Statistical Corner

Wilfred Hing-Sang WONG 黃慶生,[1] **Chun-Fan LEE** 李俊帆,[2] **Daniel Yee-Dak FONG** 方以德[2]
[1]Department of Paediatrics & Adolescent Medicine; [2]Department of Nursing Studies, The University of Hong Kong, Hong Kong

Question 1:
What is regression?

Answer :
Regression is a statistical method to examine the use of one or more factors to explain or predict outcomes of interest. Therefore, regression analysis has a priori hypothesis of a causal relationship between factors and outcomes. However, it is important to note that a causal relationship cannot be established by just running a regression analysis. We also need to have an appropriate study design.

There can be many types of regression methods. The simplest type is the linear regression that exhibits itself by a regression equation:

$$\text{Outcome} = b_0 + b_1 \text{ Factor1} + b_2 \text{ Factor2} + \ldots$$

For linear regression with more than one factor, we called it a multiple linear regression.

Question 2:
How to perform linear regression in SPSS?

Answer:
There can be two ways to perform linear regression in SPSS:

1. Analyze>Regression>Linear
2. Analyze>General Linear Model>Univariate

The first method can only accommodate factors measured at least on an ordinal scale but is equipped with automated variable selection procedure. The second method may allow nominal factors but not automated variable selection.

The use of the two methods is illustrated in Figures 1 and 2, using a dataset available at http://www.hkspr.org/. In Figure 1, education level is treated as continuous and one level increase in education leads to 25.5 mg/dl increase in cholesterol, on average. In Figure 2, education level is treated as categorical and its effects are expressed as the differences of education levels 1 and 2 with level 3 (the reference level; the one with the highest value). For example, subjects at education level 1 are, on average 50.9 mg/dl lower in cholesterol than those at education level 3. Therefore, Figure 1 examines the linear trend of education level while Figure 2 examines the difference between different education levels, which in general results in smaller standard errors as more effects are being considered.

Question 3:

How do we know a linear regression analysis is appropriate?

Answer:

It is vitally important to check adequacy of a linear regression model. It can be done by examining residuals that are readily available in SPSS. There are at least two plots we need to obtain: Normal probability plot of residuals and scatter plot of residuals against predicted values (Figure 3). Model adequacy is demonstrated when points on the Normal probability plot are close to the diagonal line and points on the scatter plot appears to be randomly scattered.

Question 4:

How should factors be selected in a linear regression model?

Answer:

Selection of factors into a regression model is not a trivial task. Automated variable selection procedure is only appropriate as an exploratory tool or in prediction. Effects of factors are preferably examined by also accounting their inter-relationships. For example, the effect of a factor (say A) should be adjusted for factors that cause A but not factors that are caused by A. Therefore, selection of factors should incorporate both statistical and clinical expertise.

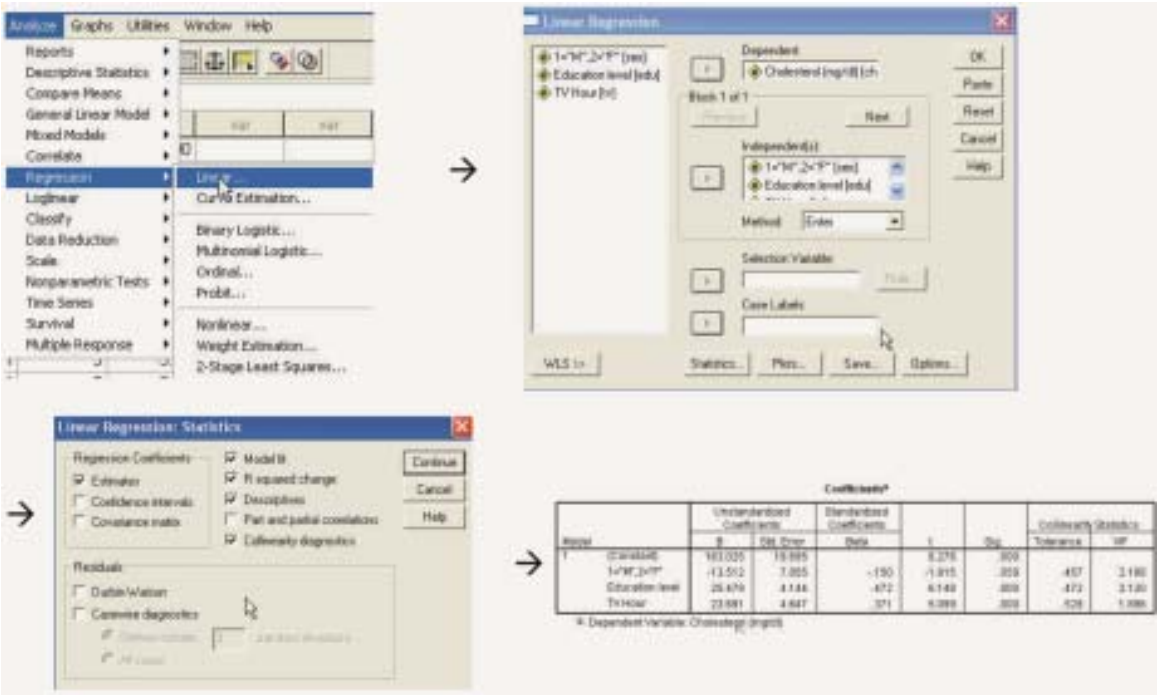
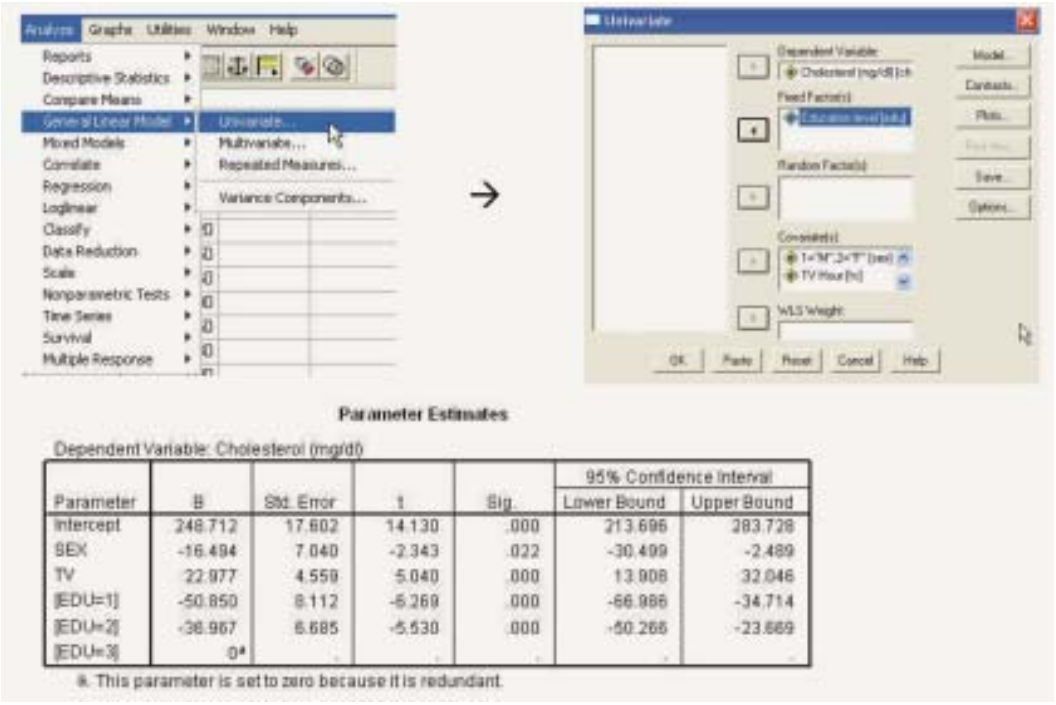


Figure 1. Using Regression>Linear in SPSS



Parameter Estimates

Dependent Variable: Cholesterol (mg/dl)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	248.712	17.602	14.130	.000	213.696	283.728
SEX	-16.494	7.040	-2.343	.022	-30.499	-2.489
TV	22.977	4.559	5.040	.000	13.908	32.046
[EDU=1]	-50.850	8.112	-6.269	.000	-66.986	-34.714
[EDU=2]	-36.967	6.685	-5.530	.000	-50.266	-23.669
[EDU=3]	0[a]	.	.	.	.	.

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

Figure 2. Using General Linear Model>Univariate in SPSS

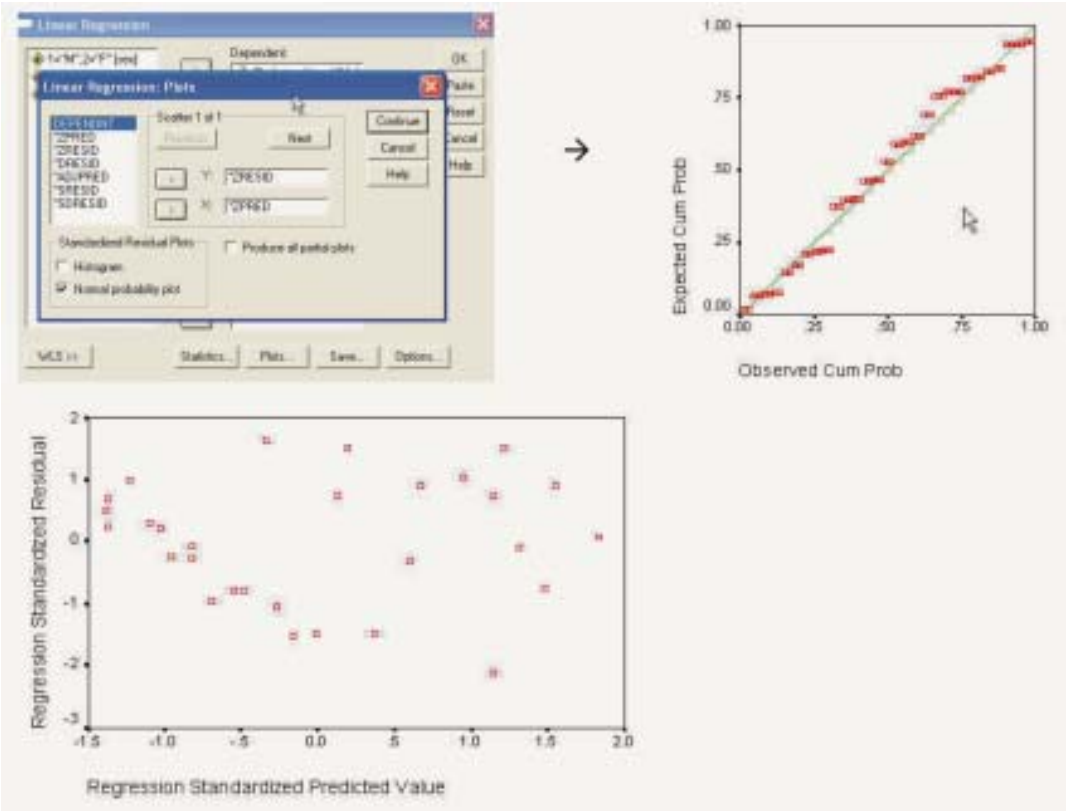


Figure 3. Obtaining residuals plots using Regression>Linear in SPSS