FACTOR PREDICTOR REGRESSION USING DUMMY VARIABLES

We can fit a factor predictor using the *Linear Regression* pulldown in SPSS by using **dummy variables**.

Suppose that a **factor predictor**, X, takes L levels, indexed by $l = 1, 2, \ldots, L$. We proceed as follows:

1. Define L **new** "dummy" variables $X_1, \ldots, X_L$, where, for $l = 1, \ldots, L$,

$$X_l = \begin{cases} 1 & \text{if } X = l \\ 0 & \text{if } X \neq l \end{cases}$$

2. Fit the multiple regression model with $L - 1$ of the dummy variables as continuous covariates, that is,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{L-1} x_{L-1} + \epsilon_i$$

   Note that we cannot include all of $X_1, X_2, \ldots, X_L$if we have an intercept β_0 in the model; we omit X_L and regard L as the baseline group.

The estimates, standard errors etc. from this model are identical to those obtained using the *General Linear Model* analysis.

EXAMPLE : Diabetes Data Set

The data set **DIABETES.SAV** has three subgroups defined by different patient characteristics. Thus $L = 3$. A subset of the data are displayed below, with the new variables X_1, X_2 and X_3 defined as above. They can be computed using the

Compute

pulldown menu, or entered by hand.

ID	glutest	group	Dummy 1	Dummy 2	Dummy 3
	y	x	x_1	x_2	x_3
1	356	3	0	0	1
2	289	3	0	0	1
3	319	3	0	0	1
4	356	3	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
87	503	2	0	1	0
88	540	2	0	1	0
89	469	2	0	1	0
90	486	2	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
113	1468	1	1	0	0
114	1487	1	1	0	0
115	714	1	1	0	0
116	1470	1	1	0	0

The analysis below indicates that the estimated coefficients and the ANOVA results are identical whether we use the *General Linear Model* or *Regression* pulldown menus.

Factor Predictor Fitted Using *General Linear Model*

Between-Subjects Factors

		Value Label	N
Clinical Group	1	Overt Diabetic	32
	2	Chemically Diabetic	36
	3	Normal	76

Tests of Between-Subjects Effects

Dependent Variable: Log(GluTest)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
~~Corrected Model~~	~~24.344(a)~~	~~2~~	~~12.172~~	~~412.568~~	~~.000~~
~~Intercept~~	~~4969.483~~	~~1~~	~~4969.483~~	~~168437.466~~	~~.000~~
group	24.344	2	12.172	412.568	.000
Error	4.160	141	.030		
~~Total~~	~~5509.040~~	~~144~~			
Corrected Total	28.504	143			

a R Squared = .854 (Adjusted R Squared = .852)

ANOVA results identical

R squared identical

Parameter Estimates

Dependent Variable: Log(GluTest)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.852	.020	297.026	.000	5.813	5.891
[group=1]	1.039	.036	28.704	.000	.967	1.111
[group=2]	.344	.035	9.905	.000	.276	.413
[group=3]	0(a)	.	.	.	.	.

a This parameter is set to zero because it is redundant.

Factor Predictor Fitted Using *Linear Regression*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.924(a)	.854	.852	.17177

a Predictors: (Constant), Group = 2, Group = 1

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.344	2	12.172	412.568	.000(a)
	Residual	4.160	141	.030		
	Total	28.504	143			

a Predictors: (Constant), Group = 2, Group = 1
b Dependent Variable: Log(GluTest)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5.852	.020		297.026	.000	5.813	5.891
	Group = 1	1.039	.036	.971	28.704	.000	.967	1.111
	Group = 2	.344	.035	.335	9.905	.000	.276	.413

a Dependent Variable: Log(GluTest)

Estimates of coefficients, standard errors etc. are identical