FACTOR PREDICTOR REGRESSION

We need to take some care when combining factor predictors and covariates in the regression model. Suppose that we have only two predictors

- A **covariate**, x_1
- A **factor predictor**, x_2, now taking L levels, with the levels being indexed by $l = 1, 2, \ldots, L$.

We want to build a model that takes into account both x_1 and x_2.

EXAMPLE : Binary Factor $L = 2$
Suppose that factor predictor x_2 takes two levels, labelled 0 and 1, that identify two data subgroups. Five models can be considered, that correspond to different straight-line models

- **MODEL 0 : Same intercept**, **slope zero**, in the two subgroups
- **MODEL 1 : Different intercept**, **slope zero**, in the two subgroups
- **MODEL 2 : Same intercept**, **same non-zero slope**, in the two subgroups
- **MODEL 3 : Different intercept**, **same non-zero slope**, in the two subgroups
- **MODEL 4 : Different intercept**, **different non-zero slopes**, in the two subgroups

We can write out the models in terms of the usual slope and intercept parameters. The general model can be written

$$y = \begin{cases} \beta_{00} + \beta_{01} x_1 + \epsilon & \text{GROUP 0} \quad (l = 0) \\ \\ \beta_{10} + \beta_{11} x_1 + \epsilon & \text{GROUP 1} \quad (l = 1) \end{cases}$$

- **MODEL 0 :** $\beta_{00} = \beta_{10} = \beta_0, \ \beta_{01} = \beta_{11} = 0$
- **MODEL 1 :** $\beta_{00} \neq \beta_{10}, \ \beta_{01} = \beta_{11} = 0$
- **MODEL 2 :** $\beta_{00} = \beta_{10} = \beta_0, \ \beta_{01} = \beta_{11} = \beta_1 \neq 0$
- **MODEL 3 :** $\beta_{00} \neq \beta_{10}, \ \beta_{01} = \beta_{11} = \beta_1 \neq 0$
- **MODEL 4 :** $\beta_{00} \neq \beta_{10}, \ \beta_{01} \neq \beta_{11}$

The numbers of parameters, p, in each model are as follows:

MODEL 0	:	$p = 1$	β_0
MODEL 1	:	$p = 2$	β_{00}, β_{10}
MODEL 2	:	$p = 2$	β_0, β_1
MODEL 3	:	$p = 3$	$\beta_{00}, \beta_{10}, \beta_1$
MODEL 4	:	$p = 4$	$\beta_{00}, \beta_{10}, \beta_{10}, \beta_{11}$

SPSS Parameterization: The default parameterization used by SPSS is different from the one described above. SPSS takes a baseline group, and looks for **differences** in the parameters compared to the baseline group. The baseline group is taken to be the last listed subgroup for the factor predictor; in the binary example above, the baseline group would be Group 1.

The interaction model is therefore written

$$y = [\beta_0 + (1 - x_2)\delta_{00}] + [(\beta_1 + (1 - x_2)\delta_{01})x_1] + \epsilon$$

- δ_{00} is the **change in intercept** from Group 1 to Group 0
- δ_{01} is the **change in slope** from Group 1 to Group 0

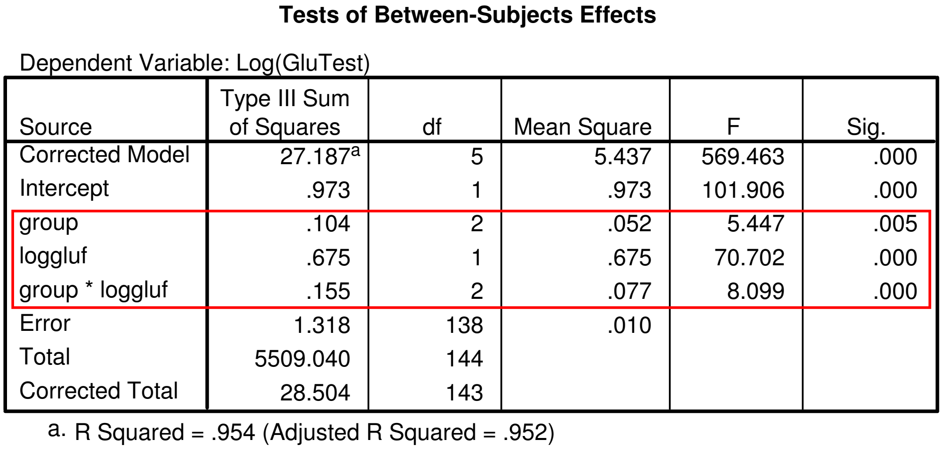
EXAMPLE : Diabetes Data Set

The data in the data set **DIABETES.SAV** contain information on 68 diabetes patients falling into two clinically different categories (overt and chemical diabetics) and 76 normal controls. Measurements of plasma glucose in blood samples when fasting and in a dietary test are recorded.

The objective is to predict the the test glucose levels from the fasting glucose levels in the three subgroups, and to find out if there is any significant difference between the subgroups.

In this analysis, there is a single response variable, one covariate and one factor predictor:

- y : **glutest**, the test glucose level
- x_1 : covariate **glufast**, the fasting glucose level
- x_2 : factor predictor **group**, the diabetes group
  - GROUP 1: Overt Diabetic
  - GROUP 2: Chemical Diabetic
  - GROUP 3: Normal Patients

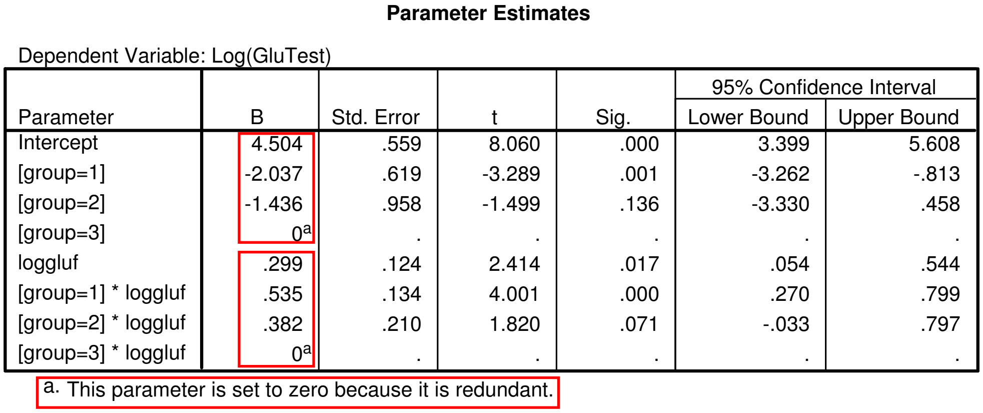


Tests of Between-Subjects Effects

Dependent Variable: Log(GluTest)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27.187[a]	5	5.437	569.463	.000
Intercept	.973	1	.973	101.906	.000
group	.104	2	.052	5.447	.005
loggluf	.675	1	.675	70.702	.000
group * loggluf	.155	2	.077	8.099	.000
Error	1.318	138	.010		
Total	5509.040	144			
Corrected Total	28.504	143			

[a] R Squared = .954 (Adjusted R Squared = .952)



Parameter Estimates

Dependent Variable: Log(GluTest)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.504	.559	8.060	.000	3.399	5.608
[group=1]	-2.037	.619	-3.289	.001	-3.262	-.813
[group=2]	-1.436	.958	-1.499	.136	-3.330	.458
[group=3]	0[a]	.	.	.	.	.
loggluf	.299	.124	2.414	.017	.054	.544
[group=1] * loggluf	.535	.134	4.001	.000	.270	.799
[group=2] * loggluf	.382	.210	1.820	.071	-.033	.797
[group=3] * loggluf	0[a]	.	.	.	.	.

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

The first ANOVA table demonstrates that there **is** a significant interaction between the covariate and the factor predictor ($F = 8.099$, p-value < 0.001). This means that there is a **significantly different slope** in **at least two of the three** subgroups.

The second table gives the slope and intercept parameters in the three groups. The SPSS parameterization is not directly in terms of the slopes and intercepts, but looks at **differences** from baseline subgroup, Group 3. For example, the Group 1 intercept and slope are, respectively,

$$\text{INTERCEPT} : 4.504 + (-2.037) = 2.467 \qquad \text{SLOPE} : 0.299 + 0.535 = 0.834.$$