**Section 2.2.6 – Categorical explanatory variables (continued)**

Odds ratios

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  from the model



in the example with the categorical explanatory variable having 4 levels. The indicator variables were coded as

|  |  |  |  |
| --- | --- | --- | --- |
|  | Indicator variables | | |
| Levels | x1 | x2 | x3 |
| A | 0 | 0 | 0 |
| B | 1 | 0 | 0 |
| C | 0 | 1 | 0 |
| D | 0 | 0 | 1 |

A common **mistake** is to interpret this odds ratio as

The odds of success are  times as large for level B than for *all of the other levels*,

where the mistake is italicized. Rather, this should be

The odds of success are  times as large for level B than for level A.

To understand this odds ratio, let’s go back to looking at the ratio of two odds – one odds for level B and one odds for level A:



Similar interpretations are found for  (compare level C to A) and for  (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:



Similarly, odds ratios can be found for comparing level B to level D as  and level C to level D as .

**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.**

Question: What would be the equation for the Wald confidence interval for the odds ratio comparing level B to level C?

Example: Control of the Tomato Spotted Wilt Virus (TomatoVirus.R, TomatoVirus.csv)

Because the interaction between Infest and Control was found to be important, we would normally focus only on the model that includes the interaction. However, it is instructive first to see how calculations are performed using the model without the interaction.

Without interaction

The estimate model is



The estimated odds ratios for the control methods are:

> exp(mod.fit$coefficients[3:4])

ControlC ControlN

0.452342 1.674025

> #Control N vs. Control C

> 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 exp(0.5152) = 1.67. Thus,

The estimated odds of plants showing symptoms are 1.67 times as large for using no control methods than for using a biological control, 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 a biological control method than for 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%.

To compare no control to chemical control, the odds ratio is



The estimated odds ratio is exp(0.5152-(-0.7933)) = 3.70.

Question: Which method – B or C – reduces the estimated odds more?

To find confidence intervals, we can use the method function corresponding to confint() of the stats package when there is only one regression parameter in the OR expression. When there are two, there are two options:

* Change the ordering of the levels for the variable of interest so that only one regression parameter is in the OR and then use the method function for confint() of stats.
* Use the mcprofile (profile LR) or emmeans (Wald) packages.

Below is the code for mcprofile:

> library(package = mcprofile)

> K <- matrix(data = c(0, 0, 1, 0,

0, 0, 0, 1,

0, 0, -1, 1), nrow = 3, ncol = 4, byrow = TRUE, dimnames = list(c("C vs. B", "N vs. B", "N vs. C"), c("beta0", "beta1", "beta2", "beta3")))

> K

beta0 beta1 beta2 beta3

C vs. B 0 0 1 0

N vs. B 0 0 0 1

N vs. C 0 0 -1 1

> linear.combo <- mcprofile(object = mod.fit, CM = K)

> ci.log.OR <- confint(object = linear.combo, level = 0.95, adjust = "none")

> ci.log.OR

mcprofile - Confidence Intervals

level: 0.95

adjustment: none

Estimate lower upper

C vs. B -0.793 -1.054 -0.536

N vs. B 0.515 0.258 0.773

N vs. C 1.309 1.031 1.589

> # R correctly applies exp() to estimates and intervals because there is a exp.mcpCI() method function

> exp(ci.log.OR)

mcprofile - Confidence Intervals

level: 0.95

adjustment: none

Estimate lower upper

C vs. B 0.452 0.349 0.585

N vs. B 1.674 1.295 2.167

N vs. C 3.701 2.805 4.901

> # Invert

> round(data.frame(OR.hat = 1/exp(ci.log.OR$estimate),

OR.low = 1/exp(ci.log.OR$confint$upper),

OR.up = 1/exp(ci.log.OR$confint$lower)), 2)

Estimate OR.low OR.up

C vs. B 2.21 1.71 2.87

N vs. B 0.60 0.46 0.77

N vs. C 0.27 0.20 0.36

For example, the 95% profile LR confidence interval comparing level N to level B is 1.30 to 2.17. Thus,

With 95% confidence, the odds of plants showing symptoms are between 1.30 and 2.17 times as large for no control methods than for a biological control (holding the infestation method constant).

Alternatively, we could also say

With 95% confidence, the odds of plants showing symptoms are between 0.46 and 0.77 times as large for a biological control method than for no control methods (holding the infestation method constant). Thus, using the spider mites (biological control) is estimated to reduce the odds of a plant showing symptoms by approximately 23% to 54%.

Below is the code for emmeans:

> library(emmeans)

> calc.est <- emmeans(object = mod.fit, specs = ~ Control, type = "response")

> confint(object = contrast(object = calc.est, method = "pairwise"), adjust = "none", level = 0.95)

contrast odds.ratio SE df asymp.LCL asymp.UCL

B / C 2.211 0.2916 Inf 1.707 2.863

B / N 0.597 0.0785 Inf 0.462 0.773

C / N 0.270 0.0384 Inf 0.204 0.357

Results are averaged over the levels of: Infest

Confidence level used: 0.95

Intervals are back-transformed from the log odds ratio scale

> confint(object = contrast(object = calc.est, method = "revpairwise"), adjust = "none", level = 0.95)

contrast odds.ratio SE df asymp.LCL asymp.UCL

C / B 0.452 0.0597 Inf 0.349 0.586

N / B 1.674 0.2198 Inf 1.294 2.165

N / C 3.701 0.5264 Inf 2.800 4.891

Results are averaged over the levels of: Infest

Confidence level used: 0.95

Intervals are back-transformed from the log odds ratio scale

Comments:

* The specs argument only contains Control because there is no interaction in the model and we are only interested in the Control variable for this odds ratio. This is why emmeans stands for “estimated marginal means” – we essentially are finding a “marginal mean” for Control during the calculations.
* The at argument was not needed in emmeans() because we are interested in all levels of Control.
* Notice how I used contrast() within confint(). This potentially can save one small step rather than creating an object that contains the results from contrast() first.
* The method = "pairwise" argument value instructs R to perform all pairwise comparisons between the levels of Control. The method = "revpairwise" argument value does the same but reverses the order of the levels (inverts the odds ratios).
* Again, profile LR intervals are preferred over Wald intervals, but we can see they are very similar due to the large sample sizes.
* Odds ratios can also be calculated for the Infest variable too. These calculations are given in the corresponding program to this example.

With interaction

The estimated model is





To understand the effect of Control on the response, we will need to calculate odds ratios where the level of Infest2 is fixed at either 0 or 1. The odds ratio comparing level N to level B with Infest2 = 0 is



The odds ratio comparing level N to level B with Infest2 = 1 is



Other odds ratios are found in a similar manner. Below are all of the estimated odds ratios and corresponding profile LR intervals for Control holding Infest2 constant:

> comp <- c("C vs. B, Infest2 = 0", "C vs. B, Infest2 = 1",

"N vs. B, Infest2 = 0", "N vs. B, Infest2 = 1",

"N vs. C, Infest2 = 0", "N vs. C, Infest2 = 1")

> K.names <- list(comp, c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5"))

> K <- matrix(data = c(0, 0, 1, 0, 0, 0,

0, 0, 1, 0, 1, 0,

0, 0, 0, 1, 0, 0,

0, 0, 0, 1, 0, 1,

0, 0, -1, 1, 0, 0,

0, 0, -1, 1, -1, 1), nrow = 6, ncol = 6, byrow = TRUE, dimnames = K.names)

> K

> K

beta0 beta1 beta2 beta3 beta4 beta5

C vs. B, Infest2 = 0 0 0 1 0 0 0

C vs. B, Infest2 = 1 0 0 1 0 1 0

N vs. B, Infest2 = 0 0 0 0 1 0 0

N vs. B, Infest2 = 1 0 0 0 1 0 1

N vs. C, Infest2 = 0 0 0 -1 1 0 0

N vs. C, Infest2 = 1 0 0 -1 1 -1 1

> linear.combo <- mcprofile(object = mod.fit.inter, CM = K)

> ci.log.OR <- confint(object = linear.combo, level = 0.95, adjust = "none")

> exp(ci.log.OR)

mcprofile - Confidence Intervals

level: 0.95

adjustment: none

Estimate lower upper

C vs. B, Infest2 = 0 0.850 0.585 1.234

C vs. B, Infest2 = 1 0.253 0.174 0.365

N vs. B, Infest2 = 0 3.083 2.116 4.517

N vs. B, Infest2 = 1 0.961 0.671 1.375

N vs. C, Infest2 = 0 3.627 2.472 5.357

N vs. C, Infest2 = 1 3.795 2.544 5.708

> # Wald

> save.wald <- wald(object = linear.combo)

> ci.log.OR.wald <- confint(object = save.wald, level = 0.95, adjust = "none")

> exp(ci.log.OR.wald)

mcprofile - Confidence Intervals

level: 0.95

adjustment: none

Estimate lower upper

C vs. B, Infest2 = 0 0.850 0.586 1.234

C vs. B, Infest2 = 1 0.253 0.175 0.367

N vs. B, Infest2 = 0 3.083 2.111 4.503

N vs. B, Infest2 = 1 0.961 0.671 1.375

N vs. C, Infest2 = 0 3.627 2.465 5.337

N vs. C, Infest2 = 1 3.795 2.535 5.681

The estimated odds ratio comparing level N to level B with Infest2 = 1 is exp(1.1260 – 1.1662) = 0.96. Thus,

The estimated odds of plants showing symptoms are 0.96 times as large for no control than for a biological control when infected thrips are released into the greenhouse.

We can also see why the interaction between Infest and Control was significant. The N vs. B and C vs. B odds ratio differ by a large amount over the two levels of Infest. However, there is not much of a difference for N vs. C over the levels of Infest.

The 95% profile likelihood ratio interval comparing level N to level B with Infest2 = 1 is 0.67 < OR < 1.38. Thus,

With 95% confidence, the odds of plants showing symptoms are between 0.67 and 1.38 times as large for no control than for a biological control when infected thrips are released in the greenhouse. Because 1 is within the interval, there is not sufficient evidence to conclude a biological control is effective in this setting.

Notice the interval for comparing level N to level B with Infest2 = 0 is 2.11 < OR < 4.52. Because the interval is above 1, there is sufficient evidence to conclude the biological control reduces the odds of plants showing symptoms when interspersing infected plants with uninfected thrips.

Comments:

* MANY more interpretations are needed to fully interpret the results of the experiment. For example, one should interpret the Infest variable for fixed levels of the Control.
* The familywise error rate can be controlled as well.

Below is how emmeans is used to compute the Wald intervals.

> calc.est.inter <- emmeans(object = mod.fit.inter, specs = ~ Infest + Control, type = "response")

> confint(object = contrast(object = calc.est.inter, method = "revpairwise", simple = list("Control")), adjust = "none", level = 0.95)

Infest = 1:

contrast odds.ratio SE df asymp.LCL asymp.UCL

C / B 0.850 0.1616 Inf 0.586 1.234

N / B 3.083 0.5959 Inf 2.111 4.503

N / C 3.627 0.7149 Inf 2.465 5.337

Infest = 2:

contrast odds.ratio SE df asymp.LCL asymp.UCL

C / B 0.253 0.0478 Inf 0.175 0.367

N / B 0.961 0.1758 Inf 0.671 1.375

N / C 3.795 0.7813 Inf 2.535 5.681

Confidence level used: 0.95

Intervals are back-transformed from the log odds ratio scale

Comments:

* The intervals match those obtained from the wald() function of mcprofile.
* Both variables are specified in specs. Including the Infest:Control interaction like in formula for glm() is not needed because specs only identifies the variables of interest.