Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Core Guide: Dummy and Effect Coding in the Analysis of Factorial Designs

Part of a series addressing common issues in statistical and epidemiological design and analysis

Background

A categorical variable can take on values that represent qualitative differences, such as religion, experimental group, ethnic group, country of birth, occupation, diagnosis, or marital status (Cohen & Cohen, 1983). These can be included as independent variables in regression analysis, but must be converted and represented quantitatively. This can be implemented through a variety of coding methods, including dummy and effect coding. Dummy coding is perhaps the most common and generally preferred coding method. However, the use of effect coding, rather than dummy coding, is suggested in factorial designs (Kugler et al., 2012). This Core Guide starts with an overview of the coding methods, with an example to illustrate the use of effect and dummy coding. This is followed by a brief introduction to factorial designs. We will then directly compare the statistical properties and implications for interpretation of these two methods in the context of the factorial design.

1. Overview of coding methods

Coding methods are used to convert the classification of a categorical variable into numeric variables in a mutually exclusive and exhaustive way (Alkharusi, 2012). For a categorical variable G that can be classified into g groups, it will take $g-1$ numeric variables to fully represent the information in the classification (Cohen & Cohen, 1983). For example, in a study where participants can be divided into 4 groups according to their highest education level completed: less than primary, primary, secondary, post-secondary (**Table 1**), we can use 3 numeric variables X_1, X_2, and X_3 to arbitrarily represent the group of participants that completed primary, secondary, and post-secondary education. The 4th group, less than primary education, can be represented implicitly by non-membership of the other 3 groups. Note that we can choose another category as the 4^{th} group, but there are no essential differences. The next step is to assign values to the newly created numeric variables. For a given observation in the dataset, we can assign a set of numerical values to each of the 3 variables that define to which level of the categorical variable that observation belongs. There are several choices of the numerical value assignment: dummy coding, effect coding, contrast coding, or even by nonsense values (Cohen & Cohen, 1983). This Core guide will only focus on dummy and effect coding.

1.1 Dummy coding

The idea of dummy coding is to assign the membership of a g-way categorical variable G using dichotomies that take on the value of 1 or 0. The values of 1 and 0 make it straightforward to represent the membership and non-membership in a particular group. To illustrate the application of dummy coding, consider the example of highest education level again. Let X_1 represent having completed primary education or not: participants are scored 1 if they completed primary education and 0 otherwise. In a similar manner, X_2, X_3, and X_4 are used to indicate whether the participant has completed secondary, post-secondary, and less than primary education, respectively.

Again, only 3 dichotomies are needed, as the 4^{th} variable is redundant, providing no additional information beyond the other 3 variables. If an individual has not completed primary, secondary, or post-secondary education, then by default that individual has less than primary education; therefore, it is not necessary to have all 4 dummies to fully represent education. The 4^{th} group, less than primary education, also serves as the reference group, receiving 0 on X_1, X_2, and X_3. In this example of education, when dummy coding is used in regression analysis, each level of education represented by the 3 dichotomies is compared to the reference group, less than primary education.

Table 1. Dummy coding for a factor of 4 groups

Education level	X_1	X_2	X_3
Less than primary	0	0	0
Primary	1	0	0
Secondary	0	1	0
Post-secondary	0	0	1

1.2 Effect coding

In effect coding, similar to dummy coding, the membership of g mutually exclusive categories of factor G is represented by $g-1$ indicator variables. However, these variables are not dichotomized, but trichotomized into -1, 0, and 1 by replacing the string of zero for the reference group in dummy coding into a string of -1. **Table 2** makes the effect coding for the 4-category education levels clear. By regressing the effect coded variables, X_1, X_2, and X_3, we are essentially comparing the outcome in a given level of education to the unweighted mean of the outcome in all levels of education.

Table 2. Effect coding for a factor of 4 groups

Education level	X_1	X_2	X_3
Less than primary	-1	-1	-1
Primary	1	0	0
Secondary	0	1	0
Post-secondary	0	0	1

2. Comparison of dummy and effect coding in factorial design

2.1 Factorial design

Factorial designs are a form of experiment where multiple factors (the experimentally controlled independent variables) are manipulated and varied (Lavrakas, 2008) to examine their main and/or the interaction effects. To illustrate the factorial design and the differences between dummy and effect coding within factorial design framework, consider the following 2×2 factorial trial. The research question of this factorial trial is to evaluate the effects of two different subsidies on the uptake of a malarial rapid diagnostic test (RDT) (O'Meara et al., 2016). The two subsidies include a direct subsidy on the RDT and a subsidy on antimalarial drugs, artemisinin combination therapy (ACT), conditional on a positive RDT result. Both RDT and ACT subsidies can be provided or not provided, forming the 2×2 factorial design with 4 experimental conditions (see **Table 3**). It is hypothesized that each of these two subsidies will increase the uptake of the diagnostic test. In this factorial experiment, roughly equal numbers of participants are enrolled into the 2×2 experimental conditions. The outcome, uptake of the diagnostic test Y_i, is binary and assumed to follow a binomial distribution. Therefore, a linear probability model specified as a generalized linear model for a binomial outcome with an identity link function is constructed as below. To match the factorial design, the interaction of RDT and ACT subsidies are included:

$$E(Y_i) = \beta_0 + \beta_1 RDT + \beta_2 ACT + \beta_3 RDT \times ACT \quad (1)$$

Using **dummy coding**, the factors RDT and ACT are coded as 1 if the corresponding subsidy is provided and 0 otherwise. The interaction term $RDT \times ACT$ is coded as 1 if both subsidies are provided and 0 otherwise. Using **effect coding**, RDT and ACT are coded as 1 if the corresponding subsidy is provided, and -1 if not. The interaction term $RDT \times ACT$ is coded as 1 if both or neither subsidies are provided and -1 if one and only one of the subsidies is offered.

Table 3. Example: 2×2 factorial experiment

	ACT Subsidy=On	ACT Subsidy=Off
RDT Subsidy=On	$n_1 = 116$	$n_2 = 107$
RDT Subsidy=Off	$n_3 = 114$	$n_4 = 107$

2.2 Derivation of main and interaction effects

2.2.1 Dummy coding

To assist the interpretation of the regression coefficients with different coding methods, the expected values of the outcome under each experimental condition are tabulated in **Tables 4-5**. When dummy coding is used (**Table 4**), β_0 represents the population mean when neither RDT or ACT subsidies is provided (condition IV). β_1 is the effect of providing RDT subsidy versus not providing RDT subsidy given no ACT subsidy. This can be shown by subtracting condition IV from condition II in **Table 4**. Similarly, β_2 is the effect of ACT subsidy given no RDT subsidy (III – IV). β_3 is the interaction effect, defined as the additional effect of RDT subsidy given ACT subsidy is offered, besides β_1 (I – III). Equivalently, it can be interpreted as the additional effect of ACT subsidy besides β_2 given RDT subsidy is offered (I – II).

Table 4. Dummy coding: population means under different experiment conditions

	Experimental condition		Population means
	RDT Subsidy	ACT Subsidy	
I	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
II	1	0	$\beta_0 + \beta_1$
III	0	1	$\beta_0 + \beta_2$
IV	0	0	β_0

2.2.2 Effect coding

Since the numbers of participants in each group are experimentally manipulated and balanced, effect coding can provide more useful interpretations. With effect coding (**Table 5**), β_0 represents the grand mean, derived as the unweighted mean of the expected outcomes across all 4 experimental conditions: I, II, III, and IV. The grand mean is the same as the true population mean if sample size in each experimental condition is balanced. β_1 is half of the effect of providing RDT subsidy versus not providing RDT subsidy, averaged across the levels of ACT subsidy. This can be derived from $\frac{1}{2}$[(I + II) – (III + IV)] and corresponds to the classical definition of main effect in factorial designs (Kugler et al., 2018). Effect coding provides an alternative interpretation for the main effect, i.e. β_1 also represents the deviation of the population mean when RDT subsidy is provided from the grand mean, β_0, averaged across the levels of ACT subsidy. To show this, calculate $\frac{1}{2}$(I + II) – $\frac{1}{4}$(I + II + III + IV). Similarly, β_2 represents half of the effect of ACT subsidy, averaged across the levels of RDT subsidy, and also the deviation from the grand mean when ACT subsidy is provided. β_3 is half of the interaction effect, irrespective of the main effects: (I – III) – (II – IV) or (I – II) – (III – IV).

Table 5. Effect coding: population means under different experiment conditions

	Experimental condition		Population means
	RDT Subsidy	ACT Subsidy	
I	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
II	1	-1	$\beta_0 + \beta_1 - \beta_2 - \beta_3$
III	-1	1	$\beta_0 - \beta_1 + \beta_2 - \beta_3$
IV	-1	-1	$\beta_0 - \beta_1 - \beta_2 + \beta_3$

The benefits of effect coding for interpretation in factorial trials include two aspects. First, as shown above, the main effects can be interpreted independently from the other factors, even when there is an interaction between two factors in the model. This is due to the property of orthogonality, which will be discussed with more details in the next section. In contrast, when dummy coding is used, the interpretation of the main effect of one factor is always conditional on certain levels of other factors. Second, the alternative way of interpreting the coefficients as deviation from the grand mean in effect coding allows for the same interpretation even under different choices of reference group, while the interpretation in dummy coding is contingent on a clearly stated reference group.

2.3 Orthogonality

By calculating population means, we have demonstrated that the interpretation of the main effects of these 2 factors can be independent of each other when effect coding is used. This can also be shown from a mathematical perspective. Effect coding forms orthogonal vectors of the factors RDT subsidy and ACT subsidy, while dummy coding does not come with this property. Two vectors are orthogonal if they are perpendicular to each other. i.e. the dot product of the two vectors is zero. Algebraically, the dot product is the sum of the products of the corresponding entries of the two sequences of numbers. As is shown in the design matrix when dummy coding is used (**Table 6**), the dot product of RDT subsidy and ACT subsidy can by calculated by multiplying each entry of the first column with the corresponding entry of the second column, and then taking the sum of all those products. The dot product is n, the sample size in each experimental condition assuming balanced design, for RDT and ACT subsidies. Similarly, the dot product of the interaction term and the main effects are also n.

In contrast, the dot products of these vectors are zero in the design matrix when effect coding is used (**Table 7**), thus these vectors are orthogonal to each other. Orthogonality between these factors implies that the estimate of the slope parameter for RDT subsidy (or ACT subsidy) in the full regression model with both explanatory variables and their interaction is the same as that in the regression models with only the main effects. The benefit of this property is tremendous as the main effect of each factor can be interpreted regardless of the presence of other factors and interactions.

However, this is only true in perfectly balanced factors. Even in factorial trials, it is not guaranteed that the numbers of observations in each experimental condition are exactly the same. Nevertheless, when the numbers of observations are unequal, the estimates are still nearly uncorrelated, except in extremely unbalanced designs (Collins et al., 2018). Outside the framework of factorial design, it is more common that study factors are not experimentally manipulated and balanced. In those cases, orthogonality does not hold, even when effect coding is used.

Table 6. Design matrix of 2×2 factorial experiment example with dummy coding

$$\begin{bmatrix} 1 & 1 & 1 \\ \vdots_n & \vdots_n & \vdots_n \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ \vdots_n & \vdots_n & \vdots_n \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots_n & \vdots_n & \vdots_n \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ \vdots_n & \vdots_n & \vdots_n \\ 0 & 0 & 0 \end{bmatrix}$$

* The three columns represent the vector for RDT subsidy, ACT subsidy, and the interaction term, respectively

**Suppose the number of participants in each group is exactly the same: n

Table 7. Design matrix of 2×2 factorial experiment example with effect coding

$$\begin{bmatrix} 1 & 1 & 1 \\ \vdots_n & \vdots_n & \vdots_n \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ \vdots_n & \vdots_n & \vdots_n \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ \vdots_n & \vdots_n & \vdots_n \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ \vdots_n & \vdots_n & \vdots_n \\ -1 & -1 & 1 \end{bmatrix}$$

* The three columns represent the vector for RDT subsidy, ACT subsidy, and the interaction term, respectively
**Suppose the number of participants in each group is exactly the same: n

2.4 Standard error

When effect coding is used to analyze a 2^k factorial experiment with equal sample sizes across experiment conditions, the standard error is the same for all regression coefficients, including main effects and interactions (Collins et al., 2018). The implication of equal standard error is the identical statistical power for the detection of main effect and interaction effect of any order. In contrast, when dummy coding is used, the standard error tends to be larger for higher order interaction terms. Therefore, the statistical power to detect non-zero higher order effects is decreased. However, it is worth noting that the equal standard error property associated with the effect coding only holds in 2^k factorial experiment.

2.5 Regression coefficients

Tables 8-9 illustrate the estimated regression coefficients with both dummy and effect coding in the model specified in **Equation 1,** along with standard errors, z test statistics, p-values, and 95% confidence intervals. Analysis was performed in Stata/SE 16.0 (StataCorp, 2019). Stata command for both regression models is shown in **Code Snippet 1**. Programming tips for creating the dummy and effect coded variables for the factors RDT, ACT, and RDT×ACT (represented as RDT_ACT in Stata) is shown in the **Appendix**. Log likelihood in both models are identical, -275.154389, indicating same model fit. The real differences lie in the interpretation of the parameter estimates and the standard errors. For example, the dummy coding estimates of β_1 represents 21% increase in the proportion of RDT testing uptake if RDT subsidy is offered, conditional on the absence of ACT subsidy. The effect of RDT subsidy in the presence of ACT subsidy is 25% ($\beta_1 + \beta_3$). Clearly, the interpretation of the main effects depends on the other factor when dummy coding is used. In comparison, when effect coding is used, the effect of RDT subsidy can be simply interpreted as 23% ($2\beta_1$) increase in the proportion of testing uptake averaged across the levels of ACT subsidy. In addition, the standard errors of both main effects and the interaction are the same under

effect coding (**Table 9**). This implies the same power for tests of significance on both the higher order interaction term and the main effects.

```
// Regression model
glm any_test RDT ACT RDT_ACT, family(binomial) link(identity)
```

Code snippet 1

Table 8. Dummy coding: estimated regression coefficients

Parameter	Estimate	Standard Error	Z statistic	P>|z|	[95% Confidence	Interval]
RDT Subsidy	.2135922	.0655693	3.26	0.001	.0850787	.3421057
ACT Subsidy	-.0286966	.0680711	-0.42	0.673	-.1621136	.1047204
RDT Subsidy*ACT Subsidy	.0322119	.0903573	0.36	0.721	-.1448853	.209309
Intercept	.5242718	.0492084	10.65	0.000	.4278252	.6207185

Table 9. Effect coding: estimated regression coefficients

Parameter	Estimate	Standard Error	Z statistic	P>|z|	[95% Confidence	Interval]
RDT Subsidy	.1148491	.0225893	5.08	0.000	.0705748	.1591234
ACT Subsidy	-.0062953	.0225893	-0.28	0.780	-.0505696	.0379789
RDT Subsidy*ACT Subsidy	.008053	.0225893	0.36	0.721	-.0362213	.0523272
Intercept	.6247726	.0225893	27.66	0.000	.5804983	.6690469

3. Conclusion

This is a brief introduction to dummy and effect coding in regression analysis, with an emphasis in factorial experiments. In factorial designs, where the researcher-controlled factors exist, effect coding is preferred due to the property of orthogonality that comes with more reasonable estimates of both the main and interaction effects and the convenience for interpretation. Outside the scope of factorial design, especially where the factors are not experimentally manipulated or sample size are extremely unbalanced, dummy coding is still the generally more preferred method.

Prepared by: Yunji Zhou, MB
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Version: 4.0, last updated 03/16/2020

References

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences (2nd ed.)* Retrieved from https://library.duke.edu/.

Kugler, K. C., Trail, J. B., Dziak, J. J., & Collins, L. M. (2012). Effect coding versus dummy coding in analysis of data from factorial experiments. University Park, PA: The Methodology Center, Pennsylvania State University.

Alkharusi, H. (2012). *Categorical variables in regression analysis: A comparison of dummy and effect coding. International Journal of Education, 4*(2), 202.

Lavrakas, P. J. (2008). Encyclopedia of survey research methods. Sage Publications.

O'Meara, W. P., Mohanan, M., Laktabai, J., Lesser, A., Platt, A., Maffioli, E., ... & Menya, D. (2016). *Assessing the independent and combined effects of subsidies for antimalarials and rapid diagnostic testing on fever management decisions in the retail sector: results from a factorial randomised trial in western Kenya. BMJ global health, 1*(2), e000101.

Collins, L. M., & Kugler, K. C. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions. Cham: Springer International Publishing. doi, 10*(1007), 978-3.

Kugler, K. C., Dziak, J. J., & Trail, J. (2018). Coding and interpretation of effects in analysis of data from a factorial experiment. In *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions* (pp. 175-205). Springer, Cham.

StataCorp. 2019. *Stata Statistical Software: Release 16.* College Station, TX: StataCorp LLC.

Sweeney, R. E., & Ulveling, E. F. (1972). *A transformation for simplifying the interpretation of coefficients of binary variables in regression analysis. The American Statistician, 26*(5), 30-32.

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). *When size matters: advantages of weighted effect coding in observational studies. International Journal of Public Health, 62*(1), 163-167.

SAS Institute Inc. 2017. SAS Note 37273. *Procedures allowing different parameterizations of CLASS predictor variables*. Retrieved from http://support.sas.com/kb/37/273.html.

Appendix

Programming tips for different coding methods in common statistical software

The use of different coding methods in common statistical software is illustrated using the 2×2 factorial trial example (O'Meara, 2016). The layout of the design is demonstrated again in **Table A.1**.

Table A.1. Example: 2×2 factorial experiment

	ACT Subsidy=On	ACT Subsidy=Off
RDT Subsidy=On	$n_1 = 116$	$n_2 = 107$
RDT Subsidy=Off	$n_3 = 114$	$n_4 = 107$

Stata:

In Stata, there are three methods available to implement dummy and effect coding for a categorical variable G with g groups. The first is to manually create g-1 variables that follow the rules specified in Section 2 and regress using the newly created variables. RDT_subsidy is the original variable, coded as 1=no RDT subsidy, 2=RDT subsidy. ACT_subsidy is coded in the same manner.

```
* Manually create dummy coded variables
   * RDT subsidy
   generate RDT =  0 if RDT_subsidy==1 /* No RDT subsidy */
   replace  RDT =  1 if RDT_subsidy==2 /* RDT subsidy */

   * ACT subsidy
   generate ACT =  0 if ACT_subsidy==1 /* No ACT subsidy */
   replace  ACT =  1 if ACT_subsidy==2 /* ACT subsidy */

   * Interaction term
   generate RDT_ACT = RDT*ACT
```

```
* Manually create effect coded variables
   * RDT subsidy
   generate RDT = -1 if RDT_subsidy==1 /* No RDT subsidy */
   replace  RDT =  1 if RDT_subsidy==2 /* RDT subsidy */

   * ACT subsidy
   generate ACT = -1 if ACT_subsidy==1 /* No ACT subsidy */
   replace  ACT =  1 if ACT_subsidy==2 /* ACT subsidy */

   * Interaction term
   generate RDT_ACT = RDT*ACT
```

As an alternative, 'igenerate' package (Te Grotenhuis et al, 2017) can be used to generate indicator variables with different coding schemes. The choices include dummy coding, effect coding, weighted effect coding, and forward/backward adjacent coding. The igenerate command will create a new numeric variable with a clear variable label for each level of the categorical variable. For example, 2 variables (RDT1 and RDT2) will be generated in this example. We need to choose the one we intend to use, in our case RDT2 and ACT2, since the other one of them (RDT1 and ACT1) will be redundant.

```
* Use igenerate package to create dummy coded variables
igenerate RDT_subsidy, gen(RDT) coding(dummy)
igenerate ACT_subsidy, gen(ACT) coding(dummy)
generate RDT_ACT = RDT2*ACT2

* Use igenerate package to create effect coded variables
igenerate RDT_subsidy, gen(RDT) coding(effect)
igenerate ACT_subsidy, gen(ACT) coding(effect)
generate RDT_ACT = RDT2*ACT2
```

Dummy coded variables can also be easily created using tab catvar, gen(newcatvar), assuming there is a categorical variable called catvar. However, effect coded variables cannot be created using this method.

Another method is to prefix the variable G with "i." in the command for the regression model. This returns the estimates for dummy coding. For effect coding, prefix the categorical variable G with "g." in the contrast command after fitting the model. Although this method does not require creating new variables, it is not recommended, since some of the more sophisticated commands, like xtgeebcv, require manually recode the categorical variables.

```
* use contrast command to generate results under effect coding
glm any_test i.RDT_subsidy##i.ACT_subsidy, family(binomial) link(identity)
contrast g.RDT_subsidy#g.ACT_subsidy
contrast g.ACT_subsidy
```

SAS:

In SAS, procedures that only allow dummy coding (referred to as GLM parameterization) include GLM, MIXED, GLIMMIX, LIFEREG, etc. Procedures that allow different parameterizations using PARAM= option in the CLASS statement include CATMOD, GENMOD, LOGISTIC, PHREG, etc. For more information, please see SAS usage note 37273 (SAS Institute, 2017).

R:

In R, categorical variables are automatically coded as dummy indicators as long as the variables are of the type factor. For effect coding, use contrasts = list(G = contr.sum) in the command for the regression model.