OPD Memorandum 2002-1

Structural Analysis of a Univariate Latent Variable (SAUL)
Theory and a Computer Program

N.D. Verhelst

H.H.F.M. Verstralen

Cito groep

POK

OPD Memorandum 2002-1

Structutal Analysis of a Univariate Latent Variable (SAUL) Theory and a Computer Program

H.H.F.M. Verstralen

N.D. Verhelst

Citogroep
Arnhem, 2002

€ 5

Contents

Chapter 1

The theory

1.1 Introduction

The name SAUL is an acronym for Structural Analysis of a Univariate Latent Variable. The purpose of such an analysis is to relate a latent variable to on a number of explanatory variables, such as gender, age, method of instruction used, etc. Conceptually, the analysis is a panel of a more general part of a multivariate analysis, which essentially consists of two parts.

The measurement model In this model the k-variate variable 'item responses' is explained by a unidimensional latent trait, usually denoted θ. There are two problems associated with the measurement model: the item parameters have to be estimated and the validity of the model has to be tested. The model that is used most of the time at Cito is the OPLM-model. The advantage of this model is that item parameters can be estimated without making any assumption about the distribution of θ in the population. This is accomplished by using the CML method of estimation.

The structural model In survey research, one is usually interested in a description of the distribution of θ in the population. Such a description may be fairly simple, such as estimating the mean and the variance of θ in the population, but may also be quite complicated by asking questions like 'is there a difference in the mean θ of boys and girls' or 'is there a change in mean θ from one period to another', or even more complicated: 'has the difference between boys and girls changed over time?' The important thing to notice is that these questions are not asked with respect to observed test scores (for example, the number of correct responses), but with respect to the latent variable θ. The

program SAUL is a tool to carry out the computations needed for such analyses.

Relation between the two models The measurement model accounts for the relation between the items and the latent variable θ, while the structural model accounts for the relation between θ and one or more background variables. It should be stressed that the approach we use in SAUL does not assume that in some way estimates of the θ-values are available. The input for the program essentially consists of two main parts: (i) the item-responses as well as the values of the background variables must be present (in one or two files, as will be explained later), and (ii) a table of the item-parameters, since the program will not estimate these but use the values supplied by the user. Usually this table is contained in a file that is produced by the program that estimates the item parameters. For the current version of SAUL the only file that is accepted is a *.PAR file produced by the estimation module OPCML. This means that within SAUL the item parameter estimates are considered as 'true' values. The disadvantage of such an approach is that estimation errors in the item parameter estimates are ignored. The advantage of the approach is that measurement model and structural model are nicely separated, such that the sample used for calibration may be different from the sample used to estimate the structural model.

1.2 Regression Analysis

1.2.1 Basics

The basic model in SAUL is a univariate regression analysis, where the dependent variable (or the criterion as it is sometimes called) is not observed. Therefore, it may be instructive to compare the regression model used in SAUL to the common regression model which is widely used in applied statistics.

We will use a simple example to demonstrate the similarities and the differences between the two approaches. Suppose we have a measure of mathematics proficiency on a sample of students. Such a measure could be the score obtained on a mathematics test. Represent the score of student i as Y_i. Suppose we want to estimate the difference between the mean score of boys and girls (in the population!), assuming we have a random sample from the boys population and a random sample from the girls population, and to decide whether the difference between the means is zero or not zero. For

the decision we can use a statistical procedure like a t-test or an analysis of variance, but we can also use a regression approach. To do this we have to recode the values of the variable gender (which are 'boy' and 'girl') to numerical values. Usually, for a variable like gender, which can take only two values, we recode to the numerical values 0 and 1. By doing this we create a numerical variable, which we will denote by X_{1i}. For the example we define this variable as follows:

$$X_{1i} = \begin{cases} 1 \text{ if student } i \text{ is a girl} \\ \\ 0 \text{ if student } i \text{ is a boy} \end{cases}$$

The regression model contains two parts. The first part is given by the regression equation:

$$Y_i = \mu + \beta_1 X_{1i} + \varepsilon_i \tag{1.1}$$

where μ and β_1 are (unknown) numbers which apply to all students and ε_i is also an unknown number, but it can vary from student to student. This number is called the residual. The first purpose of such a model is to estimate the unknown numbers μ and β_1 from the data we have. Although this is possible without adding more assumptions, it is useful to add one more assumption which makes it possible not only to estimate parameters, but also to test statistical hypotheses and building confidence intervals. This second assumption says that the residuals are normally distributed with mean zero and an unknown variance σ^2. We write this formally as

$$\varepsilon_i \sim N(0, \sigma^2) \tag{1.2}$$

The unknown numbers μ and β_1 are called regression coefficients, and σ^2 is called the residual variance. To have a nice interpretation of the regression coefficients we compute the mean test score (or what is the same: the expected test score) in the population of the girls and in the population of the boys. This gives for the boys

$$\begin{aligned} E(Y|X_1 = 0) &= \mu + \beta_1 \times 0 + E(\varepsilon_i|X_1 = 0) \\ &= \mu + 0 = \mu \end{aligned}$$

and we see that μ is the mean of the test score in the population of boys, or more generally, in the population that has received the coding zero for the variable X_1. This population is called the reference population, and the value of the original variable ('boys') is called the reference value. By the same technique we find that for the girls

$$\begin{aligned} E(Y|X_1 = 1) &= \mu + \beta_1 \times 1 + E(\varepsilon_i|X_1 = 1) \\ &= \mu + \beta_1 \end{aligned}$$

and we see that the *difference* between the mean of girls and boys is $(\mu + \beta_1) - \mu = \beta_1$. So we see that β_1 is the difference between the mean of the girls population and the reference population. If β_1 is positive, the girls do better than the boys, and if negative, the boys do better. The difference between the mean of girls and boys is also called the gender effect.

The assumption (1.2) says that the variance of the scores in the boys population and in the girls population is the same. This equality of variances is called homoscedasticity in statistical theory. It may be the case that this assumption is not true, and then more complicated techniques have to be used. In this manual, however, we will assume that assumption (1.2) is adequate.

The model that is used in the program SAUL has the same structure as (1.1) and (1.2) above, with one very important exception: the dependent variable is not observed, and in the equation it is substituted by the unobserved (latent) variable θ:

$$\theta_i = \mu + \beta_1 X_{1i} + \varepsilon_i \tag{1.3}$$

Of course, the model as it stands is not of much use, if we have no information about θ. This information is contained in the observed item-responses, and the exact relation with the item responses is given by the measurement model. If the observed item responses of student i are represented by $Y_{i1}, Y_{i2}, \ldots, Y_{ik}$, the model can be represented schematically as

$$\left.\begin{matrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{matrix}\right\} \leftarrow \theta_i = \mu + \beta_1 X_{1i} + \varepsilon_i \tag{1.4}$$

where the $\leftarrow$ represents the relation between the item responses and θ. In this way one can understand what the program needs as input: for the left hand side of (1.4) we need to give all the item responses, and for the right hand side the value of the background variable (in the example, the gender converted to a numerical code) are needed. But we must also supply the precise nature between θ and the item responses. This is done by telling the program which measurement model is used (in our case this is always OPLM), and by supplying a list of the item parameters. For the program this list is contained in the PAR file which is issued by the program OPCML. The program SAUL reads this file; so it is not required to produce a special file.

Now we discuss two important differences between the use of (1.1) and (1.4).

1. The first is on the interpretation of the residual ε_i. Sticking to the example that the background variable X_1 is gender, and assuming that θ is mathematical ability, the residual ε_i in (1.3) says that in the subpopulations of boys the mathematical ability is not constant: some boys are more able than other boys. The equation (1.2) tells us how big these differences are. This is given by the variance σ^2. If we can estimate this variance accurately we know how big the differences are in the subpopulation of boys. Of course, the same argument applies to the subpopulation of girls. Now suppose, we do not have a program like SAUL, but we decide to substitute θ in (1.3) by an estimate of θ, (The program OPCML. and the program OPLAT provide several estimates of θ, which we will indicate here as $\widehat{\theta}$.) Notice that, although we do not observe the real θ-value of students, we can produce for every student an estimate. So, this estimate is observed, and we could use the model

$$\widehat{\theta}_i = \mu^* + \beta_1^* X_{1i} + \varepsilon_i^* \tag{1.5}$$

   It will be clear that two persons having the same θ will usually not have the same estimate, because of the estimation error. This means that in (1.5) we introduce an extra source of variation, the estimation error, and the consequence will be that

$$\sigma^2(\varepsilon^*) > \sigma^2(\varepsilon)$$

   Which means that the residual ε_i^* is now a mixture of the estimation error and the true deviation from the subpopulation mean. Moreover the regression coefficients μ and β_1 in (1.3) will in general differ from the coefficients μ^* and β_1^* in (1.5). How big the difference will be depends on which estimator of θ is used.

2. The model (1.1) assumes that the variable Y_i can be meaningfully compared across students. In practice this will mean that all students have made the same test. It is possible to adapt the model a bit such that it can be used with different tests, but in such a case one will have to show that all different tests measure the same concept. If one uses (1.3), and one has done a good calibration (showing that all items measure the same θ), then the model can be used also in case different students have made different tests. It is even perfectly justified to apply the model if all students have had a different test consisting of one item. Notice by the way that different tests may be used if (1.5) is used as a model, but in the case of one item tests, the model will produce gross errors.

Summarizing then: use of model (1.3) is to be preferred because the residual is not contaminated by estimation errors of individual θ-values and because different tests, even tests consisting of one item, can be used to estimate the regression parameters and the variance σ^2. It does not follow from this that using just one item response per student will give the same results as using data where students have answered several items. We will come back to this problem in the section about statistics.

1.2.2 Dummy coding

To have a good understanding of regression analysis, it is important to make a distinction between two kinds of regressors (or independent variables). The regressors we discuss in the present section are measurements at the nominal level (qualitative variables). In the next section some words will be devoted to continuous regressors.

Measurement on a nominal level means that the categories of the variable have the role of identification labels, and that no intrinsic numerical relation exists between these categories. The standard example of such a variable is gender, which can take two different values: 'boy' and 'girl'. Even if we recode these values to numerical values '0' and '1', it does not follow in any way that girls are in any respect higher valued than boys. In particular it does not imply that girls do or will do better on the dependent variable than boys. So the codes 0 and 1 are just arbitrary numerical codes, and the conclusions of our analysis will not change if we use another coding. We could have reversed the coding (0 for the girls and 1 for the boys), or we could have used quite different values for the numerical coding, for example, 3 for the boys and 11 for the girls, as long as we do not use the same value for boys and girls (because then, the distinction between boys and girls is lost.) The advantage of choosing a coding of 0 and 1 is that the regression coefficients have a nice and easy interpretation: the additive coefficient (μ in equations (1.1) and (1.3)) is the expected value of the dependent variable in the subpopulation with the reference category 0, and the coefficient β_1 is the difference between the expected values of the subpopulation with code 1 and the reference population.

In cases the regressor variable is nominal but has more than two categories, we cannot suffice with a single numerical regressor; instead we need $m-1$ regressors $(X_1, X_2, \ldots, X_{m-1})$ if the nominal regressor has m categories. As an example, assume that we want to estimate the covariation between the test score (or the underlying latent ability) and the instruction method which has been used in the schools of the student. Assume further that in the country $m = 4$ different methods for mathematics instruction

have been used, which are called 'A', 'B', 'C' and 'D'. To estimate the influence of each method, we need $m - 1 = 3$ numerical regressors, X_1, X_2 and X_3 which can be constructed in the following way:

method	X_1	X_2	X_3
'A'	0	0	0
'B'	1	0	0
'C'	0	1	0
'D'	0	0	1

(1.6)

The regression equation, similar to (1.3), can then be written as

$$\theta_i = \mu + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

and using (1.6) we find

$$\begin{aligned} E(\theta|\text{method} = \text{'A'}) &= E(\theta|X_1 = 0, X_2 = 0, X_3 = 0) = \mu \\ E(\theta|\text{method} = \text{'B'}) &= E(\theta|X_1 = 1, X_2 = 0, X_3 = 0) = \mu + \beta_1 \\ E(\theta|\text{method} = \text{'C'}) &= E(\theta|X_1 = 0, X_2 = 1, X_3 = 0) = \mu + \beta_2 \\ E(\theta|\text{method} = \text{'D'}) &= E(\theta|X_1 = 0, X_2 = 0, X_3 = 1) = \mu + \beta_3 \end{aligned} \quad (1.7)$$

and, again we see that, for example, β_2 is the difference in expected ability between students having used method 'C' and method 'A', the reference category.

This way of coding categorical variables is called dummy coding. Of course, to give the correct interpretation we must not forget how we applied the coding, or in the example, the table in (1.6) must not be forgotten.

1.2.3 Continuous regressors

As an example, we take age as an explanatory variable (regressor) of the dependent variable. We assume that age is expressed in months, and that the average age in the population we are interested in is about twelve and a half years or 150 months. We denote the variable 'age' as X_1. A simple model is the following:

$$\theta_i = \mu + \beta_1 X_{1i} + \varepsilon_i$$

which is formally the same as (1.3), but now the variable X_1 is continuous. How can the regression coefficients now be interpreted? If we insert 0 for the variable X_1 we find that

$$E(\theta|X_1 = 0) = \mu$$

which claims that the average ability of children at their birth equals μ, which of course is ridiculous. For the interpretation of β_1, let's take two subpopulations: in the first subpopulation the age is 161 months, in the second subpopulation the age is one month less: 160 months. We compute the average ability in both subpopulations and then take the difference between the two averages:

$$\begin{aligned} E(\theta|X_1 = 161) &= \mu + \beta_1 \times 161 \\ E(\theta|X_1 = 160) &= \mu + \beta_1 \times 160 \end{aligned}$$

giving as difference

$$E(\theta|X_1 = 161) - E(\theta|X_1 = 160) = \beta_1$$

It is not difficult to see that we would have found the same difference if we had used ages of 154 and 153 respectively, or whatever ages which differ exactly one month. This gives us a nice interpretation of β_1: it is the average increase (growth) in ability per unit of age (month), or in other words, the model says that with every additional month of age, the average ability increases with the value β_1, independently of where we take this difference on the age scale. If we make a graph where age is plotted against average ability, this graph will be a straight line with slope equal to β_1. Therefore we say that the model is a linear model, and β_1 is sometimes called the slope parameter.

We add three comments to this model

1. Speaking of ability 'at birth' does not make much sense; so the parameter μ does not have a nice interpretation. To give it a nice interpretation one sometimes transforms the regressor variable: instead of taking the absolute age, one takes the deviation of the age to some reference age, which is usually a good representative value of the age of the population one is interested in. Here this could be 150 months. An the regression equation becomes

   $$\theta_i = \mu + \beta_1(X_{1i} - 150) + \varepsilon_i$$

   so that μ is now the average ability of students of 150 months old.

2. The preceding model is of course completely equivalent to the first one. A relevant critique is that the model is linear. If the total age span in the population is rather moderate (say one year or so) this model might be a good approximation to reality, but if the age span is large (say more than five years) then we know from developmental psychology

that the average ability is not developing in a linear way, but flattens off at older ages. So the graph giving the relation between age and average ability should not be a straight line but a curve which flattens as age increases. Sometimes one tries to approximate such a curve by a polynomial of degree two or higher. If one chooses degree two, one assumes that this curve can reasonably well be approximated by a parabola, which means that the relation between ability and age should also take the square of the age into account. So a suitable model for such a case might be

$$\theta_i = \mu + \beta_1(X_{1i} - 150) + \beta_2(X_{1i} - 150)^2 + \varepsilon_i$$

It is important to notice that we use here only one regressor (age), but we use it in a polynomial equation of degree two, and therefore we need the two regression coefficients β_1 and β_2. Such models are sometimes indicated as polynomial regression.

3. The program SAUL does not take continuous regressors. So if we want to use age as a regressor we will have to make it a categorical variable. Usually we do this by defining a number of intervals, say m, and then using $m - 1$ dummy variables as in the case of nominal variables. In principle there is no objection to take a large number for m (for example, by using intervals of one month), but in doing so we will need to estimate a large number of regression parameters, each based on relatively few observations, and as a consequence these estimates will have a large standard error, i.e. they will not be very stable. Therefore it is advisable to be as parsimonious as possible on the number of dummy background variables.

1.2.4 Simpson's paradox

Until now we considered only models with one single regressor, but in most applications we are interested in more than one regressor, because even in elementary approaches, we usually will be convinced that the variation in the dependent variable can be explained by more than one regressor variable. Although extension of the regression model to more than one regressor is very simple, the interpretation of the regression coefficients may become problematic and give rise to a number of wrong interpretations.

We will use an example which is partly realistic. In the Flemish part of Belgium the Catholic educational system for boys and girls in primary and secondary education was completely separated until some years ago, while the public system has always been integrated. But the catholic system is

very dominant and comprises the large majority of students. Since a number of years the separation of boys and girls has been abolished, such that every student has access now to every school. But this does not mean that from one moment to the other the distribution of gender in all schools is even. Especially in secondary education, former girl's schools still have a large majority of girls and vice versa in former boy's schools.

Let us assume that there are basically two methods of instruction used in the Flemish schools, method 'A' and method 'B' say, and let us assume that former girl's schools have a preference for method 'A' and former boy's schools have a preference for method 'B'. As a consequence of the unbalanced distribution of gender across schools, method 'A' will be used primarily for the instruction of girls and method 'B' for boys. Suppose the contingency table (the bivariate distribution) of the two variables is as follows

	boys	girls	total
'A'	0.07	0.38	0.45
'B'	0.43	0.12	0.55
total	0.5	0.5	1

(1.8)

The numbers in the table are proportions. This means that 50% of the students are boys and 50% are girls, and that the two methods are not equally popular: method 'B' is used more often than method 'A'. But the most important feature of the table is the fact that the two variables are not independent: of the girls population $100 \times 0.38/0.5 = 76\%$ use method 'A', while this is only the case for $100 \times 0.07/0.5 = 14\%$ of the boys. It is precisely this lack of independence of the background variables which causes the problems in the interpretation, as we will see next.

To set up the model we define two dummy variables: X_1 for gender ('boy' as reference category) and X_2 for method (method 'A' as reference category). To make the distinction between the two background variables clear, we will use the greek letter α_1 as symbol for the regression coefficient of the gender variable and β_1 for the coefficient of the method variable. The regression equation is then given by

$$\theta_i = \mu + \alpha_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i \tag{1.9}$$

Table (1.8) indicates that the total population is partitioned into four subpopulations. Using the definition of the dummy variables, we can compute the expected ability in each of the four subpopulations. Theses expected values are given in the next table:

	boys	girls
'A'	μ	$\mu + \alpha_1$
'B'	$\mu + \beta_1$	$\mu + \alpha_1 + \beta_1$

(1.10)

If we subtract the first row from the second row, the difference is β_1 for the boys column as well as for the girls column. So we see that the effect of method 'B' (as compared to method 'A') is β_1 or the method effect for short. Likewise, if we subtract the first column from the second, the difference is α_1 in both rows. So α_1 is the effect of the category 'girl' (as compared to the category 'boys'), or the gender effect for short.

We can choose numbers for the parameters in order to make the table a bit more concrete. Suppose $\mu = 100$, $\alpha_1 = 2$ and $\beta_1 = 5$. For this choice of parameter values we find the expected abilities in the four subpopulations being:

	boys	girls
'A'	100	102
'B'	105	107

(1.11)

and we can conclude that girls score on the average 2 units higher than boys, irrespective of the method used, and that instruction with method 'B' gives on the average a score which is 5 points higher than instruction with method 'A', irrespective of gender.

In the two subpopulations, students having had method 'A' and 'B' respectively, girls do better than boys. One could be tempted to conclude that therefore in the whole population the average of the girls will be higher than the average of the boys. But this conclusion is not correct, as can easily be shown by combining the tables (1.8) and (1.11). In the girl's population, $100 \times 0.38/0.5 = 76\%$ used method 'A' and 24% used method 'B'. The weighted average in the total girls population is therefore

$$E(\theta|X_1 = 1) = 0.76 \times 102 + 0.24 \times 107 = 103.2 \tag{1.12}$$

and by a similar reasoning we find that for the boys

$$E(\theta|X_1 = 0) = 0.14 \times 100 + 0.86 \times 105 = 104.3 \tag{1.13}$$

This paradoxical finding that girls perform worse than boys in general, while they do better in each subpopulation is known as Simpson's paradox. It is clearly caused by the fact that in table (1.8) gender and method are not independent, which means that in the subpopulations of girls and boys, the variable method does not have the same distribution. This paradox is virtually always at work in survey research because the regressors we use are almost never independent. We give some additional comments on this paradox.

1. The difference between the overall mean of girls and boys ($103.2 - 104.3 = -1.1$) is the value of the regression coefficient for gender we

would have found if we had done the regression analysis with only gender as regressor. So with only gender, we find -1.1 as gender effect, and with an additional regressor (method) we find $+2$ as gender effect. This is a quite dramatic change (because the algebraic sign reverses), but in general we speak of Simpson's paradox if the regression coefficient of a regressor changes if one or more other regressors are added to the model. So this paradox is almost always present; the only case where it is not present is the case where all regressors are mutually independent, and this usually does not occur in survey research. In the example used above we can also find out what the method effect would have been if we used it as the only regressor, leaving out gender. We find

$$E(\theta|X_2 = 1) = \frac{0.43}{0.55} \times 105 + \frac{0.12}{0.55} \times 107 = 105.44$$

and

$$E(\theta|X_2 = 0) = \frac{0.07}{0.45} \times 100 + \frac{0.38}{0.45} \times 102 = 101.69$$

giving a method effect of $105.44 - 101.69 = 3.75$, which is definitely less than the effect of 5 units that we have if gender is added as a second regressor. Finding two different values for the methods effect is also an example of Simpson's paradox.

2. The presence of this paradox is at the heart of all problems analysts encounter when they have to report results to the public or to policy makers. In such reports answers are requested to simple questions such as: "do girls perform better than boys, or not?" The answer is not simple, as is shown in the example. From (1.12) and (1.13) we know that in the general population the average ability of girls is below that of boys, but at the same time, we know (at least partially) how to explain this difference: the boys have used in great majority the best of the two available methods, and the majority of the girls has used the worst method. So we might argue that the comparison is not fair, and try to correct it by adding 'method' as a second regressor. As a result we find that if we control for the variable method, girls do better than boys. So what we would like to report is that if boys and girls would have used the same method (or more exactly if gender and method would have been independent in the population) we would find that girls do better than boys. Explaining this kind of subtle differences in the results proves to be quite hard in practice.

3. The preceding explanation can help us in formulating a general rule for the interpretation of the results of a regression analysis with multiple

regressors which are binary dummy variables. We first give the rule in the form of a formula. Suppose we use k regressors, and we concentrate on one of the regressors, regressor j say. Next we consider two subpopulations. In the first subpopulation it holds that the dummy variable $X_j = 1$, and the other dummy variables have some value, which can be chosen arbitrarily. In the second population $X_j = 0$, and the other dummy variables have the same value as in the first subpopulation. Then the regression coefficient β_j is the difference between the average ability in the first and the second subpopulation:

$$\begin{aligned} & E(\theta|X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, X_j = 1, \ldots, X_k = x_k) \\ - \quad & E(\theta|X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, X_j = 0, \ldots, X_k = x_k) = \beta_j \end{aligned}$$

We can formulate this a little bit differently: β_j the difference in average ability between a subpopulation having a 1 on the dummy variable X_j and the subpopulation having the reference category, where all other regressor variables have the same value in both populations, or in short all other things being equal (which is often expressed in Latin: ceteris paribus). In the example, we say that the gender effect (or in the notation we used in the example: the regression coefficient α_1) is the difference between average ability of girls and boys if the other regressor has the same value, or more concrete, if boys and girls use the same method. Another expression which is used in regression literature is to say this is the gender effect if we **control** for method.

4. The example of this section is also a nice starting point to warn us for unjustified optimism. If we do a regression analysis with only gender, we find a negative effect for girls; if we add method as regressor, we find a positive effect for girls, but we don't know if there isn't a third variable that would cause, when added, the girls effect become negative again or positive but larger than the effect we have hitherto. If there is such a variable which we do not include in the model, we are said to make a **specification error**, but since we are never sure wether such a variable exists, we can never conclude on the basis of regression analyses that the regression coefficient for gender expresses a difference between boys and girls that is completely due to gender and not (partly) to another variable that is confounded with gender. The only thing which we can deduce from the example given above is that the difference between girls and boys of 2 points is **not** due to differences in methods used, because we have controlled for this difference. So the only conclusions from a regression analysis of which we can be sure, are formulated in the negative.

1.2.5 Interactions

In the previous section, an example was discussed with two regressors - gender and method - and the model essentially said that the effect of method was the same for boys and girls and similarly, that the effect of gender was the same in both methods. Such a model is nice, and leads to relatively easy interpretations. For example, one could say, that the average performance of girls is two points more than the average performance of boys, and this holds irrespective of the method. But reality may be more complex: it might be that the difference between boys and girls is 2 points if method 'A' is used, but more than or less than 2 points if method 'B' is used. So the answer to the question: "do girls perform better than boys?" may need a nuance, for example: "it depends: if method 'A' is used, girls obtain on the average 2 points more, but if method 'B' is used, their score is 1 point less than that of the boys." We can make this situation visible in the following table of expected abilities in the four subpopulations (compare to table (1.11)):

	boys	girls
'A'	100	102
'B'	105	104

(1.14)

Comparing tables (1.11) and (1.14) we see a marked difference: in (1.11), the difference between second and first row is 5, and the difference between second and first column is 2. In table (1.14) this is no longer the case, and we need to give a distinction: the difference between method 'B' and 'A' is 5 points in the subpopulation of boys, but it is only 2 points in the subpopulation of girls.

If such a distinction is necessary, we say that there is an interaction between the variables gender and method. It is very important to distinguish the notion of interaction and the notion of lack of independence between the two regressors. Dependence or independence between gender and method can be decided upon without looking at the dependent variable. (This means, we can construct table (1.8) without knowing anything about the ability we measure by the test; it is the kind of information that could be provided by the national office of statistics.) If we speak about interaction we mention the joint effect of two (or more variables) on the dependent variable. If we say that there is interaction, we mean that the joint effect of gender and method on ability cannot be explained by the simple sum of the main effect of gender plus the main effect of method. We need something more, and this something is called the interaction effect.

The modeling of interaction effects when we use dummy variables is quite easy. In the case of the example, we define a third dummy variable which is

just the product of the two other dummy variables:

$$X_3 = X_1 \times X_2$$

from which it immediately follows that

$$X_3 = \begin{cases} 1 \text{ for girls under method 'B'} \\ \\ 0 \text{ otherwise} \end{cases}$$

and we add this regressor X_3 to the regression equation, giving

$$\theta_i = \mu + \alpha_1 X_{1i} + \beta_1 X_{2i} + (\alpha\beta)_{11} X_{3i} + \varepsilon_i$$

Here we use some funny notation for a regression coefficient. The symbol $\alpha\beta$ placed between parentheses does not mean a product of two coefficients, but should be read as a single symbol. Its notation immediately says that it denotes the coefficient of an interaction of the variable for which we used the symbol α (gender) and β (method) respectively. The reason why we use a double subscript (11) will be clear shortly. It is easy to check that we can reproduce table (1.14) with $\mu = 100$, $\alpha_1 = 2$, $\beta_1 = 5$ (which is the same as before) and $(\alpha\beta)_{11} = -3$

We end this discussion with an indication on how to model interaction if nominal variables with more than two categories are used. Suppose that in the example above, three methods ('A', 'B' and 'C') were used instead of two. To model main effects of gender and methods as well as the interaction between gender and method we need five dummy variables, defined as follows:

$$\begin{aligned} X_{1i} &= 1 \text{ if student } i \text{ is a girl} \\ X_{2i} &= 1 \text{ if student } i \text{ used method 'B'} \\ X_{3i} &= 1 \text{ if student } i \text{ used method 'C'} \\ X_{4i} &= X_{1i} \times X_{2i} \text{ (interaction)} \\ X_{5i} &= X_{1i} \times X_{3i} \text{ (interaction)} \end{aligned}$$

In the notation introduced earlier, we will then write the regression equation as

$$\theta_i = \mu + \alpha_1 X_{1i} + \beta_1 X_{2i} + \beta_2 X_{3i} + (\alpha\beta)_{11} X_{4i} + (\alpha\beta)_{12} X_{5i} + \varepsilon_i$$

Here we see that we need two regression coefficients (β_1 and β_2) to express the method effect: one for comparing method 'B' to the reference category (method 'A') and one to compare method 'C' to the reference category. (The

methods 'C' and 'B' can be compared by taking the difference $\beta_2 - \beta_1$. Similarly, we need two regression coefficients to model the interaction between gender and method: one which expresses the special effect of the combination girl and method 'B' ($(\alpha\beta)_{11}$) and one for the combination girl and method 'C' ($(\alpha\beta)_{12}$).

In general if we want to model the interaction between two categorical variables having m and p categories respectively, we will need to add $(m-1) \times (p-1)$ extra dummy variables and regression coefficients. Fortunately, the program SAUL automatically generates all dummy variables for main effects and interactions that are wanted.

1.3 The Statistics of Regression Analysis

1.3.1 Estimates and standard errors

In the examples of the preceding section, we always used a specific number for the regression coefficients, suggesting that these coefficients are known (at least to somebody). In practice, however, these coefficients are not known and must be estimated from the data. The resulting values, the estimates, are not the real coefficients, but the real coefficients plus an estimation error. To have an idea of the size of the estimation error, the standard error is estimated as well as the coefficient itself. The standard error as well as the estimate are used to construct statistical tests. This will be illustrated with two examples.

Example 1. Suppose we do a regression analysis with one regressor, gender, where 'boys' as before is considered to be the reference category. The regression coefficient β_1 is the difference in average ability between girls and boys. As a result of the regression analysis we find the estimate $\widehat{\beta}_1$, which is not the true β_1. If the estimate differs from zero, it might be the case that the true coefficient does not differ from zero. Therefore we may wish to test the (null) hypothesis that $\beta_1 = 0$. The test statistic is, just as in a common t-test given by

$$t = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$$

where $SE(\widehat{\beta}_1)$ stands for the estimated standard error. Theoretically the value found for t must be compared with the critical value of the t-distribution, but in practice the standard normal distribution will be a good alternative. (That is why the symbol t is often replaced by

the symbol z.) So if t is larger than 1.96 in absolute value and we are testing two-tailed at the 5% significance level, the null hypothesis (of no difference between boys and girls) will be rejected. Otherwise the null hypothesis is not rejected. It is very important that one interprets this latter statement correctly: it does not mean that we have 'proved' in some sense that there is no difference; it only says that we have no evidence that there is a difference. And these two things are not quite the same. (A good, although not perfect, comparison is in the legal area: suppose your wallet is stolen and you suspect x of having stolen it. If you cannot prove x has done it, you will not necessarily conclude that he did not steal it. Maybe you will search for further evidence; otherwise you have to keep your mouth shut. On the other hand, if you can prove that x did steal your wallet, you can start whatever legal action against him.)

Example 2. Suppose we do a regression analysis using the variable 'method' as a regressor, where 'method' refers to an instruction method, which can be method A, method B or method C. As explained above, we will have to create two dummy variables. Suppose we choose method A as the reference category, give the dummy X_1 a value of 1 in case of method B and the dummy X_2 a 1 in case of method C. Now we have to estimate two regression coefficients β_1 and β_2. Suppose we find the following estimates and standard errors:

	estimate	stand. error	t
β_1	0.80	0.34	2.35
β_2	1.20	0.37	3.24

(1.15)

The two t-values indicate that both estimates differ significantly from zero, (and since they are positive) indicate that the average ability under methods B and C are both larger than under method A. Or more generally, the testing of the significance of a regression coefficient tests always the difference between the average ability of a certain category and the reference category. But in the present case, we may ask one more interesting question. It can be seen from the table that the effect of method C is estimated to be larger than the effect of method B. But also here, we have to ask ourselves if this difference of $1.20-0.80 = 0.40$ units is statistically significant. The test statistic is

$$t = \frac{\left(\widehat{\beta}_1 - \widehat{\beta}_2\right) - 0}{SE\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)} = \frac{\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)}{SE\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)}$$

Of, course the difference between the two estimates can easily be computed from Table (1.15), but the standard error of this difference cannot be computed exclusively from the numbers given in the table, because the two estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are correlated, and to compute the standard error of the difference we need an estimate of this correlation. In the program SAUL this correlation is computed, but not printed in the output. The pairwise differences of all regression coefficients, however, are tested in the way described here, and the corresponding t-values are given in the output. So, if a variable has five categories, four tests on the four regression coefficients are computed, and six tests on the pairwise differences are computed as well.

1.3.2 The size of the standard errors

The size of the standard errors is influenced by a number of factors. In this subsection we give three of them which are common to all regression analyses and a fourth factor which is specific to latent regression analysis, i.e., a regression analysis where the dependent variable is not observed directly, but only indirectly by a number of discrete response variables (the answers to the items). We discuss these factors in turn.

1. The sample size (n). This is the most important factor. The rule which applies here is that the standard error decreases proportionally with the square root of the sample size. If one carries out an analysis on a sample, and one finds standard errors which are twice as large as one would wish, one can halve them by increasing the sample to **four** times the size of the present sample: the standard error will then be reduced by a factor$\sqrt{4} = 2$. Notice that this rule is exact, but that in practice one will not find this exact relationship, because one cannot compute the exact standard error in practice, but only estimate it. Therefore the above rule will only apply approximately to estimated standard errors.

2. The balance of the design. Suppose we do a regression analysis using a single binary regressor such as gender, and suppose the total number of observations is fixed at some value, 1000 say. One could do a regression analysis with 500 boys and 500 girls, but one could also use a sample of 800 boys and 200 girls. Both are acceptable (as long as the sampled boys are a random sample from the population of boys, and similarly for the girls) in the sense that the estimate of the regression coefficient will not be distorted systematically. (It does not matter whether the proportion of boys in the sample does or does not correspond to the

proportion in the population.) But it does matter for the accuracy of the estimation: the standard error will be smallest if an equal number of boys and girls is used. There exists a good rule to compute the loss of efficiency when one uses unbalanced designs. The rule is that the standard error in an unbalanced design is approximately equal to the standard error in a balanced design where the number of observations in each category is equal to the harmonic mean of the numbers one uses actually in the unbalanced design. In the example the harmonic mean is given by

$$H = \frac{2}{\frac{1}{200} + \frac{1}{800}} = 320.$$

This means that we would obtain an equally accurate estimate of the regression parameter with a sample of 320 girls and 320 boys, as we obtain now with 200 girls and 800 boys. (The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals; if the numbers are not all equal, the harmonic mean is smaller than the arithmetic mean.) It is important to take this aspect into account when planning a survey. Suppose the total population consists of two important subgroups (e.g. nationalities), one of which is numerically much more important than the other. If one plans to treat group membership as a regressor, the estimate of the regression coefficient will be more accurate if one samples an equal number from each group, than when one samples proportionally. Such a sampling scheme is called stratified sampling or stratification.

3. Correlation between the regressors. In the section on Simpson's paradox it was argued that the lack of independence between the regressors may cause problems in the interpretation of the regression parameters. Dependence between the regressors has also an influence on the standard errors of the estimates of the regression coefficients. In general one can state that the stronger the dependence, the larger the standard errors will be. In the limit case, where there is total dependence, the standard error goes to infinity, meaning that the regression coefficients cannot be estimated. An example of total dependence is the following: Suppose all boys were instructed with instruction method A and all girls with instruction method B. In such a case it is impossible to estimate the effect of gender and method. One says that the variables gender and method are confounded. To have an idea of the effect of the dependence of the regressors on the accuracy of the estimate, an artificial example has been investigated. Seven cases are studied using

the following design

	boys	girls	total
meth. A	n_{11}	n_{12}	500
meth. B	n_{21}	n_{22}	500
total	500	500	1000

If $n_{11} = 250$, gender and method are independent (If one of the cells is fixed, the others are fixed as well because the margins of the table are fixed). For the seven cases the value of n_{11} is consecutively equal to 250, 290, 330, 370, 410, 450 and 490. In the last case there is no confounding, but the dependence between method and gender is very strong. For each of the seven cases an artificial data set is created using 30 items, and with true effects of 0.5 for gender and 0.8 for method. The regression coefficients and their standard errors are estimated with the program SAUL. It turned out that in all cases the two standard errors were vitually equal. In Figure 1 they are plotted against the value of n_{11}.

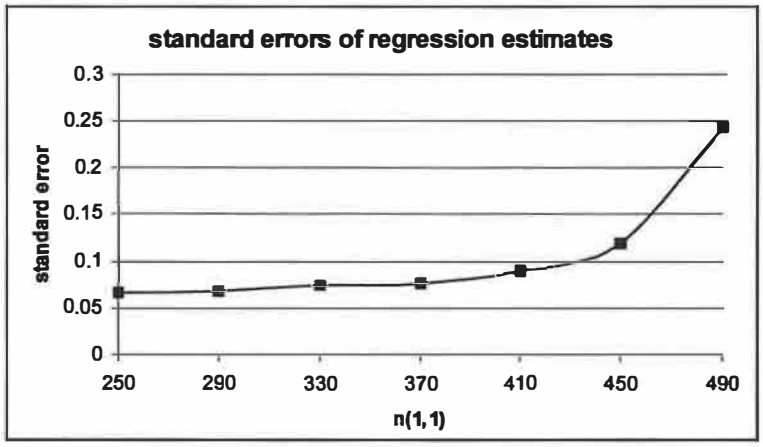

Figure 1

It is clearly seen that the standard error increases with increasing dependency, but, fortunately, the effect becomes dramatic only if the dependency is very strong. In the strongest case the standard error is four times as large as in the independence case. Moreover, the Figure suggests clearly that the standard error will grow very fast if we make the dependence still stronger. Therefore, it is always advisable to check the design for strong dependencies, and to remove regressors which are too tightly connected with other regressors. In practice this is not always easy, because it may happen that one regressor or a combination

of several regressors is strongly dependent on a combination of other regressors. There exist special multivariate techniques to check for dependencies, but discussion of them is beyond the scope of this manual. The program SAUL, however, does detect cases where regressors are confounded.

4. The fourth factor is the reliability of the dependent variable. The reliability depends primarily on the number of items. (The item parameters play an important role as well, but this will not be discussed here.) It can be expected that with a less reliable dependent variable, the regression parameters will be estimated less accurately than with a more reliable measurement instrument. To illustrate the effect of the reliability on the standard errors, a number of analyses on artificial data were carried out using two regressors with the same effect as in the previous example. For the measurement model, the Rasch model was used with all difficulty parameters being equal. A single data set was created with 50 items, then the calibration was done, and after that a series of regression analyses were carried out using different numbers of items. The results are displayed in Figure 2.

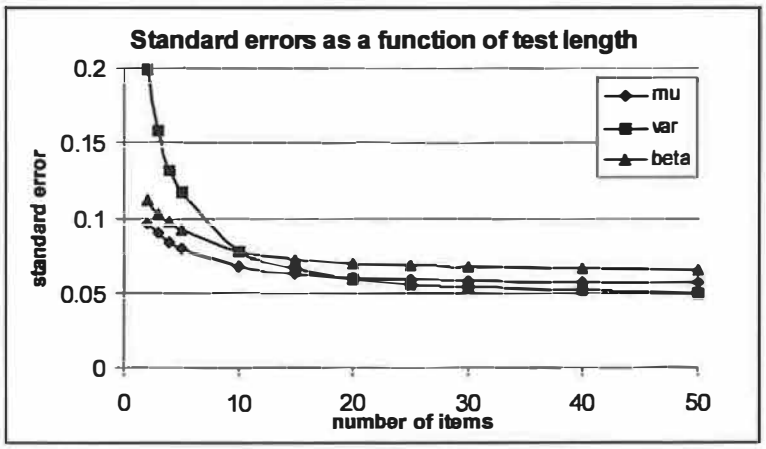

Figure 2

The standard errors for the two regression parameters were equal, but in the Figure the standard errors for the estimates of μ (mu) and σ^2 (var) are displayed as well. We give some comments on this Figure.

(a) All three curves seem to level off if the number of items increases towards a limit (which may different per parameter) but which certainly is not zero. This is quite natural: if the number of

items goes to infinity, we can estimate the θ-value of each person without error, but even in this case, this does not mean that we can estimate the regression coefficients without error, because we do have only a finite number of persons in our sample.

(b) In the example used ($n = 1000$, the two regressors are orthogonal and perfectly balanced), we see only serious changes in the standard error if the number of items is less than about 10 to 15. In the example a test with 15 items has a reliability of about 0.80. The practical implication of this is that we cannot gain very much in accuracy of estimation by making the test more reliable (by lengthening it) The real gain in accuracy has to come from increasing the sample size.

(c) For short tests (say no more than five items), the loss in accuracy for the estimate of the residual variance is much more dramatic than for the other parameters.

Chapter 2

The programs SLIN, SLPRE and SAUL

2.1 Introduction

As argued in the introduction of the previous chapter, a latent regression model is a structural model which takes an (accepted) measurement model for granted. Practically, this means that to apply latent regression, the item parameters must be known, and their values must be made available to the regression program. The program SAUL (which does the actual computations) assumes that the measurement model used is OPLM. So in order to run SAUL the item parameters (as well as the discrimination indices) must be available in a *.PAR file, created by OPCML.

To run SAUL, however, two other programs have to be run beforehand: SLIN and SLPRE. After SAUL has finished its computations, the results of the latent regression analysis are available to the user, but sometimes another result is wanted in survey research. As was explained in the first chapter, regression coefficients can be interpreted as differences in conditional means: what is the difference in performance between boys and girls if all other regressors are kept constant. But sometimes one may be interested in marginal means: what is the average performance of boys and girls in the target population. Estimates of marginal means cannot be computed from a regression analysis without further assumptions. These assumptions will be discussed at length in Chapter 3, and at the same time a fourth program of the package, RAPPORT, will be introduced. The four programs OPIN, SLPRE, SAUL and RAPPORT, however, are tightly related by a system of files. In Figure 3 a general overview of the system is given in the form of a flow chart. All programs and files will be discussed in detail in this and the

next chapter.

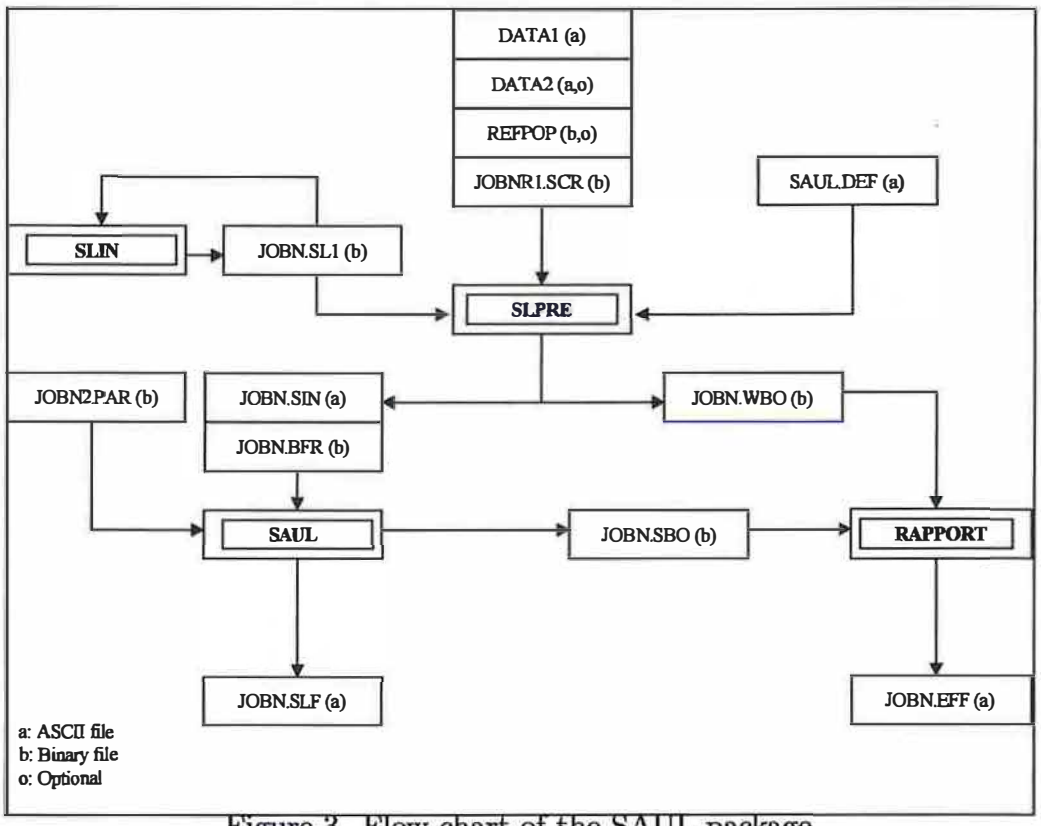

Figure 3. Flow chart of the SAUL package

We start with a short discussion of the three programs SLIN, SLPRE and SAUL.

1. SLIN. This is an interactive program which asks for the specifications of the analysis. This specifications are to be input via the keyboard and are saved in a binary file having the extension '.SLI'. The program can be used also to edit an existing *.SLI file. The program will be discussed in the next section.

2. SLPRE. This program processes the data (item responses and background variables) to compute tables of statistics according to the specifications of the user (and stored on a *.SLI file) and creates a number of files. Two of these files are the input for the program SAUL. Another file is prepared for a special application (the program RAPPORT; see next chapter). The program SLPRE is discussed in section 3 of the present chapter.

3. SAUL estimates the coefficients of the specified regression model and creates two output files: a text file (with extension '.SLF') which contains a full report on the results of the regression analysis and a binary file which is the input for the program RAPPORT. The program is discussed in detail in section 4 of this chapter.

All files created by the three programs have the same file name. The files which must be available to run the programs can have arbitrary names. As a generic name for the file name of the output files, the name JOBN will be used in this manual. There are three (optionally four) input files with arbitrary names, but two of them have compulsory extensions. For these files we will also use generic names in this manual. Here is a short overview.

1. JOBN1.SCR. This is a *.SCR file created by the program OPIN. Some of the information in this file is used, other information is ignored. The information that is used is the specific booklet structure of the item responses, the set status of the items (indicating which items are included in or excluded from the analysis) and the format of the data file. The name of the data file in the *.SCR file, however, is ignored, as well as all specifications which are specific for an OPLM analysis.

2. JOBN2.PAR. This is a (binary) parameter file created by the program OPCML. It must contain item parameter estimates of all items which have the 'on' status in the JOBN1.SCR file, but it may have more. The maximum score of these items must match in both files; the value of the discrimination index is read from the JOBN2.PAR file, the values in the JOBN1.SCR file are ignored. In standard applications the names JOBN1 and JOBN2 will be equal, but this is not compulsory.

3. The data file. In the present manual this file will be designated as DATA1. It must contain the booklet identification and the item responses in a format which corresponds to the FORMAT statement in the JOBN1.SCR file. Besides that, this file contains all or some of the regressor variables that can be used in the latent regression analysis. The name used for the DATA1 file is completely arbitrary, but it must be a legal DOS file name. If not all regressors are contained in this file, the remaining ones must be supplied in a second data file, which is discussed next.

4. The second data file, with generic name DATA2. In many instances the data in the sample are hierarchically organized, such as for example, students within schools. If one wants to use a regressor which

has the same value for each student within a school, one can put this value on each student record in the data file DATA1, but one can also prepare a separate data file which contains only the school variables. The program SLPRE will merge the two data files. For this merging to be possible, the two files DATA1 and DATA2 must contain an identification field for the school. Both data files must be sorted on the identification field in ascending order. Using this mechanism, it is also possible to submit all student regressor variables via a separate file. In such a case the two files DATA1 and DATA2 must have a unique identification field per student (i.e. per record in the DATA1 data file.) If all regressor values are contained in the DATA1 file, no DATA2 file is specified. The name used for the DATA2 file is completely arbitrary but it must be a legal DOS file name.

As to the location of the files, the following rule applies: all required input files must be in the same directory (folder), to be specified via the interactive program SLIN, and all output files, with the exception of the *.SLI file are placed in the same directory (folder) as well. If no directory name is specified, the default directory is used. The location of the JOBN.SLI file (created by SLIN) is located in a directory specified by the user in the Save Menu of the program SLIN.

2.2 The program SLIN

- Purpose: user specifications for a latent regression analysis.
- Components: an initial menu (Init Menu), a Save Menu and one application screen.
- Input files: none (everything is input via the keyboard) or an existing *.SLI file (to be edited).
- Output file: a *.SLI file or nothing.
- Command: SLIN <CR> or SLIN [path]JOBN[.SLI] <CR>

The program SLIN, as well as other programs in the package use a number of auxiliary files which must be available at run time. A complete list of these files, and instructions how to install them is given in the Appendix.

2.2.1 The Init Menu

This menu appears on the screen if the command SLIN <CR> is used. It has five options, which may be reached using the cursor, or typing the highlighted key (hotkey). Choices are confirmed by tapping the Return key (<CR>).

- Init. By choosing this field, the program switches immediately to the screen which has to be filled out completely via the keyboard.
- Directory: Choosing this option allows to specify the directory where the existing *.SLI file one wants to edit is to be found. Leaving this field blank specifies the current directory.
- File: If there exist one or more *.SLI files in the directory specified in the preceding field, the name of one of them will be displayed here. Typing <CR> moves the cursor to the name field, where the name can be changed. Typing *.* or *.SLI will produce a list of the existing *.SLI files, from which one may be chosen with <CR>. Typing <CR> when on the name field makes SLIN move to the screen, where the contents of the chosen *.SLI file are displayed and are ready to be edited.
- Start File: has the same function as file. It can be reserved for a general template of a *.SLI file.
- Quit: returns control immediately to DOS.

2.2.2 The Application Screen

The screen contains four components, three of which have to be filled out completely, and one which is optional (the Komment field). Each component consists of one or more fields, indicated by a name which contains one capital letter. The combination of the Alt key with this letter moves the cursor to this field. The key combination Ctrl-F1 (indicated as ^F1 on the bottom line of the screen itself) gives a screen with general information (in a very condensed style) on how to control cursor movements in the screen, and some other general information. The key combination Ctrl-F2 (indicated as ^F2) gives help on the meaning of the field where the cursor resides. Help screens can be left by tapping a key. When filling out a field, confirmation by typing Enter is required. The four components are discussed in turn next.

Files

- dIrectory. Specify the path where the input files are to be found and where all output files (with the exception of the *.SLI file; see the

subsection on the Save Menu) are to be stored. Leaving this field blank selects the current directory.

- Data file: the name of the DATA1 file. Upon confirmation of this field with <CR> the user is prompted for the format of the regressor values on the DATA1 file. Notice that the format for the booklet information and the item answers is read from the format specification in the JOBN1.SCR file. Example: suppose two regressor values are contained in the DATA1 file, one in position 35, and a two digit value in the columns 45 and 46. Then the two following expressions are valid and equivalent formats:

  (T35,I1,T45,I2)
  (34X,I1,9X,I2)

  The example gives the three descriptors which can legally be used in a format statement: 'Tn' means: position the reading head at position n, i.e. read from that position on. This is an absolute tab, not a relative one. 'nX' means: skip n positions. 'In' means: read an integer number of n positions. A repetition factor may be used with the I-descriptor. 7I1 means: read 7 numbers of one position each. Notice that an opening and a closing parenthesis are compulsory. Lower case letters 't', 'x'and 'i' are also allowed. Further comments on this format will be given in the subsection on background variables. It is possible to use the booklet variable or an item response as a regressor. If one wants to do so, booklet number or item response must be part of the list of the background variables (see below), and the format must correctly point to them, in the same way as for the other background variables. In fact, if the booklet number is at the same time a regressor, it is read two times when the program SLPRE reads the data from DATA1: once as booklet number, using the format specification in JOBN1.SCR and once as regressor using the present format.

- High file (optional): the name of data file DATA2. Upon confirmation of this name, the user will be prompted for the format of the background variables which are contained in DATA2. The rules are the same as with DATA1. Upon confirmation of this format the user will be prompted for the identification key which is necessary to merge the two data files. To this end, a small screen appears which asks for three pieces of information: (1) the first element is the first position of the identification string in the data file DATA1; (2) the second element is the first position of the identification string in the data file DATA2 and

(3) the third element is the length of the identification string. (maximum is 10 positions). Notice that the identification string may contain any alphanumeric character, whereas the regressor values must be numeric (and integer).

- Refp file: this file will be discussed in the chapter on the program RAPPORT. If only a regression analysis is wanted, this field may be skipped.

- format: this field cannot be reached directly. An entered format, however, can be made visible (and edited) by putting the cursor on the fields Data file or High file, and typing Ctrl-V. The format field is quit without modification by typing 'Esc'. Changes have to be confirmed with 'Enter'.

- Scr file: the name of the JOBN1.SCR file

- Par file: the name of the JOBN2.PAR file

- jobName: the common name of all output files from SLIN, SLPRE, SAUL and RAPPORT. It must be a legal DOS file name. The generic name in the present manual is JOBN. See for further comments also the section on the Save Menu.

Background variables

- #backGround variables. Enter the number of background variables which will (possibly) be used in a regression analysis. The variables used in an actual regression analysis are specified in the model component, but they must be a subset of the variables specified here.

- Nr: after confirmation of the number of background variables, a list of numbers appears in this field. For each of the background variables three additional fields must be entered (see below). The field Nr itself cannot be reached by the cursor. The first variable has a special status, which is not relevant for the regression analysis, but for the program RAPPORT. It will be discussed in the chapter on the program RAPPORT.

- laBel: This is the label for the regressor variable. The label has a maximum length of 8 characters. These labels will appear in the output of SAUL and RAPPORT.

- daTa: the number of categories the regression variable can take. The program SAUL assumes that category values are numeric and take values 1, 2,.... If this is not the case in the data files DATA1 and DATA2, the user has to recode the values on the data files to consecutive integers 1, 2,.... This can be accomplished using the recode statement, which is discussed in the subsection Komment further down. After confirmation of the number of categories, a subscreen appears, where for each category a label can be entered. While editing a screen, this subscreen can be reached by placing the cursor on the number of categories and typing Ctrl-V. The subscreen is left unaltered by typing 'Esc'; changes must be confirmed with 'Enter'.

- reFp: Usually the value entered here is the same as the value in the daTa field. The program RAPPORT, however, accepts a smaller value. This will be discussed further in the Chapter on RAPPORT. The value entered here has no influence on the regression analysis.

- leVel: for each variable, it must be indicated whether it has to be read from the file DATA1 or DATA2 by entering a 1 or a 2 respectively. The association between the numbered variables, the leVel and the data files is governed by the format statements, specified when the name of the data files are confirmed. The general rule is that the format must point to the variables in the order in which they appear in the variable list for each of the data files separately. The following example shows this for 6 background variables.

Nr	laBel	daTa	leVel	data file	position
1	Stratum	3	2	DATA2	34
2	Gender	2	1	DATA1	12
3	Weight	3	1	DATA1	11
4	Ethnicity	4	1	DATA1	14
5	Age	3	1	DATA1	15
6	Method	4	2	DATA2	32

The column position tells where the value of the regressor variable is to be found in the relevant data file. The formats appropriate for both data files are

DATA1: (T12,I1,T11,I1,T14,2I2)
DATA2: (T34,I1,T32,I1)

Notice that the tab descriptor (T) is suited for moving backwards in the record; the skip descriptor (X) cannot be used for 'back skipping'.

A valid format for DATA1 is also (11X,I1,T11,I1,2X,2I2); T11 cannot be replaced by a skip descriptor.

Model

Specifying the model is done by typing the numbers of all background variables which are to be entered as regressors. The numbers of the variables are separated by a blank or a comma. They may appear in any order. We discuss five examples, which all refer to the table in the preceding section.

- Only main effects. Typing 1 2 6 specifies a regression model with the variables Stratum, Gender and Method as regressors. Only main effects will be estimated. For these three variables, $(3-1)+(2-1)+(4-1) = 6$ regression parameters are to be estimated. Since in every analysis the additive parameter μ and the residual variance σ^2 have to be estimated as well, this yields a total of 8 parameters. The program SAUL accepts analyses with a maximum of 126 parameters to be estimated.

- Main effects and interactions (1). It is also possible to ask for the estimation of interaction effects. Suppose one wants, besides the main effects of the preceding example, to estimate also the interaction effects of Stratum and Gender. This is done by typing [1 2] 6. Notice that it is not possible to estimate interaction effects without at the same time estimating the main effects of the variables involved in the interaction. It is not necessary for the user to compute the interaction dummy variables. This is implicitly done by the program SLPRE. The data files submitted by the user are not altered.

- Main effects and interactions (2). The model specification [1 2] [2 6] makes SAUL to estimate the main effects of the three variables Stratum, Gender and Method, the interaction effects of Stratum and Gender and the interaction effects of Gender and Method. The number of interaction parameters is $(3-1) \times (2-1) = 2$ for [1 2] and $(2-1) \times (4-1) = 3$ for [2 6]. So for this model a total of $8+5 = 13$ parameters have to be estimated. The interactions discussed in this and the preceding example are called first order interactions, because only two variables are involved in each interaction.

- Higher order interactions. The model specification [1 2 6] means that the main effects of the three variables have to be estimated, as well as the three first order interactions and the second order interaction effects of the three regressors jointly. Logically this way of specifying quite

complicated models can be generalized to more than three variables. The present versions of the programs SLIN and SAUL, however, do not allow for higher order interactions. They will be implemented in future versions.

- Empty model. If no model is specified, the regression analysis estimates the additive parameter μ and the variance σ^2 of the latent variable.

Komment - skip - recode - group

The comment field can be reached by typing AltK, and is left by typing 'Esc' ('Enter' will generate a new line).. The field itself serves a double purpose: the comments entered here will be echoed in the output of SAUL. But this field also allows a quite extensive control on the way the background variables are to be treated by the program SLPRE. Therefore three commands are at the disposition of the user: skip, recode and group. These will be discussed in turn. To separate the comments from these commands, a separate line containing three slashes (///) at the first three positions must be entered. All lines in the Komment field up to the triple slash line are considered as comments; all lines following the triple slash are interpreted as commands. If there are no commands, the three slashes are optional. If there are commands but no comments, the triple slashes are compulsory.

- skip. The syntax of the skip command is

  ski[p] BackVar Val1, Val2=Val3,...

  where BackVar is the sequence number of the regressor variable (see the field Nr), Val1 indicates a separate value, and the string Val2=Val3 indicates the range from Val2 through Val3. The values Val1, Val2, etc. designate values as they appear on the data files DATA1 or DATA2. As an example, assume that the variable Age is expressed in months on the data file, and that Age has sequence number 5 in the list of background variables. The command 'skip 5 -100=143,170=999,167' will cause all records where the age is less than 144 or greater than 169 to be skipped as well as the records where the age is precisely 167 months. Since negative values may appear on the data file, the sign '=' is used to indicate ranges.

- recode. The syntax of the recode command is

  rec[ode] BackVar LowVal1,HighVal1=Cat1,...

where BackVar is the sequence number of the regressor variable (see the field Nr), LowVal1,HighVal1 indicates the range from LowVal1 through HighVal1 and Cat1 is the recoded value which applies to this range. We discuss two examples.

- Assume Gender is regressor number 2, and is coded as 0/1 on the data file. The program SLPRE requires that the coding of the regressor values starts at 1. The recode command allows the user to leave the data file as it is and to recode implicitly the values on the data file. The appropriate command is

  recode 2 0,0=1, 1,1=2

- The current version of SAUL does not allow for continuous regressors. If a variable like Age (expressed in months on the data file) is to be used as regressor, the recode command can be used to discretize the continuous Age variable. If Age is the 5th variable, the command can be, for example,

  recode 5 -100,150=1, 151,160=2, 161,999=3

  If these three age categories are effectively to be used in the regression analysis, the number of categories specified in the daTa field must equal 3. The category labels apply to these recoded values.

- group. Using nominal variables as regressors may lead to models with many dummy variables, and sometimes one might wish to use more parsimonious models. This can be accomplished by grouping some categories together into newly defined categories. The group command is suited for this purpose. The syntax is

  gro[up] BackVar LowCat1,HighCat1=NewCat1 "Label1",...

  We discuss two examples.

  - Suppose that the user wants to reduce the number of age categories, as defined in the previous example from 3 to 2, by taking the former categories 2 and 3 together into a single grouped category. Assume moreover, that the category labels in the previous example were 'young', 'middle' and 'old' respectively. The command

    group 5 2,3=2 "med&old"

causes the program SLPRE to group the recoded age variable having values 2 or 3 to be grouped together as category 2. At the same time the label of this new category will be "med&old", and the number of values the age variable can take will be reduced from 3 to 2. The values specified in the daTa field can be left unchanged.

- As in the previous example, but now one wishes to take together the original categories 1 and 2, and give this combined category the label "yng&med". This could be effected by the command

  group 5 1,2=1 "yng&med"

  The effect will be that the regression analysis program will assume that Age still has 3 categories, but of course, after the grouping no more records belonging to category 2 will be found, such that the model is not identified. The program SAUL will detect this, and will restart the analysis automatically with category 2 implicitly deleted. One can avoid this restarting with a somewhat more extended group command:

  group 5 1,2=1 "yng&med", 3,3=2 "old"

The following rules apply if one uses several of the commands skip, recode and group:

1. Each of these commands can apply to only one background variable.
2. Several commands of the same kind can be used with the same background variable. Example: the command 'skip 5 2, 9' is equivalent to the two commands 'skip 5 2' and 'skip 5 9' displayed on two lines.
3. Group and recode commands can be logically inconsistent. Although some inconsistencies may be detected by the program SLPRE, the built-in checks are not full proof. Results of undetected inconsistencies are unpredicted.
4. Each use of one of these commands starts on a new line.
5. The order in which the commands appear in the Komment field is completely arbitrary
6. The logical ordering of the commands is as follows: skip takes precedence, then comes recode and then group. The group command applies to the recoded values if any.

7. The skip command applies to the record. No other command can be applied to a skipped record, because it has virtually disappeared from the data file.

8. The commands group and recode only apply when the background variable is subsumed in the model (in other cases application would be futile). The skip command, however, applies also if the variable is not subsumed in the model.

9. Background variables which take 'out of range' values (for example having an explicit code for missing) cause the record to be skipped only if these variables are part of the model. For example: a record where the value of gender is out of range will be skipped if gender is a regressor; otherwise the observed value of gender is not relevant.

2.2.3 The Save Menu

The screen is left by typing F10; the program displays the Save Menu. The options of this menu are discussed in turn.

- Directory: the directory (folder) where to store the saved specifications. Notice that the directory specified here may be different from the one specified in the field dIrectory of the screen.

- File: the file name of the saved specification. By default this name is the jobName specified in the screen followed by the extension '.SLI'. Upon confirmation of this name (by placing the cursor on it, and typing ,CR>) one of two things will happen. If a file of this name (in the specified directory) does not exists, the file is saved and the program returns to the Init Menu. In case the file exists already, the user is warned, and asked whether the file can be overwritten or a new name is wanted. In the latter case the cursor goes back to the file field, where a new name can be typed. If the file is saved under a new name, the jobName in the specifications is altered to the name of the file, and saved with this new name. In this way consistency between file name and job name is guaranteed.

- Start file: this field acts as the file field, with the only difference that the name of this file is kept within the system, and will appear in the start file field of the init menu.

- No save: immediate return to the Init Menu. All specifications in the screen are lost. None of the existing files have changed.

2.3 The program SLPRE

- Purpose: reading and possibly transforming the data, preparing standardized input files for SAUL and creating a special file for use in the program RAPPORT (see next chapter).
- Components: a single run with phased messages on the screen in a DOS box
- Input files:
  - JOBN.SLI (created by SLIN)
  - JOBN1.SCR as specified in JOBN.SLI
  - DATA1 as specified in JOBN.SLI
  - DATA2 as specified in JOBN.SLI (optional)
  - The Refp file as specified in SLIN (optional) (see the chapter on RAPPORT)
  - The SAUL.DEF file (see the chapter on RAPPORT)
- Output files:
  - JOBN.SIN, a TXT file with information which is only relevant as input for the program SAUL
  - JOBN.BFR, a binary file, containing tables of statistics in a standardized and condensed form. This file is input for the program SAUL. Its contents may be made visible by the utility program SLLOOK, to be discussed in a separate section.
  - JOBN.WBO, a binary file which contains information for the program RAPPORT. It will be discussed in the chapter on the program RAPPORT.
- Command: SLPRE <CR> or SLPRE [path]JOBN[.SLI] <CR>. In the former case the JOBN.SLI file which is most recently edited by SLIN will be used automatically.

While the program runs, it writes three small reports to the screen. After each report it waits until the user hits a key and then continues the processing or stops. (But see the section on the SAUL.DEF file in the next chapter on how to suppress this option). We describe these reports in short.

1. Summary of the contents of the JOBN.SLI file and the SAUL.DEF file. This report allows the user to check the correctness of the specifications given in the job specification. The specified formats are displayed in an easy interpretable way by giving the start position, the end position and the field length of every variable that enters the regression analysis. Also the skip, recode and group commands are displayed in an easy to interpret lay-out.

2. Format and data checks. This report allows to check whether the format specified in the JOBN1.SCR file and the format of the DATA1 file go well together. The program prints the first record of the DATA1 file on the screen, and indicates the position of the booklet number (symbol b), the first item response (symbol d) and the position of the regressor values which are to be used in the current analysis (symbolized by their sequence number). If a DATA2 file is specified, the same information is displayed to check the format against the first record of this file.

3. Frequencies. To understand the frequency tables displayed, we need the concept of task. Suppose we use three background variables, taking two, three and four categories respectively. If the model specification is given by '1 2 3', we say that there are three tasks, each task corresponding with a background variable. For the second variable, the marginal frequency table is unidimensional, having three cells, since the variable can take three values. The report written by SLPRE will in this case consist of a unidimensional frequency table for each task. If the model specification is '[1 2] 3' we say that there are two tasks: the task [1 2] and the task '3'. For task [1 2], the marginal frequency table is a two by three table, which is displayed in a linear form. So for this task, six cells will be displayed. The rule for reading the frequencies correctly is that the categories of the last variable vary fastest. So the row and column indices of this two by three table are (1,1), (1,2), (1,3), (2,1), (2,2) and (2,3) in that order. Even if the task is specified by the user as [2 1], it will be changed implicitly by SLPRE to the task [1 2], and be reported accordingly. Or more generally, for tasks involving interactions, the component regressors are ordered according to their sequence number. For the second task '3', of course, the marginal frequency table is a unidimensional table with four cells.

2.4 The program SAUL

- Purpose: computing a latent regression analysis.

- Components: One run with a short initial dialogue
- Input files:
  - JOBN2.PAR: the parameter file specified in the screen of SLIN
  - JOBN.SIN, created by SLPRE
  - JOBN.BFR, created by SLPRE
- Output files:
  - JOBN.SLF, a text file with a complete report of the regression analysis
  - JOBN.SBO, a binary file read by the program RAPPORT
- Command: SAUL <CR> or SAUL [path]JOBN[.BFR] <CR> The former command selects the JOBN.BFR file which corresponds to the most recently edited JOBN.SLI file

To estimate the regression parameters and their standard errors, integrals must be computed frequently. Since there do not exist explicit formulae to evaluate these integrals, they must be approximated numerically. The numerical procedure implemented in SAUL is Gauss-Hermite quadrature, which can be described loosely as the replacement of a continuous variable by a discrete variable taking a limited number of values. These values are called quadrature points, and different approximations can be computed using different numbers of quadrature points. In older versions of SAUL this number was set at 20, but for some applications this number is too small to get accurate results. More details will be given in the subsection on statistical testing.

In the present version the number of quadrature points must be chosen by the user. There does not exist clear cut theory from which an sufficient number in any application of latent regression can be derived. It is even not sure that a larger number always gives better approximations than a smaller number. If the user wants to have some reassurance on the accuracy of his results, the only thing we can advise is to run the program several times with different numbers of quadrature points, and check on similarities and differences in the results. Some information on this topic from our own experience will be given further down.

The program SAUL starts with a short dialogue where it asks for the number of quadrature points to be used. This number is larger than 9 but not larger than 181.

To interpret the results of SAUL in a correct way, some theoretical concepts must be well understood. We mention the notions of origin and unit of the latent variable, the difference between significance and importance of an effect, testing the difference between several regression models and the exact meaning of interaction effects. Before discussing the output of SAUL, these topics will be discussed in short.

2.4.1 Origin and unit of the scale

The latent variable in IRT models in general is not fully identified. In OPLM the unit and the origin can be freely chosen. The origin is chosen by the so-called normalization of the item parameter estimates. The user of OPLM has control on this normalization, either by choosing the default normalization (the sum of the item parameters equals zero) of by fixing one parameter at an arbitrary value. If one runs two regression analyses with SAUL using the same data and the same parameter values except for the normalization, the only difference in the output of the two regression analyses will the additive parameter μ: adding an arbitrary constant to the item parameters will result in adding the same constant to the estimate of μ. All other values reported will be unaffected by the change in normalization.

The unit of the scale may be chosen or altered in different ways. A natural unit could be defined by fixing the standard deviation of the latent variable (for example by giving it a value of one). If conditional maximum likelihood (CML) is used for estimating the item parameters, however, the concept of the standard deviation of the latent variable does not exist. But the unit is chosen in this case by the magnitude of the discrimination parameters. This can be seen by inspecting the following identity in the OPLM-model:

$$\begin{aligned} P(X_i = 1|\theta) &= \frac{\exp[a_i(\theta - \sigma_i)]}{1 + \exp[a_i(\theta - \sigma_i)]} \\ &= \frac{\exp[ca_i(\theta/c - \sigma_i/c)]}{1 + \exp[ca_i(\theta/c - \sigma_i/c)]}, \quad (c > 0) \end{aligned}$$

If a new unit is chosen for θ by dividing it by a positive constant c, the probability of a correct answer remains unchanged if the original difficulty is changed accordingly, and the original discrimination parameter is multiplied by c. So the unit of the scale is chosen by fixing at least one discrimination index, which is always done in applications of OPLM.

Suppose we run two regression analyses using the same data, and the same measurement model where the only difference is that in the second OPLM estimation procedure the discrimination indices of the first analyses

are multiplied by a positive constant c. When we compare the two outputs we will see that in the second analysis, as compared to the first

- All regression parameters (inclusive μ) and their standard errors are divided by c
- The residual variance and its standard error are divided by c^2
- The value of the test statistics remains unchanged.
- The value of the effect size (to be discussed further) remains unchanged.

To avoid such a situation, it was decided to overrule the unit chosen by the user in the OPLM analysis, and to choose a unit which obeys the same rule in all SAUL analyses. **The scale unit in the output of SAUL is chosen in such a way that the geometric mean of the discrimination indices equals one.** (I.e, the discrimination indices are chosen in such a way that their product equals one. Of course the difficulty parameters are changed accordingly.) In this way, the estimates computed by SAUL are not affected by the order of magnitude of the discrimination indices used in the OPLM analysis. But notice that this applies only to the items which are used in the SAUL analysis.

The user should realize that this does not guarantee that the results of several models are directly comparable. We give an example of a possible pitfall. Suppose a calibration is done with twenty items, the first ten having a discrimination index of one, the others a discrimination index of two. If the measurement model is correct, one might expect that any regression model should give the same results using an arbitrary subset of the items, so that in practice the estimates resulting from any two regression analyses should be approximately the same (which can be checked by plotting the two sets of estimates against each other). But if we run an analysis using the first ten items and an analysis with the last ten items, a different unit in the reported scale is used (differing by a factor 2), so that the direct comparison of the parameter estimates becomes pointless.

2.4.2 Significance and importance

When a regression parameter estimate differs significantly from zero, one can safely (i.e., with a risk indicated by the significance level) accept that the true regression parameter is not zero. But this does not mean that the effect is important. True effects may be so small that they can be considered trivial, and yet lead to significant results. They will do so if the sample size

is sufficiently large. This means that one has to be careful when judging on the importance of a significant result.

A suitable way to judge on the importance is to compare the parameter estimate itself with some standard. The standard used in SAUL is the residual standard deviation. We consider some examples and discuss the differences in interpretation resulting from these examples. The general framework is a regression analysis with one binary regressor like gender. In the next table, four hypothetical outcomes of a regression analysis are displayed.

case	$\sqrt{\widehat{\sigma}^2}$	$\widehat{\beta}_1$	SE($\widehat{\beta}_1$)	z	effect size
A	0.30	0.25	0.05	5.00	0.84
B	0.22	0.25	0.14	1.79	1.16
C	3.86	0.85	0.12	7.08	0.22
D	6.07	0.85	0.45	1.89	0.14

The test statistic z is defined as

$$z = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$$

and the effect size is defined as

$$\text{Effect size} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2}}$$

If we look at the estimates themselves, we might be tempted to say that the cases A and B show a rather small effect and the cases C and D a rather large effect. But this would be quite misleading. First, in the cases B and D the estimate do not differ significantly from zero, meaning that whatever measure of the effect size we use, we always must admit that we do not have convincing evidence that there is any effect at all. This means that all discussions on the effect size in these cases are pointless. Second, in the cases A and C we have to bring the residual variability into consideration. In case A the numerical value of the estimate is rather small, but the residual standard deviation is small too, and the measure of the effect size is defined as the ratio between these. As a result we see that the effect size in case A is about four times as large as in case C. A probably convincing argument can

be found in the graphical display of the Figures 4 and 5.

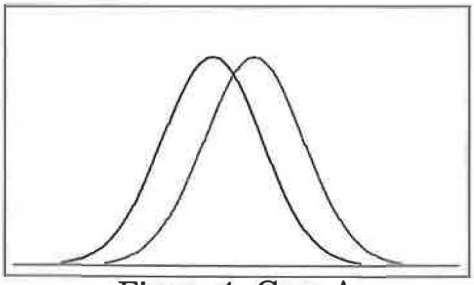

Figure 4. Case A

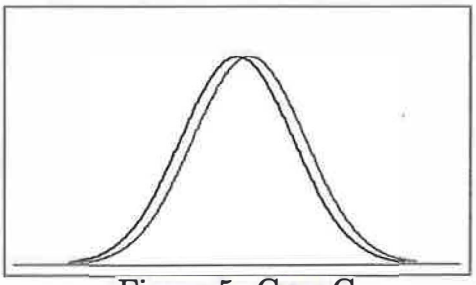

Figure 5. Case C

The two curves in each figure represent the distribution of the ability in the boys' and the girls' population. For case A we see that the two populations are reasonably well separated, while in Case C the two curves overlap a great deal.

The used definition of effect size is the same as the one adopted by Cohen (J. Cohen (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale: Erlbaum). This author proposes a labeling to classify effect sizes: an effect size of 0.2 is called 'small', 0.5 is called 'medium' and 0.8 is called 'large'.

2.4.3 Comparing regression models

The statistical tests discussed in Chapter 1 are well suited for use with a single parameter estimate. But sometimes one is interested in more global tests. Suppose for example that one regressor has six categories. This means that 5 regression parameters are to be estimated. These estimates can be tested, one by one, using the procedure described in chapter 1. Suppose none of these tests yields a significant result. It does not follow from this outcome that all true regression parameters are zero; the only justified conclusion is that we do not have enough evidence to claim that there is an effect differing from zero. If, on the other hand, one or two of these tests do yield a significant result, we must even be more careful, because we apply multiple comparisons. If we use a significance level of 5%, the probability of rejecting the null hypothesis if it is true is exactly 5% if we do a single test. If more tests are done, and the null hypothesis is true in all cases, the probability of rejecting one or more null hypotheses is larger than 5%, and in some cases dramatically larger. So, if we want to decide whether the different methods have any effect at all, the use of multiple t-tests is not very appropriate, and better methods exist. A method which can be applied with the program SAUL is the use of a likelihood ratio test.

An important requirement to apply the likelihood ratio test when comparing two models is that one of the models must be a special case of the other, or more technically, the parameter space of one of the models must be a subset of the parameter space of the other model.

In the example with the methods, the 5 regression parameters can in principle take any value. If the null hypothesis is true (methods do not have different effects), this means that the true regression parameters for all methods are zero, and we see that the model with zero parameters for the methods is a special case of the more general model where each method is allowed to have a different effect. But estimating the special case amounts to the same thing as leaving out the regressor 'methods' from the model.

In order to construct the test statistic we have to run the program SAUL two times: once for the general case where the regressor 'methods' is included, and once for the special case where 'methods' is left out. If other regressors are used, they must be the same in the two cases.

The program SAUL computes a quantity, which is called the log-likelihood at the solution (the likelihood is the probability of the observations as a function of the parameters; the log-likelihood is the logarithm of the likelihood, and the likelihood is computed as a function of the parameters at their final estimates. At this point the likelihood reaches its maximal value). The log-likelihood value in the general model cannot be smaller than the log-likelihood value in the special model, and the difference between the two is the basis for the statistical test. Denoting the log-likelihood value in the general model as $\ln L_g$ and in the special model as $\ln L_s$, the test statistic is given by

$$L = 2\,(\ln L_g - \ln L_s)$$

If the null hypothesis is true, L follows the chi-square distribution with degrees of freedom equal to the number of parameters in the general model minus the number of parameters in the special model.

In the example with the methods variable, the number of degrees of freedom equals the number of methods minus one, or $6-1=5$. This corresponds exactly to the number of regression parameters for the variable methods in the general model. If we test at the 5% level, we find in the tables of the chi-square distribution that the critical value is 11.07. So, if the value for L that we find is larger than this critical value, the null hypothesis of no differences between methods is rejected, otherwise it is not rejected. It is possible that the likelihood ratio test yields a significant result, while none of the five t-tests is significant. In such a case priority in the interpretation is to be given to the likelihood ratio test. If the likelihood ratio test is not significant, but some of the t-tests are, one should be careful in taking these significant

t-tests too serious: they may be an artefact of the multiple comparisons.

There exist a number of situations where the use of the likelihood ratio test is more appropriate than the t-tests per parameter. We give some of them.

1. When a regressor variable has many categories, and one wants to decide whether this variable has any effect at all. This case is used in the example above.

2. A more refined case is when one hypothesizes that a regressor has too many categories. Suppose the method variable can take 10 different values, but one thinks that they can be partitioned into a smaller number of categories, two say. This amounts to saying that the regression parameters for methods which belong to the same category are equal to each other. This defines the special model. One can implement this in SLIN by using the group command. To construct the test statistic L, one has to run the program twice: once with the original 10 values of the methods variable and once with the 2 grouped categories. The number of degrees of freedom in this case is $(10-1)-(2-1)=8$. Rejection of the null hypothesis means that the methods cannot be grouped into the two groups as hypothesized.

3. Models with and without interaction effects. Assume that we think there might be some interaction between two variables. Suppose the first variable can take three different values, and the second can take four. To model the interaction we have to define $(3-1)\times(4-1)=6$ extra dummy variables and to estimate 6 extra regression parameters. The model with interactions is the general model, and the special model is the same model, but with the interaction effects left out, which means that the regression parameters of the 6 extra dummy variables are set equal to zero. If the null hypothesis (no interaction effects) is true, the test statistic L is chi square distributed with 6 degrees of freedom. Rejection of the null hypothesis means that there is convincing evidence of interaction.

The computation of the log-likelihood value is rather complicated, and experience has shown that the numerical accuracy is sometimes very unsatisfactory if a small number of quadrature points is used. Since the statistic L has to be computed by the user from two independent runs of SAUL, it may happen that its computed value is negative, which is mathematically impossible if the correct values are used. Therefore it is strongly advised to use a large number of quadrature points if likelihood ratio statistics are wanted,

and to check the numerical stability with different numbers of quadrature points. The estimates of the parameters and their standard errors seem to be far less sensitive to variation in the number of quadrature points.

2.4.4 Reporting interaction effects

The disadvantage of working with nominal variables is that the number of dummy variables and hence the number of regression coefficients may grow rather fast, especially if interaction effects are estimated. Moreover, the description of the interaction effects is not always straightforward. To help the user with this interpretation, the outcomes of regression analyses involving interactions are reported in two different forms. The present section describes this output with the help of a simple example.

Assume that an analysis is requested where gender (two levels) and method (three levels: A, B and C) as well as the interaction effects of gender and method are estimated. The model is given by

$$\theta_i = \mu + \alpha_1 X_{1i} + \beta_1 X_{2i} + +\beta_2 X_{3i} + (\alpha\beta)_{11} X_{4i} + +(\alpha\beta)_{12} X_{5i} + \varepsilon_i$$

where

$$\begin{array}{lll} X_{1i} = & 1 & \text{if student } i \text{ is a girl,} \\ X_{2i} = & 1 & \text{if method B is used,} \\ X_{3i} = & 1 & \text{if method C is used,} \\ X_{4i} = & X_{1i} \times X_{2i}, & \\ X_{5i} = & X_{1i} \times X_{3i}. & \end{array}$$

The use of these two variables partitions the population into six subpopulations, and we can represent the mean ability (as a deviation from the constant μ) in the following table

	'A'	'B'	'C'
boy	0	β_1	β_2
girl	α_1	$\alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\alpha_1 + \beta_2 + (\alpha\beta)_{12}$

The first form of the output is to report the numerical value of the estimates as represented in the above table. So the output will look as follows

boy x 'A'	set to zero
boy x 'B'	β_1
boy x 'C'	β_2
girl x 'A'	α_1
girl x 'B'	$\alpha_1 + \beta_1 + (\alpha\beta)_{11}$
girl x 'C'	$\alpha_1 + \beta_2 + (\alpha\beta)_{12}$

In the second form, main effects and interaction effects are reported separately. Notice that the two variables, gender and method, have both a reference category, the effect of which could be represented by α_0 and β_0 respectively, but these effects are zero by definition. Also the interaction effects involving one or two of these categories are zero by definition. The report of the main effects looks like

main effects	
- girl	α_1
- 'B'	β_1
- 'C'	β_2

For the first order interactions we get

interaction effects	
- girl x 'B'	$(\alpha\beta)_{11}$
- girl x 'C'	$(\alpha\beta)_{12}$

Suppose there were no boys having had method 'B' in the sample, meaning that of the six combinations of gender and method only five have a non zero frequency. In such a case, as soon as SAUL detects the cell with a zero frequency, it will redefine the task set forth by the user, and estimate a model with four regression parameters instead of five: the cell boy and method 'B' is left out, and in the new model the following quantities will be estimated directly: α_1, β_2, $\alpha_1 + \beta_1 + (\alpha\beta)_{11}$ and $\alpha_1 + \beta_2 + (\alpha\beta)_{12}$. It is clear that from these four quantities β_1 and $(\alpha\beta)_{11}$ cannot be determined. In such a case the second form of the output is not given at all; only the first form appears in the output.

2.4.5 The JOBN.SLF file

The program SAUL creates a text file with a full report of the analysis. It is composed of the following parts

1. The comments provided by the user in Komment field of SLIN
2. The number of items and booklets used in the regression analysis. This information is found in the JOBN1.SCR file
3. A table of item numbers, item labels, discrimination parameters and item parameters as found in the JOBN2.PAR file (but only for the items used in the regression analysis). In the same table, a report is given on the transformed discrimination indices and the transformed difficulty

parameters. The product of the transformed discrimination indices is one. This transformation defines the scale on which the regression effects are reported.

4. A overview of the tasks requested by the user. The model specification '1 2 [3 6]' in SLIN defines three tasks: estimating the main affects of variable 1, the main effects of variable 2, and the main effects and interaction effects of variables 3 and 6.

5. Estimates of the additive constant ($= \widehat{\mu}$), the residual variance ($= \widehat{\sigma}^2$) and the standard deviation ($= \sqrt{\widehat{\sigma}^2}$), together with their (estimated) standard error. The standard error of the standard deviation is estimated from the standard error of the variance using the so-called delta method. This technique is not discussed further in this manual. The t-statistic is reported under the title 'z'. It is the estimate divided by its estimated standard error.

6. The log-likelihood value and the number of parameters estimated in the model. This information can be used to build likelihood ratio tests. The log-likelihood value is given up to a non specified additive constant (denoted as c in the output.) The value of this constant is not important for computing the statistic L. The reported value is given in so called scientific notation. For example $-0.2546D+05$. The number following the 'D' indicates that the decimal point in the preceding number must be shifted that number of positions (to the right if positive, to the left if negative). In the example this would give -25460, but we are not sure about the last digit. Therefore the log-likelihood value is also given with higher accuracy, like $-0.254598976354D+05$, which in decimal notation is the number -25459.8976354. Notice that the log-likelihood value is always negative, and that its magnitude grows linearly with the total number of observations. Running an analysis with ten times as big a sample will approximately decuple the log-likelihood value.

7. For each task implying only main effects, there is a report giving for each category, the category label, the estimate of the corresponding parameter, its estimated standard error, the number of observations in the sample falling in that category, the tests statistic 'z', and the effect size: the estimate divided by the residual standard deviation. For the reference category, the estimate is replaced by the string 'set to zero'; for categories with zero frequency in the sample, the string 'not identified' is printed. In general the first category is the reference category. If it has frequency zero, the next category is taken is the

reference category. The user can control for the reference category by using the recode command. For tasks implying interactions, two forms of output are displayed (see preceding subsection). In the first form estimates, standard errors, frequencies, z-values and effect sizes are displayed as well; in the second form frequencies are not displayed.

8. If a regressor has more than two categories, the pairwise differences between the regression coefficients are reported as well, accompanied by their standard error, the z-value and the effect size. The difference taken is indicated by the labeling. If there are for example three methods, the labeling will have the form 'method B - method C', indicating that the regression coefficient of method C is subtracted from the regression coefficient of method B, and not the reverse.

Chapter 3

The program RAPPORT

3.1 Introduction

A regression analysis reports essentially (estimates of) conditional averages. Using for example a two category and a three category regressor partitions the target population into 2×3 subpopulations and the correct combination of estimated regression parameters gives an estimate of the average ability in each of the six subpopulations. If the first variable is gender, it is immaterial whether the gender distribution in the target population is even or not and whether the numbers of girls and boys in the sample are equal or not, as long as the total sample can be seen as the merging of six random samples, one from each subpopulation. (See, however, the influence of unbalanced designs on the standard errors; Chapter 1).

In survey research, however, researchers usually want more than only an estimate of the conditional averages. In the example given, it may be a legitimate question to ask for the average ability of boys and of girls in the target population. To produce a consistent estimate of these marginal averages, one has to have information on the distribution of gender in the target population.

To fix the ideas, we will treat two examples, a simple one, which will be explained now, and a more complicated one which will be introduced later on. Suppose, for the simple example, we use two regressors, gender and method, but we used gender as a stratification variable, i.e. we sample a prespecified proportion of girls and of boys. Moreover, we assume that the sample of boys is a random sample from the population of boys and similarly for the girls. We will use the symbol p for observed proportions (in the sample) and the symbol π for 'real' proportions in the population. To keep the example simple, we will assume that π(girl)= 0.50, and that we sampled in such a

way that $p(\text{girl}) = 0.30$. Suppose now that we have the following observed proportions in the sample:

	'A'	'B'	'C'	total
boy	0.20	0.32	0.18	0.70
girl	0.15	0.08	0.07	0.30
total	0.35	0.40	0.25	1.00

which we will represent symbolically as

	'A'	'B'	'C'	total
boy	p_{11}	p_{12}	p_{13}	p_{1+}
girl	p_{21}	p_{22}	p_{23}	p_{2+}
total	p_{+1}	p_{+2}	p_{+3}	p_{++}

To estimate the joint proportions in the population, we have to correct back for the fact that gender was not drawn randomly from the population. For the boys this gives

$$\pi_{1j} = p_{1j}\frac{\pi_{1+}}{p_{1+}} \tag{3.1}$$

and for the girls

$$\pi_{2j} = p_{2j}\frac{\pi_{2+}}{p_{2+}} \tag{3.2}$$

where $\pi_{1+} = \pi_{2+} = 0.50$. This gives the 'corrected' table

	'A'	'B'	'C'	total
boy	0.143	0.229	0.128	0.50
girl	0.250	0.133	0.117	0.50
total	0.393	0.362	0.245	1.00

Notice that the marginal proportions for method are obtained by summing the corrected joint proportions columnwise: $0.393 = 0.143 + 0.250$, or with a formula:

$$\pi_{+j} = \sum_i \pi_{ij}$$

A number of remarks are in order here.

1. All the π-values reported here are estimates, except perhaps for the population proportion of 0.50 for boys and girls, which may be based on theoretical considerations. But in real life surveys, even these proportions will be estimated, for example, using very big samples used by the national office of statistics.

2. We see from the last table that gender and method are not independent (orthogonal) variables in the population: method 'C' has about the same proportion in the boys as in the girls subpopulation, while method 'B' is used far more frequently with boys and method 'A' more frequently with girls.

3. The two regressors are not treated in the same way: the variable gender is the stratification variable. This means in this example that the ratio between any two proportions in the same row is not changed by the correction. For example in the row 'boys', it holds that $0.20/0.32 = 0.143/0.229$. But this invariance does not hold for the variable method: $0.20/0.15 \neq 0.143/0.250$.

4. The numbers in the corrected table can only be trusted if it can be guaranteed that the sampled boys are a random sample from the boys' population, and similarly for the girls.

Once the corrected table has been built, the marginal averages for all categories of gender and of method can be computed in a straightforward way. As an example, we use the following estimates of regression effects (using a model without interactions):

	'A'	'B'	'C'
boy	0	β_1	β_2
girl	α_1	$\alpha_1+\beta_1$	$\alpha_1+\beta_2$

$=$

	'A'	'B'	'C'
boy	0	2	3
girl	5	7	8

and by combining the last two tables we find for example that

$$\begin{aligned}\text{Average (girls)} &= \mu + \frac{1}{\pi_{2+}}[\alpha_1 \times \pi_{21} + (\alpha_1+\beta_1) \times \pi_{22} + (\alpha_1+\beta_2) \times \pi_{23}] \\ &= \mu + \frac{1}{0.50}[5 \times 0.250 + 7 \times 0.133 + 8 \times 0.117] \\ &= \mu + 6.233\end{aligned}$$

while the average for the boys is $\mu + 1.684$. The difference between girls' and boys' average is 4.549, which is less than the regression effect estimated, but this is a consequence of Simpson's paradox.

The main purpose of the program RAPPORT is to estimate the marginal averages (in the target population) of all categories of the regression variables used in the regression analysis. At the same time, it repeats the output of the JOBN.SLF file, but now on a scale where the user can define origin and unit. The control on the program RAPPORT is rather complicated, and the user is advised to follow carefully the instructions given in the next sections.

3.2 RAPPORT and the program SLIN

If one intends to run the program RAPPORT, a number of provisions must be made while filling out the Screen in the program SLIN. These are discussed one by one now.

- Refp: In this field the name of a so-called reference population file (further: refpop file) is asked for. The name of the file is arbitrary. The file contains the joint distribution (in the population) of a subset of the background variables listed in the screen. This subset is in principle arbitrary, and does not depend in any way on the variables used in the model. The refpop file has a special format, and the easiest way to construct it is to let it be created by SLPRE. Details will be given in a separate section below.

- reFp: in this field the maximum number of categories in the population for each regressor has to be typed. This number is less than or equal to the number of categories specified in the field daTa for the corresponding variable. It may happen that for a certain variable observations are available in three categories, say, but that only two of them belong to the target population one is interested in. The categories available for the regression analysis but not in the target population must be the last ones of the regressor. If they happen to be coded in a different way, the recode statement can be used to obey this rule. The number specified in the reFp field must correspond to the number of categories used in the refpop file (in case this variable is represented in the refpop file).

- Only one variable can be used as stratification variable. This is always the first variable, i.e. the variable with sequence number 1. Of course, there may be cases where no stratification is applied. The way to provide for such a case will be discussed in the section on the SAUL.DEF file (see below).

3.3 Creating a refpop file

The refpop file is a binary file created by a special run of SLPRE. The following specifications must be given in the application screen of SLIN

- The whole screen is filled out as for a regression analysis

- In the list of variables, only the variables are represented for which the distribution in the refpop file is wanted.
- The name of the refpop file (in the field Refp) is preceded (immediately) by the at-sign ('@')
- The result is the same for all model specifications; so the empty model suffices.
- Recode and skip commands are taken into account when creating the refpop file; group commands are ignored.
- Background variables are read from the DATA1 file and optionally from the DATA2 file. Item answers in the DATA1 file are ignored.
- Weights specified in SAUL.DEF file (see below) are taken into account.
- If at a later point in time, one wants to use the created refpop file in a run of the program RAPPORT, the variables used for creating the refpop file must be the first variables used in defining the variable list for the regression analysis, and, moreover they must be given in the same sequential order. (Labels are not used to identify variables.)

As an example, we use a situation which arises frequently in the Dutch national Assessment Program for the Basic Education (PPON). When data are collected, this is usually done for two or more domains at the same time. So, in a particular year, data might be collected for the assessment of musical performance and of arithmetic proficiency. For the regression analyses some of the background variables are identical across domains, like a stratification variable which summarizes the socio-economic background of the students in a school and gender. Others, however, are domain specific, like the arithmetic method used in the school or the number of minutes devoted weekly to musical education. If we use the total sample of students available in a particular year to estimate the joint distribution of stratum and gender, we will get more stable estimates than when we use a different estimate for arithmetic and musical education.

So, to create the refpop file, we will (in this example) define a list of two variables, stratum and gender. (Stratum is the first because it is the stratification variable.) The data file needed to create the refpop file consists of the conjunction of the data files for arithmetic and music. The only thing that really matters is that the format for the background variables, stratum and gender, is valid for the two files, i.e., in both files the values of these background variables must take the same positions. Booklet numbers,

item answers and the format in the JOBN1.SCR are not relevant. But the JOBN1.SCR file must exist. The JOBN2.PAR file can be an arbitrary *.PAR file.

If later on, we want to define a regression analysis, followed by a run of the program RAPPORT, where stratum, gender and one or more other variables are used, the variable stratum must have number 1, and the variable gender must have number 2.

Notice that the refpop file is created or used by the program SLPRE. It is not read by the program RAPPORT. How it is used by SLPRE will be described in more detail in the section called 'The SAUL.DEF, refpop and JOBN.WBO files' below.

3.4 The SAUL.DEF file

In principle, the SAUL.DEF file is a system file, which resides in the directory identified by the HDIR environment variable. While it was originally planned to keep only system defaults, the steadily growing complexity of the regression models used in the Dutch National Assessment Program has necessitated to make the information in this file more and more specific. In the present version of the program, this file plays an important role, and it is necessary to check or adapt its contents for every application. **The contents of this file, however, do not influence in any way the output of the program SAUL, they only apply to the program RAPPORT.**

To make the contents of the SAUL.DEF file easier to understand, a more complicated example of a regression analysis will be discussed first. Then, the contents of the SAUL.DEF file will be described in a more formal way.

3.4.1 An example

In the national assessment program PPON, each domain is assessed every four or five years, and in a specific year data on more than one domain are collected. Apart from estimating the effect of variables like gender and age of the students or method of instruction used, the development of the ability over time is estimated as well. In principle the method is quite simple: if one has data collected at three different time points, cycle 1, 2 and 3 say, the variable cycle with three categories is introduced as regressor in the model and the average ability at the three different cycles can be estimated, with or without control of other variables depending on which regression model one wants to use.

There are, however, three complications when one wants to report marginal averages with the program RAPPORT:

1. Stratified sampling within cycle. In PPON the population (within a cycle) is stratified according to a variable which summarizes in some way the socio-economic status of the students in a school. Three different levels of this variable, henceforth called stratum, are distinguished. The first problem arises because the distribution of schools across strata needs not to be constant across cycles. It may happen, for example, that in the first cycle 40% of the schools belong to the the first stratum, while this percentage is only 35 in the second cycle and may drop to 25 in the third cycle.

2. The second problem has to do with the fact that stratum is defined at school level, and not at the individual level, while in the intended report, one usually wants to generalize on the population of students and not on the population of schools.

3. The third problem has to do with the definition of the target population. One could argue that there in fact three target populations, one for each of the cycles, but on the other hand one could also argue that the three cycles together form a kind of superpopulation. In the former case it does not matter whether the total number of students is constant across cycles or not, while in the latter case a positive or negative growth of the student population over time may influence the report produced by the program RAPPORT.

In the next subsection it will be shown how to treat these problems technically, using the SAUL.DEF file. But first some more theoretical issues will be discussed with the aid of an example.

In the assessment program PPON stratified cluster sampling is applied: the population of schools is stratified into three strata, and for each stratum a prespecified number of schools is sampled randomly. All students (of the relevant grade level) in each school participate in the assessment of one of the domains studied at that particular time. The specific domain a student is given is chosen randomly. In the next two Tables, the number of schools sampled and the total number of schools in each stratum are displayed for

three cycles of the assessment of one particular domain.

Sample

	str. 1	str. 2	str. 3	total
cycle 1	19	18	23	60
cycle 2	67	72	61	200
cycle 3	61	31	11	103

(3.3)

Population

	str. 1	str. 2	str. 3	total
cycle 1	4097	2676	1672	8445
cycle 2	4424	2197	1644	8265
cycle 3	3806	2337	932	7075

(3.4)

With respect to these tables, some remarks are in order:

1. We see that in the population stratum 1 outnumbers (in number of schools) the two other strata, while in the sample this is definitely not the case: in the first two cycles there has been an attempt to sample an (approximately) equal number of schools from each stratum. In the third cycle, this has not been the case for some reason.

2. The total number of schools is decreasing over time (there is a period of ten years between cycle 1 and cycle 3). This does not necessarily imply, however, that the total number of students in the population has decreased proportionally, since in the period reported a large number of mergers between schools has taken place, so that the average number of students per school has probably increased.

3. The numbers reported are frequencies of schools, not of students. Exact information on the number of students in the target grade level the assessment is designed for, is not available. The only information that was available is the total number of students for the whole school, and the exact number of students in the relevant grade level for the sampled schools. Although the average numbers of students per school in the sample were not constant across strata, the differences were small, and were ignored in the study. This means that the assumption was taken that the average number of students per school was constant across strata within each cycle. If this assumption is true, the numbers given in tables (3.3) and (3.4) are proportional to the number of students as well, but this proportionality holds only within a cycle.

4. Proportionality across cycles is certainly not warranted, because the total number of schools sampled has deliberately changed (cycle 2 having much more schools than cycle 1 for example), and because the average number of students per school may have changed across cycles. This was reason enough to weight the cycles equally, i.e., treating the superpopulation of the three cycles as consisting of equally numerous populations.

So the correction factors p/π for each cycle-stratum combination are constructed as displayed in the next two tables.

Sample (p)

	str. 1	str. 2	str. 3	total
cycle 1	$\frac{1}{3} \times \frac{19}{60}$	$\frac{1}{3} \times \frac{18}{60}$	$\frac{1}{3} \times \frac{23}{60}$	$\frac{1}{3}$
cycle 2	$\frac{1}{3} \times \frac{67}{200}$	$\frac{1}{3} \times \frac{72}{200}$	$\frac{1}{3} \times \frac{61}{200}$	$\frac{1}{3}$
cycle 3	$\frac{1}{3} \times \frac{61}{103}$	$\frac{1}{3} \times \frac{31}{103}$	$\frac{1}{3} \times \frac{11}{103}$	$\frac{1}{3}$

(3.5)

Population (π)

	str. 1	str. 2	str. 3	total
cycle 1	$\frac{1}{3} \times \frac{4097}{8445}$	$\frac{1}{3} \times \frac{2676}{8445}$	$\frac{1}{3} \times \frac{1672}{8445}$	$\frac{1}{3}$
cycle 2	$\frac{1}{3} \times \frac{4424}{8265}$	$\frac{1}{3} \times \frac{2197}{8265}$	$\frac{1}{3} \times \frac{1644}{8265}$	$\frac{1}{3}$
cycle 3	$\frac{1}{3} \times \frac{3806}{7075}$	$\frac{1}{3} \times \frac{2337}{7075}$	$\frac{1}{3} \times \frac{932}{7075}$	$\frac{1}{3}$

(3.6)

To make the example complete, suppose two methods of instruction, 'A' and 'B' have been used in each of the three cycles, and the frequency table of the methods used is as follows:

cycle, stratum and method: frequencies

	str. 1		str. 2		str. 3	
	'A'	'B'	'A'	'B'	'A'	'B'
cycle 1	11	8	12	6	11	12
cycle 2	30	37	32	40	28	33
cycle 3	22	39	11	20	3	8

In cycle 1, the joint proportion of method 'A' and stratum 1 is $11/60 = 0.183$. Notice that this proportion is conditional on cycle 1: computing similar proportions for all cells in the cycle 1 row of the table, and adding them will

give a sum of one. Correction for stratification (see the formulae (3.1) and (3.2)) gives a population estimate of

$$\frac{11}{60} \times \frac{\frac{4097}{8445}}{\frac{19}{60}} = 0.281$$

Computing corrected proportions for each cell gives the following result:

cycle, stratum and method: corrected proportions

	str. 1		str. 2		str. 3		total
	'A'	'B'	'A'	'B'	'A'	'B'	
cycle 1	0.281	0.204	0.211	0.106	0.095	0.103	1
cycle 2	0.240	0.296	0.118	0.148	0.091	0.108	1
cycle 3	0.194	0.344	0.117	0.213	0.036	0.096	1

We conclude this example section with two remarks, one being an important example of Simpson's paradox, the other commenting further on the assumption that the average school size within a cycle is constant across strata.

1. Suppose with the data from the three cycles, we carry out a regression analysis using cycle, stratum and method as regressors, using no interactions. The hypothetical effect estimates are given in the next table

cycle	est.	str.	est.	meth.	est.
c. 1	0	str. 1	0	'A'	0
c. 2	−0.5	str. 2	−1	'B'	5
c. 3	−1	str. 3	−2		

from which we see clearly that, if we control for stratum and method, the average ability is going down with time. Using the corrected proportions and the effect estimates, we can compute the population average in each cycle, giving

$$\begin{aligned} \text{Average(cycle 1)} &= 1.352 \\ \text{Average(cycle 2)} &= 1.596 \\ \text{Average(cycle 3)} &= 1.671 \end{aligned}$$

a clear increase with time. But a little thought will reveal that this has to be explained by the increasing popularity of the better method 'B'. The decreasing effect of the variable cycle then allows for an interpretation, that in spite of an increasing use of a much better method,

the average ability does not grow as fast as might be expected. This can be easily checked. Suppose there was no cycle effect, so that the average in each cycle is only determined by the stratum and method effect, then the cycle averages would equal 1.352, $(1.596 + 0.5) = 2.096$ and $(1.671 + 1) = 2.671$, respectively. Of course, the reason why these effects are negative cannot be deduced from this analysis.

2. Even if we are sure that the average number of students per school is constant across strata, we cannot be sure that we did not introduce an error in correcting the estimated proportions. Suppose that the schools using method 'B' have on the average less students than schools using method 'A', then the corrected proportions are biased as an estimate of the proportion of students in each cell, the 'B' cells being overestimated and the 'A' cells underestimated. Therefore, it is always risky to introduce weights according to a level (schools) as representative for another level (students).

3.4.2 Details on the SAUL.DEF file

This file is a simple text file containing an arbitrary number of comment lines and four lines with information which is relevant for the program SLPRE. The comment lines appear on top of the file, and are closed with a separate line containing the string '=*' (without the quotes) in the first two positions. The next four lines, containing relevant data for the program will be denoted as line 1 to line 4. The 4 lines only contain numbers which are read in free format (i.e. separated by one or more blanks or commas).

1. line 1: weights for each of the strata in the population. In the example of the preceding section, these are the frequencies in Table (3.4), or multiples thereof. The weights are written row-wise or column-wise (see line 4, how to make the distinction.).

2. line 2: weights for each of the strata in the sample. In the example of the preceding section, these are the frequencies in Table (3.3). Of course, they must be arranged in the same way as the weights in line 1. If no stratification is used while sampling, care must be taken in specifying lines 1 and 2 in the SAUL.DEF file. We give an example. Suppose no stratification is used, and we want to use gender as the first variable. To indicate that gender was not a stratification variable, the two lines (1 and 2) of SAUL.DEF must be equal to each other. So specifying the lines as '1 6' and '1 6' respectively will lead to no correction, because the ratio π/p equals 1 for both categories. This

issue will be discussed further when explaining the variable '#strata' in line 4.

3. line 3: two numbers to fix the origin and the unit of the reporting scale. The first number is the population average and the second number the population standard deviation. In the example of the preceding section, the population may be the superpopulation, comprising the three cycles, or some subpopulation thereof. Which one is chosen, is to be specified when running the program RAPPORT.

4. line 4 contains 5 numbers, which are symbolically denoted here as Cov, Key, #GH, #strata and SX.

   - Cov = 1: the covariance matrix of the parameter estimates is written on a special file by the program SAUL (not implemented in the present version).
   - Key = 1: The program SLPRE writes small reports on the screen, and waits each time for the user to get a key before continuing. If Key $\neq$ 1, SLPRE does not wait. The latter option may be appropriate to run several applications in batch.
   - #GH: the number of quadrature points to be used by SAUL (not implemented in the present version; the program SAUL asks explicitly for the number of quadrature points.)
   - #strata. Two remarks are in order here
     - In the example of the preceding section, the stratum variable has three levels, but in line 1 and 2, nine numbers are specified (the numbers in the Tables (3.4) and (3.3) respectively). If we specify here 3 as the number of strata, the program SLPRE will automatically assume that the three strata are repeated a number of times, defining implicitly another variable (which was called cycle in the example). The three cycles will automatically get an equal weight; so the program SLPRE will do essentially what was shown in the example. If we specify 9, the program will assume that the stratum variable has 9 levels and occurred only once. Using 9 in the example above would lead to an estimation where the cycles are weighted proportionally to the number of schools existing at assessment time. Notice that the number of weights specified in lines 1 and 2 must be an integer multiple of the number of strata specified here.

- In case there is no stratification variable, the first variable will nevertheless be treated as such. Suppose gender is used as first variable, and we specify here 1 as the number of strata, then the program SLPRE will define implicitly two repetitions of one stratum, and the weights will be equal, leading to an estimated proportion of boys and girls of 0.50 in the population, whatever the number of boys and girls in the sample. If we specify 2 as number of strata (and the lines 1 and 2 are identical) the estimated proportion of boys and girls in the population will be equal to the corresponding proportions in the sample, irrespective of the numbers used in lines 1 and 2. So specifying these two lines as '1 6' will have the same effect as using '1 1'.

- SX : this number shows how the numbers in lines 1 and 2 have to be interpreted. If SX= 1, the interpretation is that cycles are ordered within strata, or that the frequencies of Table (3.3) are stored columnwise. In this case line 2 is written as '19, 67, 61, 18, 72, 31, 23, 61, 11'. If SX $\neq$ 1, the interpretation is that strata are ordered within cycles. In this case line 2 is written as '19, 18, 23, 67, 72, 61, 61, 31, 11'.

3.4.3 The SAUL.DEF, Refpop and JOBN.WBO files

To see the relations between these three files, the preceding example will be used further. We will assume that we have three strata in each cycle, and we will use stratum, cycle, gender, age and method as regressors. Moreover we will introduce the following complications:

- We have many data from each assessment cycle which contain information about gender and age over and above the data which were collected for the specific domain we are interested here. We want to use this information to get more stable estimates on the population distribution of these variables.

- The variable Age has three levels, but only students having age level 1 and 2 belong to the target population. We do not want to discard the students belonging to age level 3, because it is interesting to know how this category of students is related to the target population. From a separate regression analysis, however, we know already that the regression coefficient for this category is pretty much the same as for age category two. For reasons of parsimony, we want to restrict the

regression model such that age levels 2 and 3 have the same regression coefficient. This can be done via the grouping command. (We could also leave out students from the third age category, but statistically this would not be efficient.)

- The variable method has also three levels, but this variable is domain specific, and data from other domains cannot be used to estimate the distribution of this variable. The only source of information we have is the data file for the domain under study.

In order to compute marginal weights, the program RAPPORT needs an estimate of the joint distribution of the variables stratum, cycle, gender, age (with two, not three categories) and method. This distribution can be stored in a five dimensional table, having in total $3 \times 3 \times 2 \times 2 \times 3 = 108$ cells. The contents of the JOBN.WBO file is precisely this table. It is created by the program SLPRE and read and used by the program RAPPORT.

Next we explain step by step how to proceed to get the right results.

1. Since we use a stratification variable (the variable stratum) which can have a different distribution in each cycle, we must create a new artificial variable which combines stratum and cycle. We will label this variable as 'strtcycl'. This variable has $3 \times 3 = 9$ levels. An indicator of this level must be present on the data files that are used in the sequel.

2. First we will create a refpop file. To this end we need a data file where information is stored with respect to the background variables 'strtcycl', stratum, cycle, gender and age. Call this data file DATA0.

3. We create a refpop file (with the generic name REFPOP), using DATA0. (For details see the section on creating a refpop file). The important thing to remember here is the correct specification of the variables list and the number of levels to be specified in the fields daTa and reFp. This table is given here:

Nr	laBel	daTa	reFp
1	strtcycl	9	9
2	stratum	3	3
3	cycle	3	3
4	gender	2	2
5	age	3	2

4. If we specify @REFPOP in the field Refp, and run SLPRE, a five dimensional table with the joint distribution of the five variables specified in the SLIN screen is created. This table is stored on the file REFPOP. The correction for stratification are applied as specified in the SAUL.DEF file. The variable age has only two values. All data records with age level 3 are skipped.

5. In this step, the screen in SLIN is adapted to the regression analysis we want. This implies the following changes:

   - The reference to the refpop file is now REFPOP instead of @REFPOP, because the refpop file exists;
   - The data file is DATA1, containing the data relevant to the domain assessment we are interested in;
   - The domain specific regressor 'method' is added to the list of variables, as variable number 6. Do not insert it somewhere between the formerly specified variables, because this would destroy the correct link between variables in the present analysis and the entries in the refpop file. Also, do not remove the 'strtcycl' variable, although it will not appear in the regression model.
   - Apply the group command in the Komment field, to group together the levels 2 and 3 of the variable Age;
   - Specify the model as ' 2 3 4 5 6' or a more complicated model involving interactions, like for example '2 3 [4 5] 6'. Notice that the stratification variable 'strtcycl' is not part of the model.

6. Run the program SLPRE. One of the tasks this program carries out, is to construct the table with the joint distribution of the background variables used in the model. In the present case this involves two logical steps, but the example is general enough to apply in all cases.

   - Since variable 1 is not part of the model, the table which resides on the refpop file is collapsed along the dimension of variable 1, thus reducing this table from five dimensions to four dimensions. In the general case, this means that the table is collapsed along all dimensions not being part of the specified regression model.
   - Variable 6 is not subsumed in the table of the refpop file. While SLPRE is processing the data from the DATA1 file, it will expand the collapsed table from the previous step along the new dimension for methods. An example will clarify this: suppose we

have an estimated proportion equal to 0.05 in a specific cell of the four dimensional collapsed table, for example for the combination stratum 1, cycle 1, boy and first age group. Suppose that there are 100 records in the DATA1 file pertaining to this cell, and that the methods applied in these 100 cases have frequencies 60, 20 and 20 for the methods 'A', 'B' and 'C' respectively. The original estimated proportion of 0.05 will then be split proportionally to these frequencies, yielding 0.03, 0.01 and 0.01. As this is done for each cell of the four dimensional collapsed table, the result will be a five dimensional table, where the five dimensions represent precisely the five variables of the regression model. This table is written to the JOBN.WBO file, and will later on be used by the program RAPPORT. Notice that the regression program SAUL does not use this file.

- It is possible that a cell in the (collapsed) table has a positive proportion, but that the proportion in the DATA1 file for this cell is zero (because the refpop file can be based on a different sample). If the collapsed table has to be expanded (along the dimension method, say), no observations are found. In such a case the estimated proportions are set equal to each other. In the preceding example this would yield three proportions of $0.05/3 = 0.01666\ldots$ for each of the three methods.
- There is a complication with the variable Age. For the regression analysis, the original categories 2 and 3 are now treated as a grouped new category, but the original category 3 is excluded from the reference population. The program SLPRE takes care of this: while constructing the table for the joint distribution, records with age level 3 are not used, but they are used for the regression analysis. The only confusing feature is the label used in the output. As the grouping command asks for a label of the grouped category, this label will be the only one which appears in the output of SAUL and RAPPORT, but it will have different meanings at different places.

To conclude this section, some remarks are in order.

- If the model specification is empty, no JOBN.WBO file is created.
- If no refpop file is available, the JOBN.WBO file is created anyway: it is based exclusively on the relative frequencies found in the DATA1 (and possibly DATA2) file. Even if the stratification variable is not

part of the model, it will be used anyway in estimating (and adjusting for stratification) the joint distribution of the background variables in the population.

- The practical implication of the preceding point is the following: if the refpop file is exclusively based on the information contained in the DATA1 (and DATA2) files, running SLPRE and RAPPORT with and without use of the refpop file will give identical results, except for possible effects of missing data. This is explained next.

- Suppose the DATA1 file contains all the information we have on the population distribution of the background variables. It has 1000 records and each record has information on all background variables, except the last one where gender is missing. Creating a refpop file (with all variables) will cause the last record to be skipped, because of incompleteness, and consequently every JOBN.WBO file will be based on 999 observations if the refpop file is used. If we do not use the refpop file (i.e. we do not specify it in the field Refp of SLIN), and we estimate a model without the regressor gender, SLPRE will create a JOBN.WBO file which is based on 1000 records.

- More will be said on the refpop file when discussing the output of the program RAPPORT in the next section.

3.5 Running the program RAPPORT

- Purpose: report of the regression analysis and of marginal averages and distributions

- Components: One run with an initial dialogue

- Input files:

  - JOBN.SBO: a binary file created by SAUL
  - JOBN.WBO: a binary file created by SLPRE

- Output file:

  - JOBN.EFF, a text file with a complete report of the regression analysis and of the marginal distributions.

- Command: RAPPORT <CR> or RAPPORT [path]JOBN[.SBO] <CR>
  The former command selects the JOBN.SBO file which corresponds to the most recently edited JOBN.SLI file. Notice that a conflict may arise here. Suppose one has run RAPPORT before, using a JOBN.WBO file. After that, the model has been converted (in SLIN) to an empty model, but the job name JOBN has not changed. Because the model is empty, no new JOBN.WBO file is created by SLPRE, but the former one continues to exist. Running SAUL will create a new JOBN.SBO file, which is not in correspondence with the old JOBN.WBO file, and a subsequent run of RAPPORT will lead to a crash.

The output file of RAPPORT consists of several components which are described in the sequence as they appear in the file. To make the exposition not too abstract we will illustrate it with an example that uses the same variables as in the example section above: an assessment running over three cycles, and using several regressors such as stratum, gender, age, and method (or a subset of them), but certainly cycle itself.

The program RAPPORT starts with an initial dialogue where three kinds of information are asked: the first conditioning variable, the second conditioning variable, and the reference variable together with the reference category of that reference variable. The concepts of first and second conditioning variable will be explained as they become relevant in the discussion of the JOBN.EFF file. The concept of reference variable and reference category is explained first.

In the SAUL.DEF file (line 3), the values for the population mean and standard deviation are specified. In the PPON project the values of 250 and 50 respectively are used. As long as we analyze data from the first cycle, this offers no special problems: the scale is transformed such that the population of the first cycle has a mean ability of 250 and a standard deviation of 50. But once we start to analyze data from different cycles, the joint data set is considered as a sample from a superpopulation, and it may appear not too appropriate to set the mean ability of this superpopulation equal to 250. Instead the user might want to keep the mean and standard deviation of the first cycle population at 250 and 50 respectively. This may be accomplished by choosing the variable cycle as reference variable and the category 'cycle 1' as reference category. The initial dialogue allows the user to choose any category of any variable as reference category or to choose none, in which case the origin and unit apply to the whole (super)population.

The sections of the JOBN.EFF file are described next.

1. General information including the comments of the user (from the Komment field in SLIN), information on missing data, and on the particular

options used in RAPPORT.

2. A table with the item parameters of the items used in carrying out the regression analysis. The choice of origin and unit (in the SAUL.DEF file), and the choice of the reference variable and category (in the initial dialogue of RAPPORT) automatically imply a transformation of the discrimination indices and the difficulty parameters of the items. The table presented here gives the original values, as found in the JOBN.PAR file and the transformed parameters. In the table, two extra columns are added, with headings P80 and P50 respectively. The P80 column gives for each item the value of the latent variable on the transformed scale that corresponds to a 80% probability of success on the item. For binary items, the 50% probability point is found at the value of the transformed difficulty parameter (and is not repeated in the P50 column); for polytomous items, the 50% probability point is computed by the program RAPPORT and displayed in the P50 column.

3. The results of the regression analysis. This part of the output has essentially the same format as the output of SAUL (contained in the JOBN.SLF file). Here, we discuss only the differences and a few new elements.

   - The output echoes the value of the mean and standard deviation as specified in the SAUL.DEF file, together with the reference variable and reference category as chosen in the initial dialogue.
   - Changing unit and origin defines a linear transformation of the scale, using the rule

     $$\text{new values} = B \times \text{ original values } + A$$

     The values of B and A are displayed under the heading 'transformation constants' as 'multiplicative' and 'additive' respectively. This transformation holds for estimates of regression parameters and averages. Standard errors and discrimination indices are only affected by the value of B; test statistics and effect sizes do not change at all.
   - Log-likelihood value and number of parameters have the same value as in the JOBN.SLF file.
   - The additive regression parameter μ is given with the name 'additive constant'. The mean (ability) is displayed with the heading 'mean'. If no reference category is chosen in the initial dialogue,

this mean will be equal to the mean specified in the SAUL.DEF file, otherwise it will in general be different. If we choose 'cycle 1' as the reference category, and 250 as the average value in the SAUL.DEF file, this will cause a transformation of the scale such that the average of the cycle 1 population is 250; the average ability of the superpopulation (the three cycles together and equally weighted), however, will in general take another value.

- A similar reasoning as for the mean applies to the standard deviations.

4. One or more tables with marginal proportions, means and standard deviations for each variable which was used as a regressor in the model. The default option of RAPPORT is to display the mean and standard deviation (in the population) for each category of each regressor. For the regressor gender this amounts to the proportions, means and standard deviations of the boys and the girls. For the variable cycle, the proportions, means and standard deviations are displayed for each cycle. This latter case is probably meaningful to the user, but it is doubtful if the former is: it gives the mean ability of boys and girls, averaged over all other regressors, inclusive the cycle. Perhaps it is more valuable to display the proportions, means and standard deviations for boys and girls in each cycle. But doing this is equivalent to conditioning on the variable cycle, and this is controlled by the user by choosing cycle as the first conditioning variable in the initial dialogue. The effect is that for each category of each variable the conditional proportions, means and standard deviations are displayed, inclusive the conditioning variable itself. For example, in the subtable pertaining to the first category (cycle 1) of the variable cycle, the proportion of cycle 1 equals one, and the proportions of the other two cycles are zero. Of course, for the latter no means and standard deviations exist.

5. A table with the estimates in the joint distribution of the background variables in the population. This amounts to displaying the contents of the JOBN.WBO file. The table is displayed with one cell per record; the record consists of a cell identification followed by the estimated proportion. Even with a moderate number of regressors, this table may be quite large and will not show interesting features by mere inspection. But it can be used as input table for more refined analyses, such as log-linear analysis, to investigate relations between background variables and their change over time. The table is preceded by a legend telling how to interpret the cell identifications.

6. It is a quite widespread misunderstanding that the usual assumption of normally distributed residuals in regression analysis entails a normal distribution of the dependent variable (the ability). With a simple example, it can be seen that this is not true. Suppose, we use a (valid) regression model with one binary regressor (such as gender). The model assumes that the ability in the boys' population and in the girls' population is normally distributed. If the gender effect is zero, the ability distribution in the total population will of course be equal to either distribution, but if the effect is not zero, the marginal distribution will be a mixture of the two distributions, and this distribution is not normal. If the effect is small the normal distribution might be a reasonable approximation, but if the effect size is large, the marginal distribution will become bimodal, and an approximation by a normal distribution will be erroneous. (See Figures 4 and 5 for a graphical example.) If we use five binary regressors, the marginal distribution is a mixture of $2^5 = 32$ normal distributions, each having a weight defined by the joint distribution of the background variables, and this mixture distribution may deviate quite substantially from the normal distribution. The program RAPPORT computes and displays the correct percentiles (1 to 99) in such a mixture. The user can control to a certain extent the (sub)populations of which the percentiles are to be computed by the choice of no, one or two conditioning variables in the initial dialogue of RAPPORT.

   - If no conditioning variables are specified in the initial dialogue, a single table of percentiles is computed and displayed. It is the distribution of the ability in the total population. In the example with different cycles it gives the percentiles in the superpopulation.
   - If only the first conditioning variable is specified (but no second one), a percentile table is displayed for each category of this conditioning variable. For example, if cycle is the first conditioning variable, the percentiles in each cycle are computed and displayed in a separate table for each cycle.
   - If a first and a second conditioning variable are specified, the same output as in the preceding bullet is generated, and moreover the percentiles for each category of the second conditioning variable are computed. If cycle is the first conditioning variable and gender the second, there will be a table for each cycle. Each table contains the percentiles for the cycle as a whole (first column), followed by the percentiles for the boys in that cycle and for the girls.

- If only the second conditioning variable is specified (but no first one), the percentiles for the total population will be displayed, followed by the percentiles for each category of the conditioning variable. Example: if cycle is the second conditioning variable (and there is no first one), the percentiles in the total population and in each cycle will be displayed in a single table. Notice that in this case, cycle will not have the effect of the first conditioning variable when reporting proportions, means and standard deviations.

When the user is asked to answer to the questions in the initial dialogue, a list of variable or category labels is displayed on the screen as a memory aid. Each label is preceded by a number. The answer requested from the user is a number, not a label.

We conclude this chapter with a warning to the user. In computing and reporting a regression analysis, much attention has been given to a reasonable estimate of the standard errors of the parameter estimates. In the reports on the marginal means and the percentiles, no standard errors are mentioned (in the present version). A marginal mean is a weighted average of regression parameters. If the weights are fixed, the standard errors of the marginal means can easily be computed using the variance-covariance matrix of the regression parameter estimates (and this matrix is estimated consistently by the program SAUL). But the problem is complicated by the fact that the weights are the estimated proportions in the table of the joint distribution of the regressors, and these weights also have an estimation error, which are not easily computed in the general case where different amounts of information are used for different regressors (see the section on the construction of the JOBN.WBO file). Because we could not solve this problem in a satisfactory way, any report on the standard errors for marginal results has been omitted. But this does not imply that the results reported are error free. The only advice we can give to the user is to be careful with these results, especially in cases where many regressors are used with a moderate sample size.