Binary Variables: Understanding Multiple Response & Dummy Coding

Binary variables have two values, ⓪ and ①. They have the properties of both categorical and numeric variables.

q7a. Do you like Math?

- ○ Like = ①
- ○ Don't Like = ⓪

*one **binary** variable*

Indicator Variable

Binary variable about 1 quality or characteristic
① presence (has it)
⓪ absence (does not)

Tip: Variables with 2 possible values should be *treated* as an Indicator and *coded* as Binary.

Tip: Name indicators after the meaning of ①: likes_math, has_pets, is_male, non_voter.

q7. Which of these subjects do you like? *(check all that apply)*

a. ☐ Math → q7a
b. ☐ English → q7b
c. ☐ History → q7c
d. ☐ Science → q7d
e. ☐ None of these

*choices are **variables***

Multiple Response

A set of indicator q's with a different look.
☑ = ① presence
☐ = ⓪ absence

To determine if the question was skipped, add "None of these". If it is also unchecked, set values to missing.

	q7a	q7b	q7c	q7d
id	math	english	history	science
101	1	0	0	0
102	0	1	0	1
103	0	0	1	0
104	1	0	0	1
105	0	0	0	0
106	.	.	.	.

4 Indicator Variables

q8. Which of these subjects do you like **best**? *(choose one)*

- ○ Math = ①
- ○ English = ②
- ○ History = ③
- ○ Science = ④

*choices are **values***

Single Response

Has choices that are **mutually exclusive** or dependent. One and only one is selected.

This creates 1 variable that is nominal and multinomial (3+ values).

			q8_1	q8_2	q8_3
id	q8		math	english	history
101	1	→	1	0	0
102	2	→	0	1	0
103	3	→	0	0	1
104	4	→	0	0	0
105	.	→	.	.	.

4 Values → 3 Dummy Variables

In **Regression**, independent (x) variables must be numeric. Binary variables are numeric, and the coefficient is interpreted as the effect of being ① *vs* ⓪. It's possible to include *any* categorical variable by using multiple binary variables.

Dummy Coding is the easiest and most-used method for including nominal variables with 3+ values in a regression, but not the only option. See also *effect coding* and *contrast coding.*

1. Select a **Reference Category**. Choose the...
   - control or status-quo ("normal") group
   - less interesting or comparison group
   - modal (biggest) or "middle" group
2. Create indicator variables for each category (value) but **ignore** the reference. Why?
   - It is redundant: if the others are not chosen we know it is. See also *degrees of freedom*
   - It makes the dummy variables independent: coefficients are the effect of changes in a value when all other variables *remain* at 0.
3. Add the indicators to your Regression and interpret coefficients as the effect of selecting that option *vs* the Reference Category.

Reference Category
Identified by ⓪ on all the dummy-coded indicator variables. Here it is ④ Science.