds ratios can be found for comparing level B to level D as $e^{\beta_1 - \beta_3}$ and level C to level D as $e^{\beta_2 - \beta_3}$.

- Again, please remember that an odds ratio is just the ratio of two odds.
- Whenever you have difficulty understanding an odds ratio, go back to the basics and form the ratio.

Example

- For illustration, let's start with a model without the interaction, though the interaction between infestation and control is significant

The estimate model is

$$\text{logit}(\hat{\pi}) = -0.6652 + 0.2196\text{Infest2} - 0.7933C + 0.5152N$$

The estimated odds ratios for the control methods are:

```
> exp(mod.fit$coefficients[3:4])
ControlC ControlN
0.452342 1.674025
```

```
> exp(mod.fit$coefficients[4] - mod.fit$coefficients[3])
ControlN
3.700795
```

For example, the estimated odds ratio comparing level N to level B is $e^{0.5152} = 1.67$.

The estimated odds of plants showing symptoms are 1.67 times as large for <u>using no control methods than using a biological control</u>, where the infestation method is held constant.

Because we would prefer to REDUCE the proportion of plants showing symptoms, it may be of more interest to invert the odds ratio:

> The estimated odds of plants showing symptoms are 1/1.67 = **0.5973 times** as large for using a biological control method than using no control methods, where the infestation method is held constant.
>
> **Thus, using the spider mites (biological control) is estimated to reduce the odds of a plant showing symptoms by approximately 40%.**

In order to compare the no control to the chemical control, the odds ratio is

$$OR = \frac{Odds_{C=0,N=1}}{Odds_{C=1,N=0}} = \frac{e^{\beta_0+\beta_3}}{e^{\beta_0+\beta_2}} = e^{\beta_3-\beta_2}$$

The estimated odds ratio is $e^{0.5152-(-0.7933)} = 3.70$.

Berkeley SCHOOL OF INFORMATION