Creating Indicator Variables for Four Categories

Below is an illustration of how to set up indicator variables when an independent variable has multiple nominal categories. For the table below, I use the following abbreviations: D1 - D3 are dummy codes; E1 - E3 are unweighted effects codes; W1 - W3 are weighted effects codes. The referent group is the group for which all dummy variables are equal to 0, all unweighted effects codes are equal to -1, or all weighted effects codes are equal to a proportion ratio.[1] To compute the weighed effects codes for the referent group, P1, P2, P3, P4 are the proportion of cases in categories 1, 2, 3, and 4, respectively.[2]

	Dummy Codes			Unweighted Effects Codes			Weighted Effects Codes		
Participants in Category:	**D1**	**D2**	**D3**	**E1**	**E2**	**E3**	**W1**	**W2**	**W3**
1 = Social	1	0	0	1	0	0	1	0	0
2 = Developmental	0	1	0	0	1	0	0	1	0
3 = I/O	0	0	1	0	0	1	0	0	1
4 = Quant	0	0	0	-1	-1	-1	-P1/P4	-P2/P4	-P3/P4

Coding Schemes for Psychology Area and Life Satisfaction Example

i	Satisfaction	Area	D1	D2	D3	E1	E2	E3	W1	W2	W3
1	80.000	1	1	0	0	1	0	0	1	0	0
2	82.000	1	1	0	0	1	0	0	1	0	0
3	82.000	1	1	0	0	1	0	0	1	0	0
4	90.000	1	1	0	0	1	0	0	1	0	0
5	77.000	1	1	0	0	1	0	0	1	0	0
6	85.000	1	1	0	0	1	0	0	1	0	0
7	82.000	2	0	1	0	0	1	0	0	1	0
8	76.000	2	0	1	0	0	1	0	0	1	0
9	86.000	2	0	1	0	0	1	0	0	1	0
10	80.000	2	0	1	0	0	1	0	0	1	0
11	84.000	2	0	1	0	0	1	0	0	1	0
12	85.000	3	0	0	1	0	0	1	0	0	1
13	92.000	3	0	0	1	0	0	1	0	0	1
14	88.000	3	0	0	1	0	0	1	0	0	1
15	66.000	3	0	0	1	0	0	1	0	0	1
16	75.000	3	0	0	1	0	0	1	0	0	1
17	77.000	4	0	0	0	-1	-1	-1	-1.5	-1.25	-1.25
18	65.000	4	0	0	0	-1	-1	-1	-1.5	-1.25	-1.25
19	83.000	4	0	0	0	-1	-1	-1	-1.5	-1.25	-1.25
20	93.000	4	0	0	0	-1	-1	-1	-1.5	-1.25	-1.25

To compute weighted effects codes for the referent group use the proportions of cases in each group. There are 20 cases total, 6 social students, 5 developmental, 3 I/O students, and 4 quant students. The proportions of the total for each of these groups are:

Social: $P1 = 6/20 = .30$
Developmental: $P2 = 5/20 = .25$
I/O: $P3 = 5/20 = .25$
Quant: $P4 = 4/20 = .20$

Calculate the weighted effects codes for the quant referent group using these proportions:

[1] If all groups have equal *n*, then unweighted and weighted effects are the same.
[2] Note that this is one way to calculate weighted effects codes and this method differs from the prior version of the handout. There are other ways to compute codes that accomplish the same goals. For a more complete discussion see, Hardy, M.A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage

$$-\frac{P1}{P4} = -\frac{.30}{.20} = -1.5$$
$$-\frac{P2}{P4} = -\frac{.25}{.20} = -1.25$$
$$-\frac{P3}{P4} = -\frac{.25}{.20} = -1.25$$

Comparison of Results from ANOVA and the Three Coding Methods

When three or more groups are represented by two or more indicator variables, the test of a full set of $g - 1$ set of indicator variables will produce results equivalent to ANOVA. Taken together, the full set of indicator variables, whether coded with dummy codes, unweighted effects codes or weighted effects codes, will produce the same R^2, (which is equivalent to the η^2 in ANOVA) and the same F-test of its significance, with the same p-value.[3] Regardless of the coding scheme used, the overall test of differences will be identical. A significant result for this test indicates that there are differences among the g groups, but the test of the R^2 does not indicate which groups differ from one another. We must look at the test of the regression coefficients for that. In the case of one dichotomous variable, we saw that although the unstandardized coefficients differ, the standardized coefficients and the significance tests of each coefficient did not differ across coding methods. Once we have two or more indicator variables, however, the standardized coefficients and the significance tests of the coefficients will differ across coding methods, because they have different interpretations—comparison to referent group (dummy), comparison to the unweighted mean (unweighted effects), and comparison to the weighted mean (weighted effects).

The SPSS syntax below illustrates how I set up the dummy codes (there are other equivalent ways to do this). For the unweighted and weighted effects codes, the syntax would be identical, except new values would replace the 0 and 1 code values used for each indicator variable.

```
do if area eq 1.
compute SVQ=1.
compute DVQ=0.
compute IVQ=0.
else if area eq 2.
compute SVQ=0.
compute DVQ=1.
compute IVQ=0.
else if area eq 3.
compute SVQ=0.
compute DVQ=0.
compute IVQ=1.
else if area eq 4.
compute SVQ=0.
compute DVQ=0.
compute IVQ=0.
end if.
```

ANOVA Approach (using original multicategory psychology area variable)

Report

satis life satisfaction

area psych area	Mean	N	Std. Deviation
1.00 Social	82.6667	6	4.45720
2.00 Developmental	81.6000	5	3.84708
3.00 I/O	81.2000	5	10.56882
4.00 Quant	79.5000	4	11.70470
Total	81.4000	20	7.40128

[3] The meaning of the indicator variables is lost unless a full set is used together, so there is virtually no reason to ever examine dummy (or effects) variables separately.

Tests of Between-Subjects Effects

Dependent Variable: satis life satisfaction

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	24.467 [a]	3	8.156	.128	.942	.024
Intercept	129310.205	1	129310.205	2035.713	.000	.992
area	24.467	3	8.156	.128	.942	.024
Error	1016.333	16	63.521			
Total	133560.000	20				
Corrected Total	1040.800	19				

a. R Squared = .024 (Adjusted R Squared = -.160)

Regression Approach
Dummy Codes

For the dummy coding example, I use more meaningful labels to illustrate how I name my variables in practice. Giving each indicator variable a meaningful name helps with reading the output and remembering which groups are compared. So, D1, D2, and D3 above are renamed in my data set to be SVQ (social vs. quant), DVQ (developmental vs. quant), and IVQ (I/O vs. quant), respectively.[4]

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.153 [a]	.024	-.160	7.97000

a. Predictors: (Constant), IVQ, DVQ, SVQ

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.467	3	8.156	.128	.942 [b]
	Residual	1016.333	16	63.521		
	Total	1040.800	19			

a. Dependent Variable: satis life satisfaction

b. Predictors: (Constant), IVQ, DVQ, SVQ

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	79.500	3.985		19.950	.000	71.052	87.948
	SVQ	3.167	5.145	.201	.616	.547	-7.739	14.073
	DVQ	2.100	5.346	.126	.393	.700	-9.234	13.434
	IVQ	1.700	5.346	.102	.318	.755	-9.634	13.034

a. Dependent Variable: satis life satisfaction

Unweighted Effect codes

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.153 [a]	.024	-.160	7.97000

a. Predictors: (Constant), e3, e1, e2

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.467	3	8.156	.128	.942 [b]
	Residual	1016.333	16	63.521		
	Total	1040.800	19			

a. Dependent Variable: satis life satisfaction

b. Predictors: (Constant), e3, e1, e2

[4] Another common system is to name each dummy variable after the group (e.g., SOCIAL), which requires you to remember which group is the comparison.

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	81.242	1.801		45.119	.000
	E1	1.425	2.922	.138	.488	.632
	E2	.358	3.097	.033	.116	.909
	E3	-.042	3.097	-.004	-.013	.989

a. Dependent Variable: satis life satisfaction

Weighted Effects codes

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.153 [a]	.024	-.160	7.97000

a. Predictors: (Constant), w3, w2, w1

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.467	3	8.156	.128	.942 [b]
	Residual	1016.333	16	63.521		
	Total	1040.800	19			

a. Dependent Variable: satis life satisfaction

b. Predictors: (Constant), w3, w2, w1

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	81.400	1.782		45.675	.000
	W1	1.267	2.722	.152	.465	.648
	W2	.200	3.087	.021	.065	.949
	W3	-.200	3.087	-.021	-.065	.949

a. Dependent Variable: satis life satisfaction

R Code for Constructing Indicator Variables

Note that the code below assumes you start with a numeric variable for the original multi-category variable. The last category (4=Quant) does not need to be mentioned as zeros are created except when `area` equals 1, 2, or 3.

```
#may be needed if area is assumed to be a factor in the data set
#d$area = as.numeric(d$area)

#base R
d$svq<-ifelse(d$area==1, 1,0)
d$dvq<-ifelse(d$area==2, 1,0)
d$ivq<-ifelse(d$area==3, 1,0)
```

Example Write-Up

I discuss only the dummy coded results here. A multiple regression model tested the differences in life satisfaction among graduate students in the four specialty areas. Three dummy variables were entered simultaneously, with the quantitative psychology group serving as the referent for each. The differences among the groups was not significant, accounting for only approximately 2.4% of the variance in life satisfaction, R^2 = .024, F(3,16) = .128, p = .94. Moreover, although the mean group difference on the life satisfaction scale ranged between approximately 1.70 points and 3.17 points, none of the individual comparisons with quantitative psychology was significant, B = 2.167, 95%CI[-7.739, 14.073], β = .201, p = .55 (social psychology), B = 2.10, 95%CI[-9.234,13.434], β = .126, p = .70. (developmental psychology), B = 1.70, 95%CI[-9.634,13.034], β = .102, p = .76 (industrial/organizational psychology). Overall, the results suggest life satisfaction is equally poor among all psychology students.☹ [*Note: sad faces are not permitted in APA format*].