

Discrete Response Model

Lecture 3

datascience@berkeley

Categorical Explanatory Variables

Introduction

- As we saw earlier with the change variable in the placekicking dataset, a categorical explanatory variable with two levels can be represented simply as a binary variable to reflect these two levels.
- When there are q levels (where $q > 2$), only $q - 1$ binary (often referred to as "indicator") variables are needed to represent the variable, just like those in classical linear regression.

Formulation (via an Example)

Suppose an explanatory variable has levels of A, B, C, and D. Three indicator variables can be used to represent the explanatory variable in a model:

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

Notice how each level of the explanatory variable has a unique coding. The logistic regression model is

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

Informally, we could also write out the full model as

$$\text{logit}(\pi) = \beta_0 + \beta_1 B + \beta_2 C + \beta_3 D$$

where it is assumed that B, C, or D are corresponding indicator variables in the model.

Interpretation

For example, A is the “base” level, and it is represented in the model with $x_1 = x_2 = x_3 = 0$ so that

$$\text{logit}(\pi) = \beta_0$$

For category B, the model becomes

$$\text{logit}(\pi) = \beta_0 + \beta_1$$

Thus, β_1 measures the effect of level B when compared to level A.

How R Treats Categorical Variables

- R treats categorical variables as a *factor* class type.
- By default, R orders the levels within a factor alphabetically, where numbers are ordered before letters and uppercase letters are before lowercase letters:

0, 1, 2, ..., 9, ..., a, A, b, B, ..., z, Z

- To see the ordering of any factor, the `levels()` function can be used.
- This ordering of levels is important because R uses it to construct indicator variables with the “set first level to 0” method of construction.

Example: Control of the Tomato Spotted-Wilt Virus


- Plant viruses are often spread by insects.
- This occurs by insects feeding on plants already infected with a virus and subsequently becoming carriers of the virus.
- When they feed on other plants, insects may transmit this virus back to these new plants.
- To better understand the tomato spotted-wilt virus and how to control thrips that spread it, researchers at Kansas State University performed an experiment in a number of greenhouses.
- 100 uninfected tomato plants were put into each greenhouse, and they were introduced to the virus ("infested") in one of two ways (coded levels of the corresponding variable are given in parentheses):
 1. Interspersing additional infected plants among the clean ones, and then releasing "uninfected" thrips to spread the virus (1).
 2. Releasing thrips that carry the virus (2).

Example

To control the spread of the virus to the plants, the researchers used one of three methods:

- 1) Biologically through using predatory spider mites (B)
- 2) Chemically using a pesticide (C)
- 3) None (N)

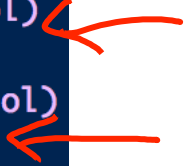
The number of plants not displaying symptoms of infection were recorded for each greenhouse after eight weeks.



```
> tomato<-read.table(file = "TomatoVirus.csv", header = TRUE, sep = ",")
> head(tomato)
```

	Infest	Control	Plants	Virus8
1	1	C	100	21
2	2	C	100	10
3	1	B	100	19
4	1	N	100	40
5	2	C	100	30
6	2	B	100	30

Both the Control and Infest explanatory variables are categorical in nature.

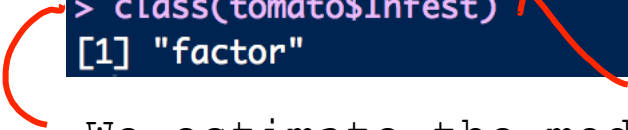


```
> class(tomato$Control)
[1] "factor"
> levels(tomato$Control)
[1] "B" "C" "N"
```


Example

For demonstration purposes, we will change the Infest variable to be a factor in the data frame:

```
> tomato$Infest<-factor(tomato$Infest)
> class(tomato$Infest)
[1] "factor"
```



We estimate the model incorporating both the Infest and Control variables:

Example

```
> mod.fit<-glm(formula = Virus8/Plants ~ Infest +
+ Control, family = binomial(link = logit), data =
+ tomato, weight = Plants)
> summary(mod.fit)
```

Call:

```
glm(formula = Virus8/Plants ~ Infest + Control, family = binomial(link = logit),
    data = tomato, weights = Plants)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.288	-2.425	-1.467	1.828	8.379

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.6652	0.1018	-6.533	6.45e-11	***
Infest2	0.2196	0.1091	2.013	0.0441	*
ControlC	-0.7933	0.1319	-6.014	1.81e-09	***
ControlN	0.5152	0.1313	3.923	8.74e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 278.69 on 15 degrees of freedom
 Residual deviance: 183.27 on 12 degrees of freedom
 AIC: 266.77

Number of Fisher Scoring iterations: 4

```
> levels(tomato$Control)
[1] "B" "C" "N"
> levels(tomato$Infest)
[1] "1" "2"
```

Example

Because the response variable is given in a binomial form, we used the **weight argument** along with the success/trials formulation in the formula argument. Estimated model:

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

- Based on the positive estimated parameter for Infest2, **the probability of showing symptoms** is estimated to be larger in greenhouses where infestation method 2 is used.
- Based on the estimated parameters for C and N, the **estimated probability of showing symptoms** is lowest for the chemical control method and highest for when no control method is used.
- Note that these interpretations rely on their not being an interaction between the explanatory variables in the model.

Hypothesis Testing

- As with classical linear regression, all indicator variables must be included in a hypothesis test to evaluate the importance of a categorical explanatory variable.

Consider again the example with the categorical explanatory variable having four levels:

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

To evaluate the importance of this explanatory variable, we need to test

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a: \text{At least one } \beta \text{ not equal to } 0$$

Note: Three separate Wald tests of $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$ are not appropriate.

One could do one overall Wald test, but we will focus instead on using a LRT because it performs better.

Interactions including categorical explanatory variables

Multiply each indicator variable by the model terms representing the other explanatory variable(s).

Berkeley

SCHOOL OF
INFORMATION