Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, http://www.osirjournal.net

The Grammar of Science: "Dummy" That Is Not So Dummy!

Jaranit Kaewkungwal

Mahidol University, Thailand

Corresponding author email: jaranitk@biophics.org

"Dummy" means many things. In the Merriam-Webster dictionary, "Dummy" is a person who is incapable of speaking, habitually silent, or a stupid person.[1] It could also mean an imitation, copy, or likeness of something used as a substitute. "Dummy" may have the appearance of being real, apparently acting for oneself while really acting for or at the direction of another. So how the smart "Dummy" that is not a stupid dummy gets into analyzing the data in research?

Types of Measurement in Research

Let's start at the very beginning. The measurement is the core of science which have evolved profoundly during the past century. Different methods of scaling and estimation were introduced by behavioral psychologists and statisticians.[2] In 1946, Stanley Smith Stevens wrote an article titled "On the Theory of Scales of Measurement" providing definition of measurement as the assignment of numerals to objects or events according to rules.[3] Stevens classified four types of measurement scales: (1) nominal, (2) ordinal, (3) interval, and (4) ratio scales. The scales are defined in terms of their mathematical transformations that can be conducted without changing their properties and the statistical operations that are considered permissible for each.[2,3] The scales form a specific hierarchy from the statistical point of view.

The simplest and lowest measurement is a nominal scale. As the word implies, "nominal" scale is the "name" given to two or more exhaustive categories. Numeric numbers could be assigned to each of the categories to represent different characteristics; for example, "Gender" could be categorized as 1=Male or 2=Female; "Race" as 1=White, 2=Black, 3=Asian. As the number is simply the name of the assigned category; thus, it does not mean that #1 < #2 or #2 < #3. The statistical methods which can be used with nominal scales are mostly the non-parametric statistics.

An ordinal scale is next level of measurement scaling. As the word implies, "ordinal" scale is the numeric numbers assigned to the categories based on their "order", the simplest form of "ranking". For ordinal scale, the categories are ordered along a continuum, for example, "Severity of disease" are categorized as 1=Mild, 2=Moderate, 3=Severe. The numbers assigned to the ordered categories represent degree of the differences, but not equal distances between the numbers. In the meaning of severity, #1 < #2 and #2 < #3 but the distance between #1 & #2 may not be the same as #2 & #3. Ordinal data are typically analyzed using non-parametric statistics.

Interval scale is the measurement scale representing "continuous" numbers with the "same distance/interval" between the two numbers; the distances are the same between #1 & #2 and #2 & #3, and so on. Interval scale has no "absolute zero". For example, "Temperature" can be measured in Centigrade, Fahrenheit, or Kelvin scales; each scale has its own zero point. The arbitrary zero of degree in Celsius scale is at the -32 degree in Fahrenheit scale. Interval scale data could be analyzed using either parametric or non-parametric statistical techniques:

Ratio scale has the same properties as interval scale but it has a "true zero point". When number 0 is assigned to the characteristic measured, it means the measured entity is presumed to be absent. For example, in measuring "Weight" in either Kilogram or Pound scales, 0 means no weight in both scaling units. Ratio scale and interval scale are generally used inter-changeably and analyzed with the same statistical methods.

Use of Interval/ratio Scaled Predictors in Regression Analysis

The purpose of regression analysis is to quantify the relationship between an outcome variable with one or more predictors that are measured in different types of

measurement scales. As an example, in linear regression model the outcome is measured as continuous variable (interval/ratio scale) and the predictor variables could be measured in any type of the measurement scales. When a predictor is interval or ratio scale in the regression model, it denotes how much difference might be for the outcome variable when comparing the predictor with a one-unit difference.

The following examples, used a dataset from a textbook, show a linear regression model quantifying several risk factors of mothers on their baby birth weight.[4] As shown in the model, when comparing mothers with "Age" difference for one year (21 vs. 20 years old), their "Baby weight" are different for about 12.36 grams and not statistically significant difference (p-value=0.219) (Figure 1).

bwt	Coef.	Std. Err.	t	P>|t|
age	12.36433	10.02055	1.23	0.219
_cons	2657.333	238.804	11.13	0.000

$$Linear\ Model: \quad Estimated\ Mean(bwt) = \beta_0 + \beta_1 age$$
$$= 2657.333 + 12.36433\ age$$
$$when\ \ age = 20 \quad Estimated\ Mean\left(bwt_{age=20}\right) = \beta_0 + \beta_1\ (20)$$
$$= 2657.333 + 12.36433(20) = 2904.620$$
$$when\ \ age = 21 \quad Estimated\ Mean\left(bwt_{age=21}\right) = \beta_0 + \beta_1\ (21)$$
$$= 2657.333 + 12.36433(21) = 2916.984$$

$$thus\ \ Mean\ difference:$$
$$Estimated\ Mean\left(bwt_{age=21}\right) - Estimated\ Mean\ \left(bwt_{age=20}\right) = (\beta_0 + 21\beta_1) - (\beta_0 + 20\beta_1) = \boldsymbol{\beta_1}$$
$$= 2916.984 - 2904.620 \quad = 12.364$$

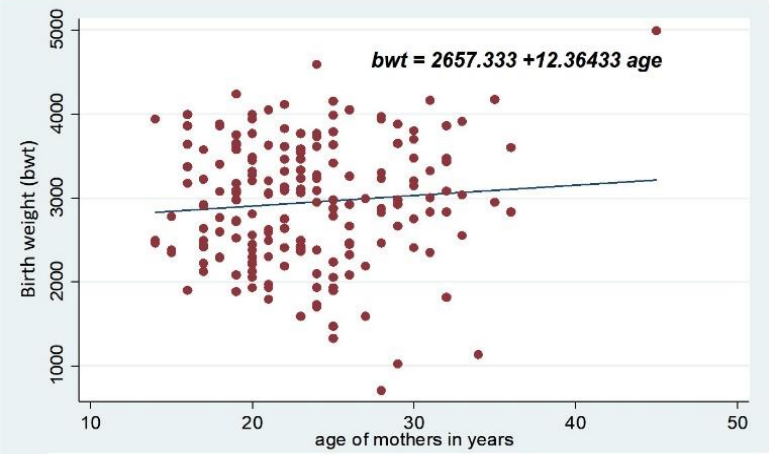

Figure 1. A linear regression model quantifying the effect of maternal age on baby birth weight

Use of Nominal/ordinal Scaled Predictors in Regression Analysis

When the predictor is measured in nominal or ordinal scale, it constitutes as a fixed scale and not equal distances between numbers. In such case, the nominal/ordinal predictor should be transcribed as the so-called "Dummy" variable.

Dummy variable is sometimes called "indicator" variable, "design" variable, "Boolean" indicator, or "proxy" variable.[5,6] As implied by the name, "Dummy" can be considered as a stand-in for a real person, an artificial attribute of the characteristics. A dummy variable in regression analysis is a numeric stand-in for a qualitative fact or a logical proposition.[6] Dummy variable is generally coded as 0 and 1; code 1 stands for "this unit belongs to category X" and 0 stands for "this unit does not belong to category X".[7] Thus, the dummy variable acts like "switch" that turn the category on and off.[8] For example, "Smoking" may be coded as a dummy variable as: 1=smoking vs. 0=non-smoking.

In a regression model, a dummy variable with a value of 0 will cause its coefficient to disappear from the equation. As shown in the model, when comparing "Smoking status" of the mothers (Yes-1 vs. No-0), their "Baby weight" are different about 281.71 grams and statistically significant difference (Figure 2).

bwt	Coef.	Std. Err.	t	P>|t|
smoke_0_1	-281.7133	106.9687	-2.63	0.009
_cons	3054.957	66.93324	45.64	0.000

$$Linear\ Model: \quad Estimated\ Mean(bwt) = \beta_0 + \beta_1 smoke\_0\_1$$
$$= 3054.957 - 281.7133\ smoke\_0\_1$$
$$when\ \ smoke = 0\ (no)\ \ Estimated\ Mean(bwt_{smoke=0}) = \beta_0 + \beta_1\ (0)$$
$$= 3054.957 - 281.7133\ (0) = 3054.597$$
$$when\ \ smoke = 1\ (yes)\ \ Estimated\ Mean(bwt_{smoke=1}) = \beta_0 + \beta_1\ (1)$$
$$= 3054.957 - 281.7133\ (1) = 2773.243$$
$$thus\ \ Mean\ difference: \quad Estimated\ Mean(bwt_{smoke=1}) - Estimated\ Mean\ (bwt_{smoke=0})$$
$$= (\beta_0 + 1\beta_1) - (\beta_0 + 0\beta_1) = \boldsymbol{\beta_1}$$
$$= 2773.243 - 3054.957 \ = \ -281.714$$

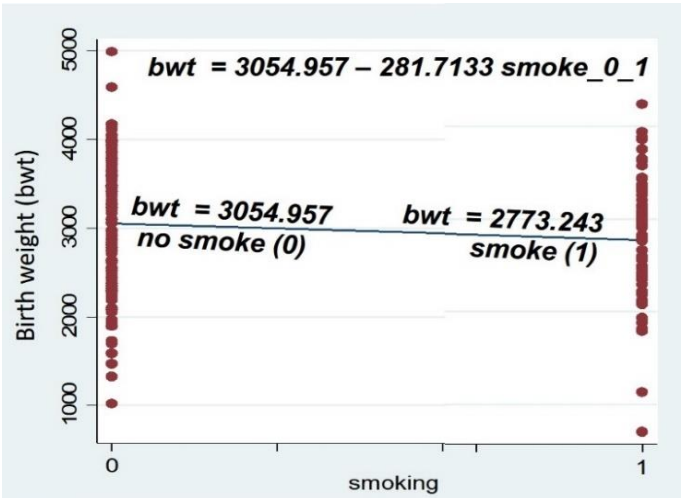

Figure 2. A linear regression model of the effect of smoking status (0,1) on baby birth weight

When the predictor is a nominal scale with two categories and the coding of the category is one-unit apart, the model will result in the same outcome estimate, but the intercept of the model will be different. When smoking status is coded as No-1 and Yes-2, the mean difference between the two groups is still -281.71 but the intercept is 3336.67 instead of 3054.95 grams (Figure 3). If the coding of the predictor is not one-unit apart, the mean difference and the intercept will be different in compensation for the values assigned to the two categories. It is more practical to use dummy variable coded as (0,1) rather than other coding scheme.

bwt	Coef.	Std. Err.	t	P>|t|
smoke_1_2	-281.7133	106.9687	-2.63	0.009
_cons	3336.67	157.7418	21.15	0.000

$$Linear\ \ Model: \quad Estimated\ Mean(bwt) = \beta_0 + \beta_1 smoke\_1\_2$$
$$= 3336.67 - 281.7133\ smoke\_1\_2$$
$$when\ \ smoke = 1\ (no)\ \ Estimated\ Mean(bwt_{smoke=0}) = \beta_0 + \beta_1\ (1)$$
$$= 3336.67 - 281.7133\ (1) = 3054.597$$
$$when\ \ smoke = 2\ (yes)\ \ Estimated\ Mean(bwt_{smoke=2}) = \beta_0 + \beta_1(2)$$
$$= 3336.67 - 281.7133\ (2) = 2773.243$$

Figure 3. A linear regression model of the effect of smoking status (1,2) on baby birth weight

When a predictor variable composes of more than two categories, more than one dummy variable must be generated to represent all characteristics. For example, "Race" variable is originally categorized as:1=White, 2=Black and 3=Asian. When creating dummy variable for Race, one can create a dummy variable called "White" and assign the coding 1= "is White" and 0= "is not White" and create other dummy variables as "Black" and "Asian" in the same fashion (Figure 4).

ID	bwt (g)	Age	Original Var
			Race
1	2523	19	1 (white)
2	2551	38	1 (white)
3	2662	28	3 (asian)
4	2600	21	3 (asian)
5	2498	17	2 (black)
6	2567	41	2 (black)

→

ID	bwt (g)	Age	Dummy Var (New)		
			White	Black	Asian
1	2523	19	1	0	0
2	2551	38	1	0	0
3	2662	28	0	0	1
4	2600	21	0	0	1
5	2498	17	0	1	0
6	2567	41	0	1	0

Figure 4. An example of dummy variable creation with three values

However, there is a redundancy in the above coding scheme; if we know that someone is not "White" and not "Black", then they are "Asian". Using all created dummy variables in a regression model would lead to a "dummy variable trap" with multicollinearity, i.e., one dummy variable can be predicted with the help of other dummy variables.[9,10] So, in this case, the regression model should be designed to include only two of the three dummy-coded variables as predictors. That is, generally, the number of dummy-coded variables needed in the model is k-1 dummy variables, where k stands for the total number of categories. The category that is left out is called the "reference" category. Choosing which category of the dummy variable to be a reference group is arbitrary, depending on the researcher's logic. As shown in the model, when "Race" variable with three categories is transformed into two dummy variables as "Black" and "Asian", with "White" as a reference groups, the estimated means of the three groups can be calculated (Figure 5). The mean difference refers to the difference of the outcome estimates between the other two groups against the reference group

ID	bwt (g)	Age	Original Var
			Race
1	2523	19	1 (white)
2	2551	38	1 (white)
3	2662	28	3 (asian)
4	2600	21	3 (asian)
5	2498	17	2 (black)
6	2567	41	2 (black)

→

ID	bwt (g)	Age	Dummy Var (Reference: White)		Dummy Var (Reference: Age ≤25)	
			Black	Asian	Age 26–35	Age >35
1	2523	19	1	1	0	0
2	2551	38	1	1	0	0
3	2662	28	0	0	1	0
4	2600	21	0	0	1	0
5	2498	17	0	0	0	1
6	2567	41	0	0	0	1

bwt	Coef.	Std. Err.	t	P>|t|
race_w1_b2~3				
black	-384.0473	157.8744	-2.43	0.016
asian	-299.7247	113.6776	-2.64	0.009
_cons	3103.74	72.88169	42.59	0.000

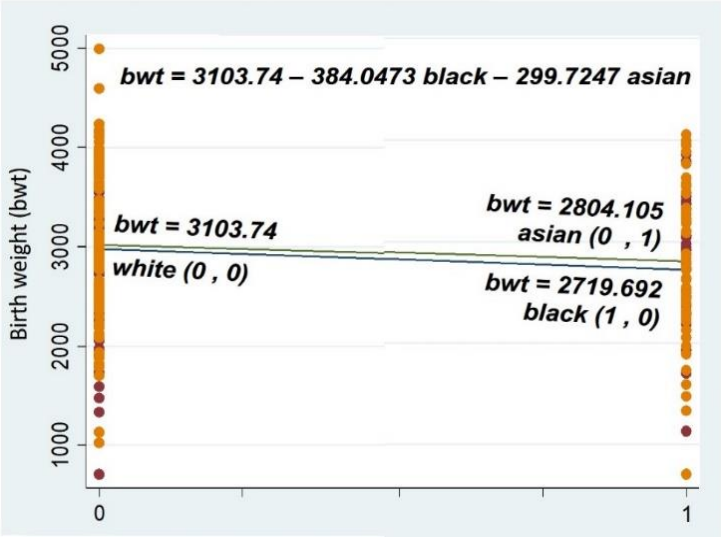

Figure 5. A linear regression model of the effect of race on baby birth weight using dummy variable creation

$$Linear\ Model: \quad Estimated\ Mean(bwt) = \beta_0 + \beta_1 black + \beta_2\ asian$$
$$= 3103.74 - 384.0473\ black - 299.7247\ asian$$
$$when\ \ race = white \quad Estimated\ Mean(bwt_{race=white}) = \beta_0 + \beta_1\ (0) + \beta_2\ (0)$$
(black = 0 ; asian = 0) $= 3103.74 - 384.0473\ (0) - 299.7247\ (0) = 3103.740$
$$when\ \ race = black \quad Estimated\ Mean(bwt_{race=black}) = \beta_0 + \beta_1\ (1) + \beta_2\ (0)$$
(black = 1 ; asian = 0) $= 3103.74 - 384.0473\ (1) - 299.7247\ (0) = 2719.692$
$$when\ \ race = asian \quad Estimated\ Mean(bwt_{race=asian}) = \beta_0 + \beta_1\ (0) + \beta_2\ (1)$$
(black = 0 ; asian = 1) $= 3103.74 - 384.0473\ (0) - 299.7247\ (1) = 2804.015$

$$thus\ \ Mean\ difference: Estimated\ Mean(bwt_{race=black}) - Estimated\ Mean\ (bwt_{race=white})$$
$$(black - white) \quad = (\beta_0 + 1\beta_1 + 0\beta_2) - (\beta_0 + 0\beta_1 + 0\beta_2) = \boldsymbol{\beta_1}$$
$$= 2719.692 - 3103.740 \ = \ -384.048$$
$$Mean\ difference: Estimated\ Mean(bwt_{race=asian}) - Estimated\ Mean\ (bwt_{race=white})$$
$$(asian - white) \quad = (\beta_0 + 0\beta_1 + 1\beta_2) - (\beta_0 + 0\beta_1 + 0\beta_2) = \boldsymbol{\beta_2}$$
$$= 2804.015 - 3103.740 \ = \ -299.725$$

Figure 5. A linear regression model of the effect of race on baby birth weight using dummy variable creation (cont.)

Tips for Using Dummy Variables in Statistical Analysis

It is not good to have a dummy variable for every (k-1) category when there are too few observations in certain category because a dummy variable for such category would be too rare to be meaningful and statistically significant. Thus, a created dummy variable may represent mixed or combined categories. For example, "Race" could be coded as: 1=White, 2=Black, 3=Others (combined all other races besides White and Black).

With the ordinal scale data, the "ranks" of the category, sometimes called "bins", could be formulated into dummy variables. Bins or ranks can act as sets of different characteristics, representing categorical, non-probabilistic set membership.[6] For example, a variable of "Severity of disease" (measures as ordinal scale) could be assumed as three types/sets of membership (nominal scale) of patients in either one of the categories of: 1=Mild, 2=Moderate and 3=Severe. By transforming an ordinal variable to dummy variables, although ordinality of the variable will not be directly considered in the regression equation, researchers can still observe the effect of ordinal nature of the variable on the outcome variable by looking at the pattern of regression coefficient values. The regression coefficients of the transformed dummy categories may reflect levels of strength of association, say dose-response pattern, between the outcome and the inherent exposure levels.

For the predictor that is continuous variable, a one-unit difference of the predictor values might have small and not significant effect on the outcome estimates, thus another way to develop a more meaningful regression model is to use independent dummy variable that represent the continuous values in a set of levels. For example, "Age" of-mother in the above example could be transformed as a set of age risk factors that might affect "Baby-weight"; two dummy variables of age group could be generated as "Age 26–35", "Age >35" with "Age ≤25" as a reference group. The cut-off points for grouping depends on the researcher's logic; it could be based on biological, clinical, social or other point-of-views.

Not only in regression model, dummy variables can be used in any statistical analysis when the researchers want to assess the effect of such variable in the models. A dummy variable is used as an independent variable in t-test, ANOVA; or as a predictor in a linear regression model. On the other hand, a dummy variable is used as a dependent or outcome variable in Binary logistic regression, Poisson regression; or as the endpoint variable in Cox's proportional hazard regression.

After all, dummy variable is not just a fake dummy of the real person. It is important and even crucial to apply dummy variables intelligently in the statistical procedures in order to have the meaningful study results.

Suggested Citation

Kaewkungwal J. The grammar of science: "dummy" that is not so dummy! OSIR. 2022 Dec;15(4):138–43.

References

1. Merriam-Webster. Dummy [Internet]. Springfield: Merriam-Webster.com dictionary; [cited 2022 Dec 10]. <https://www.merriam-webster.com/dictionary/dummy>

2. Newman EB. On the origin of "scales of measurement". In: Moskowitz HR, Scharf B, Stevens JC editors. Sensation and measurement. Dordrecht (NL): Springer; c1974.

p. 137–45. <https://doi.org/10.1007/978-94-010-2245-3_12>

3. Bandalos DL. On the theory of scales of measurement. In: Salkind NJ, editor. Encyclopedia of research design. Thousand Oaks (CA): SAGE Publications Inc.; 2010. p. 972–73. <https://dx.doi.org/10.4135/9781412961288.n292>

4. Datasets for Stata base reference manual, release 17: estat classification / estat gof [Dataset on the Internet]. College Station (TX): Stata Press; c2020–2022 [cited 2022 Dec 10]. <https://www.stata-press.com/data/r17/lbw.dta>

5. Stephanie Glen. Dummy variables / indicator variable: simple definition, examples [Internet]. [Jacksonville (FL)]: StatisticsHowTo.com: c2022 [cited 2022 Dec 10]. <https://www.statisticshowto.com/dummy-variables/>

6. Garavaglia S, Sharma A. A smart guide to dummy variables: four applications and a macro [Internet]. Los Angeles (CA): Statistical Consulting Group, UCLA; c2021 [cited 2022 Dec 10]. 10 p. <https://stats.oarc.ucla.edu/wp-content/uploads/2016/02/p046.pdf>

7. Grotenhuis MT, Thijs P. Dummy variables and their interactions in regression analysis: examples from research on body mass index [Internet]. [Ithaca (NY)]: arXiv; 2015 Nov [cited 2022 Dec 10]. 22 p. <https://arxiv.org/ftp/arxiv/papers/1511/1511.05728.pdf>

8. Trochim WMK. Research methods knowledge base: dummy variables [Internet]. Sydney: Conjointly; c2022 [cited 2022 Dec 10]. <https://conjointly.com/kb/dummy-variables/>

9. Jain D. ML | Dummy variable trap in Regression Models [Internet]. Uttar Pradesh (IN): GeeksforGeeks; 2021 Sep 8 [cited 2022 Dec 10]. <https://www.geeksforgeeks.org/ml-dummy-variable-trap-in-regression-models/>

10. Zach. What is the dummy variable trap? (definition & example). [place unknown]: Statology; 2021 Feb 2 [cited 2022 Dec 10]. <https://www.statology.org/dummy-variable-trap/>