STATISTICAL REPORT

Logistrule:
A knowledge-based system for logistic regression

by

Jan H. Aarseth and Ivar Heuch

Report no. 28
July 1996

Department of Mathematics
UNIVERSITY OF BERGEN
Bergen, Norway

Department of Mathematics
University of Bergen
N-5007 Bergen
Norway

ISSN 0333-1865

Logistrule:

A knowledge-based system for logistic regression

Jan H. Aarseth and Ivar Heuch

Statistical Report No. 28
July 1996

Table of contents

1. Introduction

The program system Express constitutes a tool for constructing knowledge-based statistical systems requiring repeated cycles of statistical analysis on given data sets. Express makes it possible to specify chains of rules in such a way that intermediate results determined by execution of standard statistical packages may affect the type of analysis to be carried out in subsequent steps. These intermediate results are assumed to occur in the ordinary output file produced by the packages involved, and they are stored as slot values in a special data base maintained by Express. If the rules refer to statistical quantities which have not yet been found for the data set in question, the system will automatically generate control language for the appropriate external package, initiate execution of this package, scan the output, and store the relevant results. In combination with a particular stack of rules, the data base constitutes the working memory of Express. At any stage in the analysis, this memory keeps track of the current state, what has already been done and what should be done next.

A slot may be defined as the value of a test statistic or the response to a particular question, given by the user or the set of rules itself. A third possibility is a "block slot" containing, for example, a plot extracted from the output produced by a package, or any other collection of text. Slots in the knowledge base are defined by giving the particular quantity in question a name, and by providing instructions indicating which rule can determine its value. These definitions are located in a particular code file. In addition to the code file, the knowledge base contains several sets of rules. During execution of the set of rules, the inference engine will consult the data base to determine whether the value of the slot is already known. If this is not the case, the rule number attached to the slot will be read from the code file, which makes it possible to execute the proper rule. A previous version of Express was described by Carlsen and Heuch (1985). General information about the present version was given by Heuch *et al.* (1990) and Aarseth and Heuch (1996a).

The knowledge that can be incorporated into Express is structured as different sets of rules relating to various areas in statistical data analysis. Several simple sets of rules have been written. In her thesis for the cand. scient. degree, Irgens (1991) constructed an extensive set of rules for model building in logistic regression, which could, however, only handle 2 independent variables. This limitation was imposed mainly because of technical restrictions in the version of Express available, which have subsequently been removed. In addition to minor adjustments, the main difference is the capability of the current version to handle sets of rules for which the number of variables has not been fixed in advance. This change required fundamental modifications of the basic programming structure, especially for the interface between the system and the external packages.

In testing the new version of Express, it was natural to attempt to extend the rules relating to logistic regression. We wanted to increase the maximum number of independent variables to 10. In contrast to Irgens (1991), we decided to concentrate mostly on the model selection procedure proposed in the book "Applied logistic regression" by Hosmer and Lemeshow (1989). The main purpose was to investigate whether it was possible to represent the knowledge in Hosmer and Lemeshow's procedure within the Express framework. Constructing an improved user interface was also important. During the revision we found it necessary to rewrite all rules, and many additional rules were implemented. The new set of rules has been given the name Logistrule, which now constitutes a knowledge-based system, constructed using Express as the main tool. The technical details of the knowledge base are described in a separate report (Aarseth, 1996).

The statistical program system BMDP is widely used and incorporates all necessary tools for model fitting applied in the analysis performed by Logistrule. This is why BMDP was selected as the underlying statistical software.

The current standard version of Express runs on a PC with DOS 3.0 or higher. The installation of Express, described by Aarseth and Heuch (1996a), also includes Logistrule.

2. Description of Logistrule

Logistic regression is a well known method applied in several areas. In medical research, the method is typically used to describe relations between different potential risk factors and the occurrence of disease. Assume that the components of a $p\times1$ vector $\boldsymbol{x}$ represent the different risk factors, which are regarded as independent variables in the regression model. The outcome variable Y is equal to 1 if a particular disease is present, otherwise 0. Let the conditional probability for disease be denoted by $P(Y = 1 \mid \boldsymbol{x}) = \pi(\boldsymbol{x})$. Since Y can only take on the two values 0 or 1, this implies that $E(Y \mid \boldsymbol{x}) = \pi(\boldsymbol{x})$. The logistic regression model is of the form

$$\pi(\boldsymbol{x}) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

where β is a $p\times1$ vector of parameters. The frequently used logit transformation of $\pi(\boldsymbol{x})$ is

$$g(\boldsymbol{x}) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta^T \boldsymbol{x} .$$

This is the logarithm of the odds, which in turn is the probability of disease being present divided by the probability of disease not being present. The model described above tacitly assumes that the independent variables are defined on an interval scale. For categorical variables, new design variables must be created, representing the contributions of the separate levels. The number of design variables for a categorical variable with k levels is equal to k-1.

Given a collection of independent variables, the aim of Logistrule is to select a subset of essential variables, to be included, on the correct scale, in a logistic model. The strategy for building a particular model as proposed by Hosmer and Lemeshow (1989) can be divided into five stages:

(1) Univariate analysis of each independent variable.
(2) Multivariate fit and selection of independent variables for the final model.
(3) Scaling of each selected variable and inclusion of interactions.
(4) Investigation of goodness of fit for the final model.
(5) Interpretation of the final model.

Each stage is divided into substages. We will next describe each stage as it has been implemented in Logistrule. Deviations from the procedure described by Hosmer and Lemeshow will be pointed out.

2.1 Univariate analysis

The selection of independent variables for the final model begins with a univariate analysis of each variable. Usually the first step is a contingency table analysis of outcome (y=0,1) versus the k levels of the independent variable. For independent variables with very many different levels (in reality continuous variables), a logistic model containing the dependent variable and the selected independent variable only is fitted.

During execution of Logistrule, the user will notice the additional work carried out to maintain control of the data. Several files are generated during the analysis to keep a simple interface to the external program BMDP. Most files needed are created by the BMDP Data Manager in BMDP format. One file is generated at the start of the univariate analysis for each independent variable. Each such file also includes the dependent variable.

The dependent variable

Before the analysis of the separate independent variables can begin, the outcome variable must be prepared. If this variable has only two levels, the lower one will be taken as absence of disease (and is recoded with the value 0) and the higher value is understood to represent presence of disease (and is recoded with the value 1). If the number of levels exceeds 2, the user is asked for a cut-point. If not given by the user, the cut-point is chosen by Logistrule as the mid-point between the minimum and maximum values.

During the preparation, BMDP is used for recoding, if necessary, to create a BMDP file with codes 0 and 1. This file is used by the BMDP Data Manager when other files including the dependent variable are created.

User selections regarding independent variables

The selection of independent variables for inclusion in the multivariate model should not only depend on statistical results. Thus, there may be several biologically important covariates which must be included in the model in any case. Furthermore, we often want to concentrate on another variable of particular interest, e.g. a potential risk factor for a disease. Logistrule must take this into consideration. For these reasons we found it natural to subdivide the set of independent variables into three categories:

1 - The study variable.
2 - Adjustment variables.
3 - Ordinary variables.

At the beginning of the univariate analysis, the user must select for each independent variable one of the three types indicated. The user is free to decide the type of any particular variable, but there must be one and only one study variable. If the decision is left to Logistrule, type 3 is selected. The study variable is supposed to be the independent variable of primary interest when the relations with the dependent variable are examined. This is why the study variable will always be included in the final model. Furthermore, this variable plays an important role in the study of interactions and in the final interpretation. Typically we also want to incorporate additional variables into the model, regardless of statistical significance. These must be defined as adjustment variables. One example is age, which is regarded as important in almost any epidemiological research. The last group of variables, referred to as ordinary variables, will only be included on statistical grounds.

The other decision left to the user regarding any independent variable is whether the variable is nominal or interval scaled. It is difficult for a program system to detect the reasons behind the coding of a variable. This is why the user should preferably give this information. If Logistrule must make the decision, two properties of the variable are taken into account. First the variable in question is examined for observations coded with decimals. As it is unlikely that a nominal variable should be coded with decimals, the variable will be regarded as interval scaled if decimals are present. If no decimals are found and there are no more than 8 distinct values, the number of observations divided by the number of levels is used in the decision. If this ratio is less than 2, Logistrule will not regard the variable as nominally coded. In this case it is assumed that the number of observations in at least one cell in a categorical analysis will be too low to produce valid results.

If the user decides that a variable should be treated as nominal, the highest number of levels accepted is 8. If this limit is exceeded, a warning is given and the variable is excluded from the further analysis.

Plot of observed risks

Regardless of the selections made so far, the system will continue with logistic regression using BMDP LR, with the relevant variable for the time being considered as categorical. This is done in order to start the univariate analysis with a plot of observed risks. Such a plot may reveal important structures in the data. If the variable has more than 8 different levels, the median, upper and lower quartiles (Q_3 and Q_1) will be used as cut-points for the categorical variable. In addition to the plot of observed risks, the results of the fitted logistic model are extracted from the output. These are likely to be useful in subsequent analysis.

The statistical significance slot

An important feature of the univariate analysis is a special slot which is used to keep track of the statistical significance of a variable. This slot, designated the "summary indicator of significance", can be assigned the values 1, 2 or 3. The lowest value shows that the variable is not statistically significant. The value 3 indicates statistical significance, and in this case the variable will be included automatically at the start of the multivariate analysis. The value 2 represents an intermediate situation with doubt about the significance. Exact definitions are given below. It should be kept in mind that the study and adjustment variables may be considered in the multivariate analysis even without a clear statistical significance.

Univariate analysis for nominal variables

In the remainder of the univariate analysis, a distinction is made between nominal and interval scaled variables. For nominal variables we already have obtained the results of the appropriate logistic analysis, but as recommended by Hosmer & Lemeshow (1989), we also want to study the data by a traditional contingency table approach if possible. Expected values will be calculated in order to determine whether the chi-square approximation in the contingency table is adequate or not. If the variable considered has only 2 levels, the contingency table analysis is always performed, using the Fisher exact test. Otherwise the following guidelines are commonly accepted (Brown, 1990): No cell should have an expected value less than 1, and not more than 20% of the cells should have expected values less than 5. If these assumptions are satisfied, BMDP 4F is used by Logistrule to execute the analysis. The value of the "summary indicator of significance" for the variable will be determined on the basis of the likelihood ratio chi-square statistic if the number of levels exceeds 2, otherwise the Fisher exact test is used.

In contrast, if the assumptions above are violated, we simply use the likelihood ratio test already computed in the univariate logistic regression. It should be noted that this procedure leads to exactly the same *p*-value as found by BMDP 4F, although the contingency table is also intended for more informal interpretation of risk by the user. Anyhow, the fitted model may have unstable estimates in this case and the chi-square approximation may be poor. This warning is included by Logistrule in the summary of the analysis, but no action is taken to deal with the problem. This could be rectified by performing exact analysis or Monte Carlo estimation (StatXact, 1989). Irgens (1991) incorporated such methods into her set of rules, using the package StatXact, but this procedure is beyond the scope of the current set of rules. Table I summarizes which tests are executed in the case of a nominal variable.

In this situation, the "summary indicator of significance" slot can only take on the values

1 or 3. If the proper univariate test in Table I gives a *p*-value < 0.25, the value 3 is assigned, otherwise 1. Among ordinary variables, only those with "summary indicator of significance" equal to 3 will be considered in the multivariate analysis. It may seem strange to use such a high value as 0.25 as the significance level, but it has been shown that a lower, more traditional level often fails to identify variables which are known to be important (Bendel & Afifi, 1977; Mickey & Greenland, 1989).

	# levels = 2	# levels > 2
Assumptions fulfilled	Fisher exact test	Likelihood ratio test in contingency analysis
Assumptions not fulfilled	Fisher exact test	Likelihood ratio test in logistic regression

Table I Tests used in the univariate analysis for a nominally coded variable. The selection depends on the conditions for satisfactory chi-square approximation for the likelihood ratio test in a contingency table.

Univariate analysis for interval scaled variables

If a variable is interval scaled, our strategy is different. The test selected will again depend on the number of distinct values of the variable and the conditions for an adequate approximation to the chi-square distribution in a contingency table analysis. As shown in Table II, the Fisher exact test is once more selected if the independent variable is dichotomous. In this case the "summary indicator of significance" is given the value 3 if the *p*-value is less than 0.25, otherwise the value 1.

	# levels = 2	2 < # levels ≤ 8	# levels > 8
Assumptions fulfilled	Fisher exact test	Likelihood ratio test in contingency table and test for trend	Likelihood ratio test in logistic regression and test for departure from a linear model
Assumptions not fulfilled	Fisher exact test	Likelihood ratio test in logistic regression and test for departure from a linear model	

Table II. Tests carried out in univariate analysis for an interval coded variable. For variables with less than 8 levels the chi-square approximation is critical when selecting test.

If the number of levels exceeds 2 but is less than or equal to 8, we use, as recommended by Hosmer and Lemeshow (1989), an analysis based on a contingency table if the conditions are satisfied. This approach provides a rather descriptive analysis, even for interval scaled variables, with increased possibilities for user interaction. Two tests are applied. First the usual likelihood ratio test statistic is computed for homogeneity. As the independent variable has levels in a natural order, the overall chi-square test may waste important information. A more sensitive way of detecting alternative hypotheses is to test for a trend in disease risk (Breslow & Day, 1980). BMDP 4F carries out the linear trend test proposed by Cochran (1954). These two tests determine the statistical importance of the variable as shown in Table III.

	Overall likelihood ratio test	
Test for trend	p-value < 0.25	p-value ≥ 0.25
p-value < 0.25	3	3
p-value ≥ 0.25	2	1

Table III. Assignment of values to the "summary indicator of significance" slot when the number of levels is in the range 3 to 8 and assumptions for contingency table analysis are met.

If there is a significant trend, Logistrule will assign the value 3 to the "summary indicator of significance", regardless of the result of the overall likelihood ratio test. The variable will then be passed on to the multivariate analysis without any change in scale. In contrast, variables without a significant trend are assigned slot values 1 or 2, depending on the likelihood ratio test. With a *p*-value less than 0.25, a relationship, probably non-linear, is indicated with the dependent variable. Thus the independent variable should not be excluded from the remaining analysis but rather be considered in the multivariate logistic model as categorical. This also applies to study or adjustment variables with a "summary indicator of significance" equal to 1. Variables for which Logistrule decides to change the type from interval scaled to categorical, are grouped according to the general rules given below.

If the assumptions for contingency table analysis are not fulfilled, the single degree of freedom likelihood ratio test for an association in the logistic regression model is applied. Hosmer and Lemeshow (1989) simply state that it is common practice at this stage to assume linearity in logit. Logistrule follows this approach but also performs a global test of departure from the linear model. This test compares the likelihood (by computing the difference in deviance) of the two models including the independent variable as categorical and interval scaled, respectively. Table IV shows how the "summary indicator of significance" slot depends on the two tests.

	Test for departure from linear model	
Test for Trend	p-value < 0.25	p-value ≥ 0.25
p-value < 0.25	2	3
p-value ≥ 0.25	2	1

Table IV. Assignment of values to the "summary indicator of significance" slot for an interval coded variable with more than 8 levels, or with fewer levels when assumptions for contingency table are not fulfilled.

If an interval scaled independent variable has more than 8 levels, a similar strategy is used, but the global test for departure from the linear model involves a derived grouped variable. The lower quartile, the median and the upper quartile are used as cut-points in this grouping. The "summary indicator of significance" slot depends on the two tests in the same way as for interval scaled variables with 8 or fewer levels when the assumptions for contingency table analysis are not fulfilled (see Table IV). Thus only when the test for trend is significant and there is no evidence for departure from a linear model is the "summary indicator of significance" assigned the value 3. If no trend is found and no evidence for departure from linearity, this indicator should clearly be set to 1. It is more difficult to decide on the correct action if departure from a linear model is indicated. In view of the possible violation of assumptions, Logistrule then sets the "summary indicator of significance" equal to 2, regardless of the outcome of the trend test. This indicates that the variable considered should be regarded as a candidate for the final multivariate model, and variables with at most 8 levels are regarded as categorical after grouping.

Computation of tests

For variables coded on a nominal scale, Logistrule extracts the *p*-values for the respective tests directly from the output produced by BMDP. This is also done when an interval scaled variable has less than 9 levels and the assumptions for contingency table analysis are fulfilled. The other tests require results from several logistic regression models which must be run separately by BMDP. Thus, for the single degree of freedom test for trend for an interval scaled variable, we extract from the BMDP output the log likelihood l_0 for the model containing only the constant and the log likelihood l_1 for the model containing the independent interval scaled variable. With corresponding likelihood values L_0 and L_1, the test statistic can be expressed as

$$G = -2\ln\left[\frac{L_0}{L_1}\right] = -2[\, l_0 - l_1]$$

This statistic is computed by Logistrule, and a separate Fortran function for probabilities in the chi-square distribution (Bhattacharjee, 1970) is used to find the *p*-value. The test for departure from linearity requires a similar computation comparing likelihood values from one model including the relevant variable as categorical and another model where the variable is interval scaled.

Summary slots

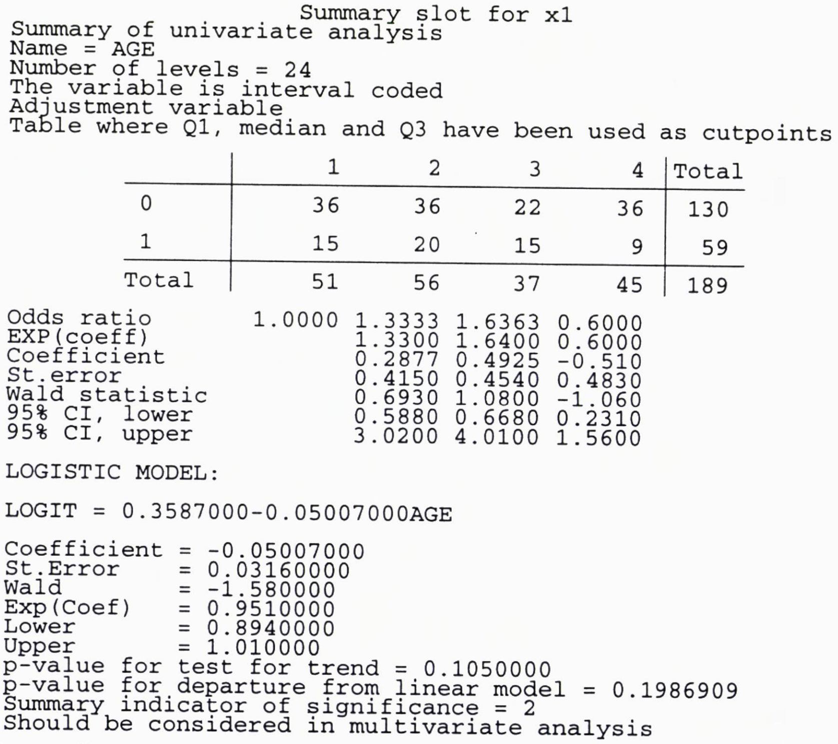
Several runs of BMDP must be started in order to obtain all values of interest, and a number of results appear on the screen. In view of the large amount of information, Logistrule sets up a summary of the univariate analysis for each independent variable.

Figures 1 and 2 give two examples of summary slots. Univariate results are presented for two variables in the analysis of the low birth weight data set used by Hosmer and Lemeshow (1989; Appendix 1) as an example of model building. The data set includes information about 189 women and was collected at Baystate Medical Center, Springfield, Massachusetts, during 1986. The outcome variable, "low birth weight", is of interest partly because infant mortality and birth defect rates are very high for low birth weight babies. The study tried to determine which factors relating to a woman's behaviour during pregnancy could alter the chance of delivering a baby of low birth weight. The independent variables were (with names used in Logistrule): Age of the mother (AGE), weight in pounds at the last menstrual period (LWT), race (RACE), smoking status during pregnancy (SMOKE), history of premature labour (PTL), history of hypertension (HT), presence of uterine irritability (UI) and number of physician visits during first trimester (FTV). In our analysis, LWT will be the study variable and AGE is an adjustment variable. The remaining variables are regarded as ordinary.

Figure 1 shows how an interval scaled variable with more than 8 levels is treated by Logistrule. Following standard information about variable name, number of levels and variable type, a frequency table is printed. Because the variable AGE has more than 8 levels, the lower quartile, median and the upper quartile are used as cut-points. The log file from this run shows that the cut-points are 19, 23 and 26 years. A subroutine in Logistrule (not using external packages) determines frequencies and computes crude odds ratios with the lowest level as reference. These values are listed below the frequency table. The row marked "EXP(coeff)" displays odds ratios found by BMDP LR, calculated from the coefficients in the model incorporating the grouped categorical variable AGE. Furthermore, regression coefficients of the corresponding design variables are given, with standard errors. The Wald statistic, defined as

$$W = \frac{\beta_1}{S\hat{E}(\beta_1)}$$

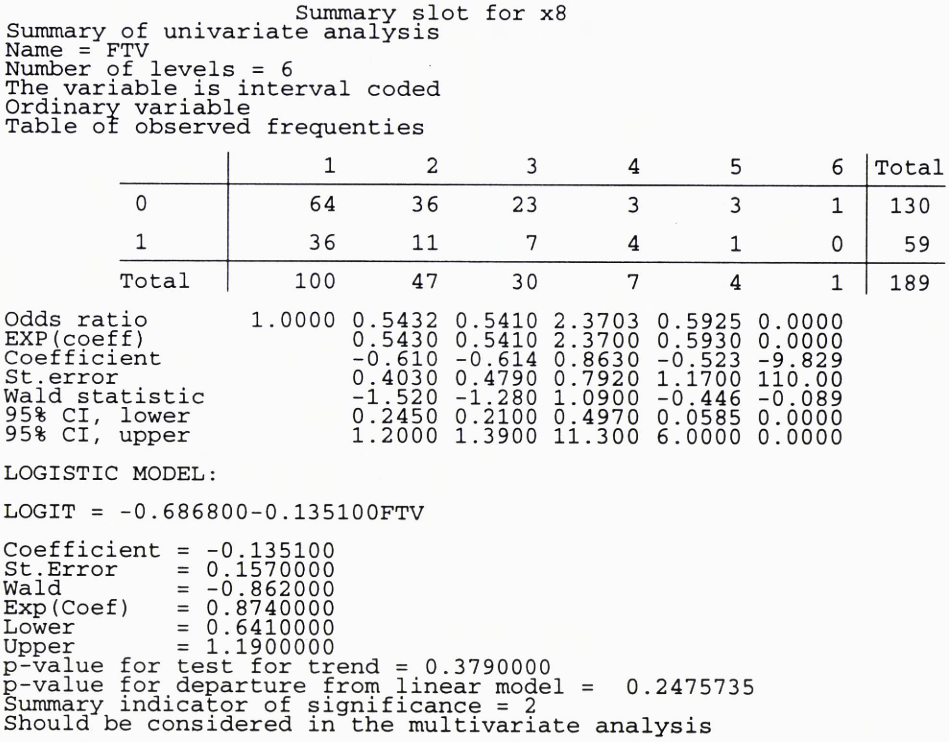
can be used to test the significance of the variable, assuming that W follows an approximate standard normal distribution under the hypothesis that the parameter β_1 is equal to zero. The last pair of rows in the summary give lower and upper 95% confidence limits for the odds ratio. Results are then presented from the fitted model with AGE regarded as interval scaled, considering all 24 levels without grouping. At the end of the summary, the *p*-values of tests used to determine the "summary indicator of significance" are printed. In this case non-linearity in logit is suggested, so the indicator is assigned the value 2. In fact, an informal interpretation of the crude odds ratios shown in Figure 1 suggests that the risk of a low birth weight baby may increase in this data set with an increasing mother's age, with a drop in risk, however, for mothers older than 26 years.



```
                       Summary slot for x1
Summary of univariate analysis
Name = AGE
Number of levels = 24
The variable is interval coded
Adjustment variable
Table where Q1, median and Q3 have been used as cutpoints
                 |     1       2       3       4 | Total
        ---------+-------------------------------+------
          0      |    36      36      22      36 |  130
          1      |    15      20      15       9 |   59
        ---------+-------------------------------+------
        Total    |    51      56      37      45 |  189
Odds ratio          1.0000 1.3333 1.6363 0.6000
EXP(coeff)                 1.3300 1.6400 0.6000
Coefficient                0.2877 0.4925 -0.510
St.error                   0.4150 0.4540 0.4830
Wald statistic             0.6930 1.0800 -1.060
95% CI, lower              0.5880 0.6680 0.2310
95% CI, upper              3.0200 4.0100 1.5600

LOGISTIC MODEL:

LOGIT = 0.3587000-0.05007000AGE

Coefficient = -0.05007000
St.Error    = 0.03160000
Wald        = -1.580000
Exp(Coef)   = 0.9510000
Lower       = 0.8940000
Upper       = 1.010000
p-value for test for trend = 0.1050000
p-value for departure from linear model = 0.1986909
Summary indicator of significance = 2
Should be considered in multivariate analysis
```

Figure 1. Summary slot for the variable AGE from the low birth weight data analysed by Express.

Figure 2 shows the summary slot for the interval scaled variable FTV, which represents the number of physician visits during first trimester. This variable has only 6 levels and thus no grouping is needed. The assumptions for analysis by a contingency table are not met, and the tests used to determine the value of the "summary indicator of significance" are those shown in Table IV. This slot has been assigned the value 2 because the *p*-value 0.379 in the test for trend is above 0.25 and the test for departure from a linear model gives a *p*-value 0.248, just below 0.25.



```
                         Summary slot for x8
Summary of univariate analysis
Name = FTV
Number of levels = 6
The variable is interval coded
Ordinary variable
Table of observed frequenties

         |     1      2      3      4      5      6 |Total
---------+------------------------------------------+------
 0       |    64     36     23      3      3      1 |  130
 1       |    36     11      7      4      1      0 |   59
 Total   |   100     47     30      7      4      1 |  189

Odds ratio          1.0000 0.5432 0.5410 2.3703 0.5925 0.0000
EXP(coeff)                 0.5430 0.5410 2.3700 0.5930 0.0000
Coefficient                -0.610 -0.614 0.8630 -0.523 -9.829
St.error                   0.4030 0.4790 0.7920 1.1700 110.00
Wald statistic             -1.520 -1.280 1.0900 -0.446 -0.089
95% CI, lower              0.2450 0.2100 0.4970 0.0585 0.0000
95% CI, upper              1.2000 1.3900 11.300 6.0000 0.0000

LOGISTIC MODEL:

LOGIT = -0.686800-0.135100FTV

Coefficient = -0.135100
St.Error    = 0.1570000
Wald        = -0.862000
Exp(Coef)   = 0.8740000
Lower       = 0.6410000
Upper       = 1.1900000
p-value for test for trend = 0.3790000
p-value for departure from linear model =   0.2475735
Summary indicator of significance = 2
Should be considered in the multivariate analysis
```

Figure 2. Summary slot for the variable FTV from the low birth weight data analysed by Express.

Grouping

Those interval scaled variables which are candidates for the multivariate model despite the lack of a significant trend in the univariate analysis, are subjected to a particular procedure for combining categories. For a variable with more than 8 levels, linearity in logit is assumed until the end of the selection of variables (Section 2.3). However, for variables with 8 or fewer levels, a grouping is performed to avoid categories with very low frequencies. Such variables are considered as categorical for the remainder of the analysis.

As Hosmer & Lemeshow (1989) do not provide any guidelines, a particular procedure for grouping of categories was devised. The procedure is iterative, with one category being combined with another neighbouring category in each cycle. At the beginning of each iteration, only categories with less than 10% of the observations and categories for which all individuals belong to the disease state or to the non-disease state (zero count cells), are considered candidates for grouping. The actual decision whether a particular category should be combined with a neighbouring category or not, is made on the basis of a particular grouping score computed for each candidate category. Low values of the score indicate a high priority for grouping.

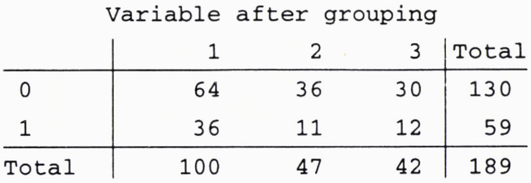
The score consists of four terms:

$$G = I_{zc} + I_{\text{freq}} + B + C$$

Here I_{zc} is equal to 0 if the group includes a zero count cell, otherwise 1. Similary I_{freq} is equal to 0 if the total count in the group is less than 10% of the observations. The term B is designed to make it easier for neighbouring groups with rather similar odds ratios to be combined. First, the odds ratios associated with the next higher and next lower levels are computed, using the level in question as reference. Any odds ratio less than 1.0 is inversely transformed so that odds ratios are comparable. If the lowest neighbouring odds ratio, defined in this way, is less than 2.0, then B is assigned the value 1, if it is greater than 3.0, then B is set to 3, otherwise B is assigned the value 2.

The last term C is supposed to make it easier for the extreme lowest level of the variable to be combined with the second lowest one. Thus C is defined as 1 for the lowest level, 2 for the second lowest and 3 for the remaining levels. The particular definition for the second lowest level is needed when this category has relatively few observations and should be combined with the lowest one. In certain situations, however, the lowest level is a particular ''non-exposed level'' that should not in any case be combined with other categories, as for example, non-smoking compared to different levels of smoking. In general, the user is given the opportunity to prevent such combinations. Although not implemented in the grouping procedure, a similar contribution to G could also be defined for the extreme highest level.

When the score G has been computed for all categories with a low percentage of observations or a zero count cell, the group with the smallest value of G is combined with the neighbouring group with the most similar risk. The next iteration cycle is then started, with the calculation of percentages in each group and a check for zero count cells. The process terminates when all categories contain at least 10% of the observations and no zero count cells remain. Figure 3 indicates how Logistrule presents the grouped variable corresponding to FTV in the example. In this case, levels 1 and 2 have been left unchanged, while the original levels 3-6 have been combined into a new group. More accurate information about the grouping is given in the log file.



```
               Variable after grouping
              |        1        2        3  |Total
--------------+-----------------------------+------
0             |       64       36       30  | 130
1             |       36       11       12  |  59
--------------+-----------------------------+------
Total         |      100       47       42  | 189
```

Figure 3. Variable FTV after grouping at the end of the univariate analysis.

2.2 Multivariate fit and selection

An initial multivariate model is now set up including the study variable, all adjustment variables and any ordinary variables which were not discarded in the univariate procedure. If no study variable was specified during the univariate analysis, the first variable considered in the multivariate model will be selected by Logistrule as a default. A new BMDP file is created, including all relevant variables. This file is used throughout the remainder of the analysis, although the file may be modified if it is found necessary to rescale a variable, or if observations are removed in the assessment of the fit of the model.

Multicolinearity

Before the model fitting can start, a simple check is made for multicolinearity among the independent variables. Multicolinearity occurs if any such variable can be expressed approximately as a linear combination of the remaining independent variables. A model incorporating multicolinearity can be highly unreliable. Hosmer and Lemeshow (1989) discuss how similar problems are handled in linear regression and recommend in particular a corresponding in-depth investigation if the fit results in extremely large estimated standard errors (or sometimes very large estimated coefficients).

Although special tools for diagnosing colinearity have been extended to logistic regression (Wax, 1992), they are not yet incorporated into standard software. As a rather crude diagnostic, Logistrule computes ordinary correlations between all independent variables. If any pair of variables has a absolute correlation higher than 0.98, one of the variables involved may be removed. The user can decide which one, if any. If the decision is left to the system, the following default rule applies. First, the study variable cannot be removed. Second, if a particular variable is included in several highly correlated pairs, this variable will be removed before any others. If two or more variables are included in the same number of highly correlated pairs, ordinary variables will be removed before adjustment variables. Although this approach is not very efficient for categorical variables without any natural ordering, a high correlation detected in the data still serves as a warning.

The multivariate model

The multivariate analysis is based on the "E,V,W model" of Kleinbaum (1994). If the maximum number of independent variables is restricted to 10, as required by Logistrule, the logit can be expressed in this model as

$$\beta_0 + \beta_1 e + \beta_2 v_1 + \beta_3 v_2 + \ldots + \beta_{10} v_9 + \beta_{11} ew_1 + \ldots + \beta_{14} ew_4 ,$$

where β_1 is the coefficient of the "exposure variable" e, the terms $\beta_2, \beta_3, \ldots, \beta_{10}$ are the coefficients of other independent variables $v_1, v_2, \ldots, v_9$, and $\beta_{11}, \ldots, \beta_{14}$ are coefficients of the interaction effects. Our "study variable" corresponds to the "exposure variable" of Kleinbaum (1994), whereas adjustment and ordinary variables belong to Kleinbaum's second category. Pairwise interactions are only considered between the study variable and any other independent variable.

Each potential interaction term handled by Logistrule requires about 100 slot definitions and 50 records for storage of slot values in internal files. In view of the space limitations, it is assumed that at most 4 interaction terms are needed in the model. This assumption is unlikely to be violated in practical applications.

The expression for the logit must be modified in an obvious way for categorical variables. A single term is then replaced by several terms involving design variables, introduced according to the standard conventions in BMDP LR (Brown, 1990). In particular, the reference category is the one with the lowest code.

Selection of independent variables

The selection of variables starts with fitting a model which includes all independent variables remaining after the univariate analysis.

Table with Wald statistic step 1

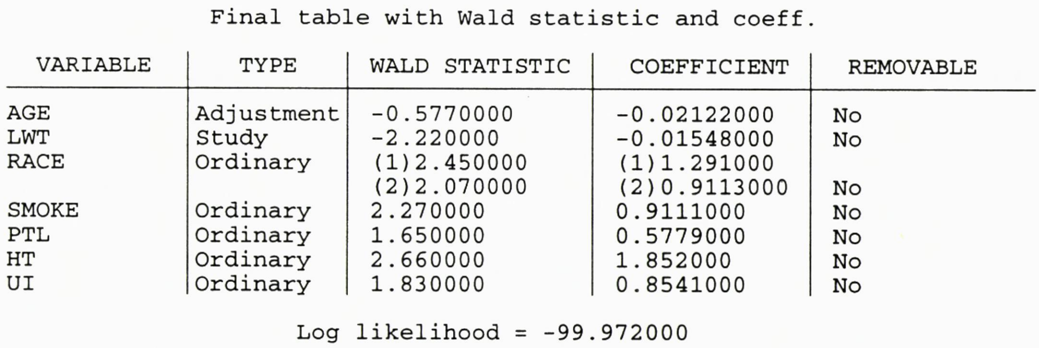
VARIABLE	TYPE	WALD STATISTIC	COEFFICIENT	REMOVABLE
AGE	Adjustment	-0.5710000	-0.02152000	No
LWT	Study	-2.010000	-0.01460000	No
RACE	Ordinary	(1)2.390000 (2)1.970000	(1)1.282000 (2)0.9062000	No
SMOKE	Ordinary	1.990000	0.8395000	No
PTL	Ordinary	1.860000	0.6924000	No
HT	Ordinary	2.510000	1.817000	No
UI	Ordinary	1.760000	0.8235000	No
FTV	Ordinary	(1)-0.6770000 (2)0.1550000	(1)-0.3164000 (2)0.08212000	Yes

Log likelihood = -98.704000

Figure 4. Multivariate model containing all variables not excluded in the univariate analysis.

Figure 4 shows how Logistrule summarizes the fit in a particular slot. The variables RACE and FTV are considered as categorical, FTV grouped as in Figure 3 and RACE with codes 0 (white), 1 (black) and 2 (other). The last column indicates whether any variables can be removed from the model. As recommended by Hosmer & Lemeshow (1989), the Wald statistic is used for this test, with threshold values ±1.645, corresponding formally to a significance level of 0.1, based on a standard normal distribution. Thus, ordinary variables may be removed if their Wald statistic is below 1.645 in absolute value. Adjustment and study variables are always retained. A categorical variable can only be removed if all corresponding design variables are non-significant. In our example, the variable FTV can be removed.

The next step is to refit the model without variables marked as removable. However, before we can conclude that these variables make no essential contribution, potential confounding effects must be investigated. These will be expressed by differences in the coefficients of the study variable, according as covariates are included or not. A necessary condition for confounding is an association between the covariate and both outcome variable and study variable. As recommended by Hosmer & Lemeshow (1989), the value of the coefficient of the study variable for the full model is compared to that for the model excluding the potentially confounding variables. The change is measured by the ratio of the larger to the smaller of the two absolute coefficients. If this ratio exceeds 1.2, some removed variables should be reintroduced into the model. This is done by refitting several models, reentering the removed variables one-by-one. For each such variable, the change in the regression coefficient of the study variable is measured in the same way to decide whether confounding is present or not. If any variables are still removed at the end of this process, a new model is fitted and non-significant variables are again eliminated unless they are confounders. When this process has finished, the user is presented with a summary slot showing a model including only those variables that are significant (by our formal criteria) in the multivariate version.



```
                 Final table with Wald statistic and coeff.

  VARIABLE     |    TYPE    | WALD STATISTIC  |   COEFFICIENT   |  REMOVABLE
---------------------------------------------------------------------------
AGE            |Adjustment  | -0.5770000      | -0.02122000     | No
LWT            |Study       | -2.220000       | -0.01548000     | No
RACE           |Ordinary    | (1)2.450000     | (1)1.291000     |
               |            | (2)2.070000     | (2)0.9113000    | No
SMOKE          |Ordinary    | 2.270000        | 0.9111000       | No
PTL            |Ordinary    | 1.650000        | 0.5779000       | No
HT             |Ordinary    | 2.660000        | 1.852000        | No
UI             |Ordinary    | 1.830000        | 0.8541000       | No

                      Log likelihood = -99.972000
```

Figure 5. Model containing only variables which are significant in the multivariate analysis.

Figure 5 indicates which variables were selected by this procedure in our example with low birth weight. The removed variable FTV was not found to be a confounder.

2.3 Scaling and interactions.

Scaling of independent variables

When a model has been obtained which presumably includes all essential independent variables, the assumption about linearity in logit is checked for each interval scaled variable. One procedure proposed by Hosmer and Lemeshow (1989) is to add the term $x\ln x$ to the model, and if the coefficient for this variable is significant, non-linearity in x is indicated. This test is carried out by Logistrule and the results are presented, although no decisions are made on this basis. Hosmer and Lemeshow also suggest introducing three design variables using the quartiles as cut-points and comparing estimated odds ratios (with the lowest group as reference). This is an intuitive procedure which is difficult to incorporate in the set of rules.

We have adopted a different approach in Logistrule, comparing two multivariate models for each interval scaled variable. In the first model, the variable in question is regarded as categorical, in the second model as interval scaled. A likelihood ratio test for global departure from the linear model is carried out. If the particular independent variable has more than 8 levels, the quartiles will be determined and a grouped variable is used in assessing departure from linearity. As in the univariate situation, a *p*-value is computed separately using the deviance. If the *p*-value is less than 0.10, the conclusion is that the linearity assumption is not met. In that case, the variable in question is grouped and considered as categorical during the remainder of the analysis.

How the grouping is actually carried out at this stage, depends on the number of levels. Several factors must be taken into account when an interval scaled variable is grouped (Breslow and Day, 1980; Section 3.3). Introducing about 4 exposure levels is usually sufficient. A simple dichotomous variable may on the other hand conceal more information that it reveals.

For variables with more than 8 levels, Logistrule uses the quartiles as cut-points, creating 4 categories. No account is taken of a possible "non-exposed category", as factors with so many levels are more likely to be genuine continuous variables. In contrast, for variables with at most 8 levels, exactly the same grouping algorithm is applied as in the univariate analysis. In this case, the "non-exposed category" remains separate for the study variable. For other variables, a lowest level with a zero count cell or very few observations may be combined with higher levels. Logistrule goes through all variables separately, considering the scale. A new model, designated the "model after refinement", is then fitted and presented.

Figure 6 shows the corresponding model in the low birth weight example. The variables LWT and PTL have changed their status in comparison to the model in Figure 5. The 75 levels of LWT have been combined into 4 groups. The variable PTL had originally 4 levels but only 2 remain after grouping.

```
                  Table with Wald statistic after scaling
   VARIABLE   |    TYPE    | WALD STATISTIC  |   COEFFICIENT   |  REMOVABLE
------------------------------------------------------------------------------
AGE           | Adjustment | -0.9760000      | -0.03748000     | No
LWT           | Study      | (1)-1.810000    | (1)-0.8833000   |
              |            | (2)-1.660000    | (2)-0.8436000   |
              |            | (3)-2.000000    | (3)-1.096000    | No
RACE          | Ordinary   | (1)2.180000     | (1)1.141000     |
              |            | (2)1.890000     | (2)0.8538000    | No
SMOKE         | Ordinary   | 1.870000        | 0.7636000       | No
PTL           | Ordinary   | (1)2.950000     | (1)1.394000     | No
HT            | Ordinary   | 2.280000        | 1.543000        | No
UI            | Ordinary   | 1.650000        | 0.7810000       | No
                        Log likelihood = -97.306000
```

Figure 6. Fitted model after rescaling.

Other methods could have been used to test for non-linearity in logit, for example, fitting models with standard non-linear terms such as ln$x$ or x^2. Examination of the marginal distributions of covariates may indicate which terms are relevant (Kay and Little, 1987).

Inclusion of interaction terms

After scaling, checks are made for interactions in the model. In general, interaction terms should only be included in the final model if they are statistically significant (Hosmer and Lemeshow, 1989). Logistrule fits up to 9 models, including the main effects of all essential variables in addition to a single interaction term. This term is created by multiplying the study variable with another variable in the model, so that the factors in the interaction are also always represented by main effects. The difference in log-likelihood is determined with a *p*-value. Any interaction term with a *p*-value less than 0.15 is included in the final model. Figure 7 shows the "interaction summary slot" in our example. Three significant interactions are included in the final model. In the case of interactions involving categorical variables, the parametrization of BMDP LR is adhered to in the summary.

```
                       Overview of interactions
  Interaction    |  Log-likelihood  |      G      |  df  |   p-value
---------------------------------------------------------------------------
 Main effects    | -97.306000       |             |      |
 LWT*AGE         | -94.221000       | 6.169998    | 3    | 0.1036261
 LWT*RACE        | -94.173000       | 6.266006    | 6    | 0.3940611
 LWT*SMOKE       | -94.130000       | 6.352005    | 3    | 0.09568536
 LWT*PTL         | -96.994000       | 0.6239929   | 3    | 0.8909184
 LWT*HT          | -94.921000       | 4.770004    | 3    | 0.1894343
 LWT*UI          | -90.065000       | 14.481990   | 3    | 0.002317369
```

Figure 7. Overview of different interactions that will be included in the model if statistically significant.

Following scaling and tests for interaction, the final model is refitted, including rescaled variables and significant interactions (Figure 8).

```
                           Overview of full model
   VARIABLE       |    TYPE     | WALD STATISTIC   |   COEFFICIENT   |   EXP(COEFF)
-----------------------------------------------------------------------------------
AGE               |Adjustment   | 1.410000         | 0.1038000       | 1.110000
LWT               |Study        | (1)1.410000      | (1)3.804000     | 44.900000
                  |             | (2)1.630000      | (2)4.658000     | 105.000000
                  |             | (3)0.1500000     | (3)0.3985000    | 1.490000
RACE              |Ordinary     | (1)2.100000      | (1)1.220000     | 3.390000
                  |             | (2)1.550000      | (2)0.7984000    | 2.220000
SMOKE             |Ordinary     | -0.5820000       | -0.4538000      | 0.6350000
PTL               |Ordinary     | (1)2.960000      | (1)1.652000     | 5.220000
HT                |Ordinary     | 2.090000         | 1.531000        | 4.620000
UI                |Ordinary     | -0.2630000       | -0.2160000      | 0.8060000
LWT*AGE           |             | (1)-0.2818000    | (1)-2.170000    |
                  |             | (2)-0.2966000    | (2)-2.330000    |
                  |             | (3)-0.1252000    | (3)-1.150000    |
LWT*SMOKE         |             | (1)2.146000      | (1)2.030000     |
                  |             | (2)0.6687000     | (2)0.5780000    |
                  |             | (3)2.026000      | (3)1.810000     |
LWT*UI            |             | (1)1.165000      | (1)0.8920000    |
                  |             | (2)12.400000     | (2)0.3460000    |
                  |             | (3)0.2437000     | (3)0.1490000    |
Constant          |             | -1.610000        | -2.792000       | 0.06130000
                   Log-likelihood for full model = -83.858000
```

Figure 8. The final model fitted by Express in the low birth weight example.

2.4 Goodness of fit

Before the final model can be interpreted, its overall adequacy is assessed. The goodness of fit is evaluated by overall measures incorporating the differences between observed values $\boldsymbol{y}$ and fitted values $\hat{\boldsymbol{y}}$. The contributions of all separate pairs $(y_i, \hat{y}_i)$ are also examined.

Summary statistics

For assessment of overall fit, Logistrule presents the *p*-value for the Pearson chi-square statistic

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2 ,$$

where the summation extends over all covariate patterns in the data set. The term $r(y_j, \hat{\pi}_j)$ is the Pearson residual, measuring the difference between observed and fitted values:

$$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}$$

Here m_j denotes the number of subjects with covariate pattern $\boldsymbol{x}_j$ and y_j is the number of positive responses among the m_j subjects. A similar statistic can be based on deviance residuals. The goodness of fit chi-square can give misleading results when cell frequencies are small. In that case, the number of levels is typically large and the assumptions for the Pearson chi-square test may not be satisfied. The Hosmer and Lemeshow test (Hosmer and

Lemeshow, 1980; Lemeshow and Hosmer, 1983) accounts for this problem by regrouping the data into a 2×10 table and then executing the Pearson chi-square test. The cells in the 2×10 table are defined by the estimated probabilities $\hat{\pi}(\boldsymbol{x})$. If the number of distinct covariate patterns exceeds 10, Logistrule relies on the Hosmer and Lemeshow test to decide on goodness of fit, otherwise the Pearson test is applied.

A *p*-value for goodness of fit below 0.10 is regarded as unsatisfactory, and the user is asked whether the multivariate analysis should be restarted. A poor fit may be explained in different ways. The logistic model may provide an inadequate approximation to the true relation between $E(Y \mid \boldsymbol{x}_j)$ and $\boldsymbol{x}_j$, or an essential variable may be incorrectly scaled or missing altogether from the model. If the multivariate part of the analysis is restarted in such situations, the limits for entering variables into the model are less restrictive. The Wald statistic is still used to decide about significance. If this statistic had a value in the interval [-1.64, 1.64] during the first cycle of the multivariate analysis, the variable was regarded as making a non-significant contribution. Following a restart, the corresponding interval is [-1,64 + 0.5, 1.64 - 0.5], with a formal significance level of about 0.25. Thus, variables of marginal significance are more easily included. When the second multivariate analysis has been completed, the goodness of fit statistic may still indicate a poor fit. The user can restart the process again, with even lower interval limits for inclusion. This loop can be carried out at most three times.

Influential observations

In addition to the overall fit, the user should also be concerned about the influence of each covariate pattern on the fitted model. The three statistics used for this purpose in Logistrule are $\Delta\hat{\beta}$, ΔX^2 and ΔD (Pregibon, 1981). These are measurements of change when a particular covariate pattern is omitted, considering the estimated coefficients ($\Delta\hat{\beta}$), the Pearson goodness of fit statistic (ΔX^2) and the goodness of fit statistic based on deviance residuals (ΔD), respectively. The quantity $\Delta\hat{\beta}$ is obtained as the standardized difference between the maximum likelihood estimates $\hat{\beta}$ and $\hat{\beta}_{(-j)}$, computed using all covariate patterns and excluding the m_j subjects with pattern $\boldsymbol{x}_j$, respectively (j=1,2,...,J). Thus

$$\Delta\hat{\beta}_j = \left[\hat{\beta} - \hat{\beta}_{(-j)}\right]^T \left[X^T V X\right] \left[\hat{\beta} - \hat{\beta}_{(-j)}\right]$$

where **V** is a $J \times J$ diagonal matrix with elements

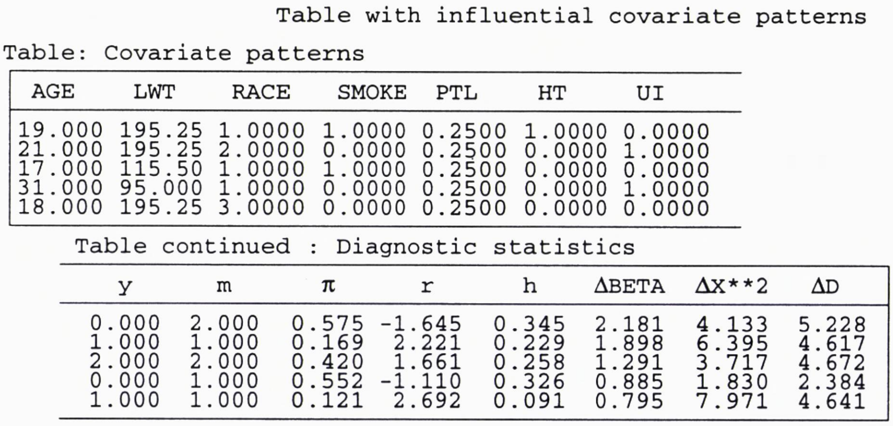
$$v_j = m_j \hat{\pi}(x_j)\,[1 - \hat{\pi}(x_j)]$$

and **X** is the design matrix. The two other statistics measure the decrease in the value of the Pearson chi-square statistic and the deviance, respectively. Through these statistics, covariate patterns may be identified which fit poorly to the adapted model (resulting in large values of ΔX^2

or ΔD), and also patterns with an excessively large influence on the estimated coefficients (reflected by large values of $\Delta\beta$).

It was emphasized by Hosmer and Lemeshow (1989) that any diagnostic should be visual as well as numerical. The visual part of the analysis in Logistrule includes several plots of the relevant statistics. In plots of ΔX^2 or ΔD against the observed probability, we can visually identify patterns with a poor fit. However, this identification of outliers is not as simple as in linear regression. As pointed out by Jennings (1986), a possible outlier in logistic regression is typically the observed success with the lowest estimated success probability, or the failure with the highest estimated success probability. Using a prior definition of outliers, Jennings (1986) found that removal of outliers and refitting produced biased estimates.

Hosmer and Lemeshow (1989) distinguished between two types of covariate patterns. Observations with very low or very high estimated probabilities should not be removed from the model, as the observed outcomes simply represent unusual events. An exclusion would not change estimated coefficients substantially or alter the conclusions. In contrast, patterns with a large influence on the estimated coefficients should be carefully examined. If the objective is risk factor analysis, these patterns should be excluded from the model only if they are of little biological interest.



```
                    Table with influential covariate patterns
   Table: Covariate patterns
   AGE     LWT     RACE    SMOKE   PTL     HT      UI
   19.000 195.25 1.0000 1.0000 0.2500 1.0000 0.0000
   21.000 195.25 2.0000 0.0000 0.2500 0.0000 1.0000
   17.000 115.50 1.0000 1.0000 0.2500 0.0000 0.0000
   31.000 95.000 1.0000 0.0000 0.2500 0.0000 1.0000
   18.000 195.25 3.0000 0.0000 0.2500 0.0000 0.0000
        Table continued : Diagnostic statistics
          y      m      π      r      h     ΔBETA  ΔX**2   ΔD
        0.000  2.000  0.575 -1.645  0.345  2.181  4.133  5.228
        1.000  1.000  0.169  2.221  0.229  1.898  6.395  4.617
        2.000  2.000  0.420  1.661  0.258  1.291  3.717  4.672
        0.000  1.000  0.552 -1.110  0.326  0.885  1.830  2.384
        1.000  1.000  0.121  2.692  0.091  0.795  7.971  4.641
```

Figure 9. Table produced by LOGISTRULE with covariate patterns that have a large influence on the estimated coefficients.

Figure 9 shows how the most influential covariate patterns with diagnostics are presented by Logistrule in the low birth weight example. Here y denotes the number of successes, m is the number of observations, π is the estimated success probability, r is the Pearson residual and h is the leverage value (the diagonal element of the hat matrix). Up to 5 influential patterns are presented, selected in this way: First, patterns for which $\Delta\beta < 0.8$, $\Delta X^2 < 4$ and $\Delta D < 4$ are not regarded as influential. The remaining patterns are sorted, considering the value of $\Delta\beta$, and the 5 patterns with the highest values are selected (if they exist).

After the presentation of influential covariate patterns, the user can decide whether they should be removed. The multivariate analysis is then repeated with the modified data set, written to a new file. Removal of covariate patterns may change the number of variable levels, and all such quantities must be recalculated before the multivariate analysis is restarted. The model building strategy applied during the new cycle coincides with the one already described, except for grouping of categories. Variables for which categories were already combined, will not be considered as candidates for regrouping.

2.5 Interpretation

The final step is the interpretation of results, as shown in Figure 10 for the low birth weight example. A table is printed with odds ratios for variables not included in any interactions. If there are no significant interactions, this is the only table presented. If an independent variable is included in the final model as categorical, separate odds ratios are given for each category relative to the lower reference category.

Logistrule then displays odds ratios for the study variable. If any interactions are included in the final model, these values cannot be calculated simply by exponentiation of the the study variable coefficient, as the effect depends on values of other variables included in the interactions. To compute odds ratios, we consider differences in logit inserting particular values of these variables. For two different levels of the study variable, say x_1 and x_2, this difference is:

$$\begin{aligned}\Delta\text{logit} &= \beta_0 + \beta_1 x_1 + \beta_2 v_1 + \beta_3 v_2 + \ldots + \beta_{10} v_9 + \beta_{11} x_1 w_1 + \ldots + \beta_{14} x_1 w_4 \\ &- (\beta_0 + \beta_1 x_2 + \beta_2 v_1 + \beta_3 v_2 + \ldots + \beta_{10} v_9 + \beta_{11} x_2 w_1 + \ldots + \beta_{14} x_2 w_4) \\ &= \beta_1 (x_1 - x_2) + \beta_{11}(x_1 - x_2) w_1 + \beta_{12}(x_1 - x_2) w_2 + \beta_{13}(x_1 - x_2) w_3 + \beta_{14}(x_1 - x_2) w_4\end{aligned}$$

(with suitably modified terms for categorical variables).

There may be several odds ratios of interest. Logistrule computes tables of odds ratios corresponding to various combinations of fixed values of the independent variables included in any interaction. In each case, the odds ratios is found comparing different levels of the study variable (corresponding to x_1 and x_2 in the equation above). In each table, one particular independent variable is allowed to go through a set of relevant values. If this is a categorical variable, Logistrule will compute the odds ratio of the study variable for each value. If the variable is interval scaled, the minimum, the mean and the maximum values are used when calculating odds ratios. Other independent variables included in interactions are kept constant during these calculations. The values selected for such variables again depend on the scale. For categorical variables, the group with highest number of observations is used, otherwise the minimum value is inserted.

In our example Logistrule produces three tables presenting odds ratios for the different groups of LWT (our study variable), using the lowest group as reference. The three tables show results for different fixed levels of the independent variables AGE, SMOKE and UI.

```
                    Overview of the interpretation
Odds ratio for variables not included in interactions:
      VARIABLE          |          ODDS RATIO
-----------------------------------------------------
RACE    (Group 1)       | 1.0000
        (Group 2)       | 3.390000
        (Group 3)       | 2.220000
PTL     (Group 1)       | 1.0000
        (Group 2)       | 5.220000
HT                      | 4.620000

Odds ratio for the study variable:

            AGE
         ------------------------------
  LWT   |  MIN      AVR      MAX      |
-------------------------------------
Gr1-Gr2 |   .868     .065    .1E-03   |
Gr1-Gr3 |  1.658     .108    .2E-03   |
Gr1-Gr4 |   .258     .081      .005   |
Fixing remaining variable(s) at these values:
SMOKE =      .000
UI    =      .000
            SMOKE
         ------------------------------
  LWT   |  MIN      AVR      MAX      |
-------------------------------------
Gr1-Gr2 |   .065     .150      .554   |
Gr1-Gr3 |   .108     .140      .211   |
Gr1-Gr4 |   .081     .180      .618   |
Fixing remaining variable(s) at these values:
AGE   =   23.212
UI    =      .000
            UI
         ------------------------------
  LWT   |  MIN      AVR      MAX      |
-------------------------------------
Gr1-Gr2 |   .065     .076      .208   |
Gr1-Gr3 |   .108     .635    .3E+05   |
Gr1-Gr4 |   .081     .084      .104   |
Fixing remaining variable(s) at these values:
AGE   =   23.212
SMOKE =      .000
```

Figure 10 Odds ratio for variables in the final model.

3. Experience gained by constructing Logistrule

For a complex system such as Logistrule, it is important to investigate the overall behaviour of the rules when the implementation has been completed. In a more general context, Logistrule can also be used to assess the potential of Express as a shell for developing knowledge-based systems in statistics. Express does not rely on more advanced tools commonly applied in the area of artificial intelligence, such as Prolog, Lisp or object oriented programming, but is implemented in Fortran and Assembler running under DOS 3.0. The choice of implementation is partly historical and partly reflects deliberate decisions. In any case, this implementation has made it possible to evaluate the usefulness and flexibility offered by an ordinary third generation language for such purposes.

For each domain, the system includes a knowledge base comprising a code file and a set of rules. Specifications in the code file define slots, representing statistical quantities or other information. The data base is used as a storage of the values assigned to the slots. If the value of a particular slot is referred to but has not yet been determined, the system retrieves a rule number from the code file, indicating which rule should be executed to determine the slot value. Often an external statistical package must be executed to obtain the information needed. Executing packages and handling statistical quantities involves a considerable amount of technical processing. This is why we have found it particularly useful to apply Fortran in the specification of rules.

In his book "The integration of expert system into mainstream software", Gillies (1991) discussed how the best qualities of more traditional software systems could be combined with the flexibility of intelligent systems. He also considered the general question of what defines an expert system. He gave the following answer:

> *A system is not an expert system because it enshrines a particular language, data structure or paradigm. It is an expert system because it provides expert level performance through whatever means it can.*

Thus a system such as Express may certainly be able to compete with systems designed using more advanced tools, provided that it can embody the statistical knowledge needed to solve the practical problems encountered. In the following sections, we consider the performance of Logistrule in view of external standards.

3.1 Comparison with Hosmer and Lemeshow's procedure

The performance of Logistrule is first tested by comparing the results found in the low birth weight example with those presented by Hosmer and Lemeshow (1989). For each separate independent variable, the results of the univariate analysis are identical for the two approaches. The minor discrepancies seen for the likelihood, other statistics and *p*-values are most likely caused by use of different software, or in some cases by rounding errors in the data. Logistrule decides that all independent variables should be considered as candidates in the multivariate model, while Hosmer and Lemeshow exclude the variable FTV. This difference is caused by the special handling in Logistrule of independent variables which do not satisfy the criteria concerning linearity in logit. Rather than eliminating such a variable, Logistrule performs a grouping and treats the variable as categorical in the multivariate fit. As can be seen in Figure 4, FTV is not significant in the multivariate model. Thus FTV is also removed from the model by Logistrule, although this occurs at a later stage. In any case, there is no difference caused by limitations in Express. The multivariate analysis leads to selection of the same variables and the models fitted are almost identical.

In the selection of a suitable scaling, Logistrule pays special attention to the variables AGE, LWT and PTL, as do Hosmer and Lemeshow. Their solution is to keep AGE as continuous while LWT and PTL are grouped. Both LWT and PTL are transformed into dichotomous variables. For LWT this happens because an analysis of the grouped variable, based on quartiles as cut-points, shows that the odds ratios are almost identical for all categories except the first. For PTL the decision to dichotomize is based on the number of observations at the different levels. As few women had previously experienced premature labour, Hosmer and Lemeshow found it sensible not to distinguish between the number of times it had occurred. The variable LTW is found by Logistrule to depart from linearity in logit and is therefore grouped, considering quartiles. It is treated as categorical in the remainder of the analysis. No further action is taken in Logistrule after grouping of a continuous variable, mainly because Hosmer and Lemeshow (1989) do not clearly describe their general approach. For the variable PTL, Logistrule applies the grouping algorithm described in Section 2.1, producing the same categories as considered by Hosmer and Lemeshow.

As interactions are selected in Logistrule using the E,V,W model of Kleinbaum (1994), the procedure differs from that of Hosmer and Lemeshow (1989). Whereas Hosmer and Lemeshow detected only 2 significant interactions in the example, Logistrule found 3, involving an additional term between LWT and UI. As shown in Figure 7, this interaction is clearly statistically significant. The *p*-value quoted by Hosmer and Lemeshow for including the two interactions is 0.06, while a corresponding computation for the models fitted by Logistrule results in a *p*-value equal to 0.0014. Hosmer and Lemeshow (1989) do not clearly define a default strategy but attempt to include those interactions which would be suspected on prior grounds to be significant. This approach could of course have been adopted in Logistrule, expecting the user to list all potentially interesting interactions. However, for a novice, this approach is difficult.

In the final model fitted, there are two essential differences between the results of Logistrule and those of Hosmer and Lemeshow. A different scaling is used for LWT, and Logistrule includes the additional interaction between LWT and UI. As a consequence, there are also minor differences between the remaining coefficients in the two models, in particular for terms involving LWT and UI.

In the assessment of overall fit, neither Logistrule nor Hosmer and Lemeshow's procedure detects any evidence of a poor fit. As the final models do not coincide, the influential covariate patterns also differ. As recommended by Hosmer and Lemeshow, Logistrule checks the logistic regression model by computing suitable statistics and sets up plots for visual assessment. In Logistrule, the selection of covariate patterns which may be excluded is mainly based on the value of $\Delta\beta$. Thus, although Hosmer and Lemeshow stress that the assessment should rely on visual impressions, the action taken is almost identical to that chosen by Logistrule. The user may decide whether influential observations should be removed and the analysis restarted. In our example, we decided not to discard any observations.

The last and perhaps most difficult part is the interpretation of the fitted model. In Logistrule, we again rely on the E,V,W model, with a study variable of particular interest. Logistrule concentrates on computing odds ratio for this variable. If any interaction is present, the effect of the study variable is investigated through calculation of changes in logit (Section 2.5). In this calculation, other variables included in the interaction are fixed. Logistrule computes the odds ratios for a selection of fixed values. In the low birth weight example, the odds ratios determined by Logistrule, for independent variables not included in any interactions, are of the same magnitude as those given by Hosmer and Lemeshow (1989). As the variable LWT is scaled differently and the number of interaction terms is not the same, results produced by Logistrule cannot be compared directly to those of Hosmer and Lemeshow, but the main conclusions are similar.

3.2 Assessment of statistical software on the basis of principles for expert systems

Most people working in the area of artificial intelligence will agree that it is extremely difficult to construct a successful expert system. A basic problem common to all such systems is the formalization of knowledge. According to Streitberg (1988), it is impossible to build systems for statistical analysis with an expert performance, simply because the experience of a statistician cannot be made completely available. In a response to this claim, Molenaar (1988) pointed out that a considerable amount had been learnt already about statistical consultation and knowledge representation during the first years of research in this field. The research should therefore go on, despite some disappointments in connection with early statistical expert systems. Perhaps initial expectations were too high. Some statisticians may even have regarded the new systems as competitors rather than useful software tools. In the following, we consider the prospects of solving problems in statistical data analysis by an expert system.

Gillies (1991) refers to four characteristics of general problems which can be solved efficiently by expert systems:

(a) No deterministic solution.
(b) Clearly defined limits.
(c) Machine representable knowledge.
(d) Available expertise.

As a background for assessing the performance of Express, we will discuss whether these conditions are typically satisfied by statistical problems. On the basis of our experience with Logistrule, we will also investigate whether reasonable solutions can be found using Express.

Different parts of a statistical analysis

Some textbooks give the impression that statistics is simply a diverse collection of techniques, and that statisticians have a duty is to select the correct technique and apply this to the data. Hand (1987) called this impression the "cookbook effect", and he listed the responsibilities of the statisticians to show that there are no cookbook solutions to real statistical problems. First, the statistician should interact with the client to refine the research objectives. Second, relevant techniques must be identified, followed, third, by an analysis of ıhe data and, fourth, by a discussion of the conclusions with the researcher. Hand's description of the separate steps shows that there are no purely deterministic solutions to ordinary practical statistical problems.

In the design of Logistrule we have concentrated on the third and fourth responsibilities, as our primary objective was to investigate whether rules formulated in Express were adequate for representing statistical knowledge. In using the book by Hosmer and Lemeshow (1989) as a basis for the knowledge, we already restricted ourselves to logistic regression. During the development of Express, a considerable amount of time was spent constructing a satisfactory interface between statistical packages and the sets of rules. Thus Express is particularly useful for data analysis. In fact, none of the sets of rules developed so far relate to the planning stage of an experiment. Other kinds of tools are more likely to be needed for problems of such a different nature.

Several knowledge-based systems have been developed for design of experiments. Nys *et al.* (1992) described a system that starts with an initial design and checks whether effects are testable and if the design is powerful. The researcher may change the design and watch the implications. DEXPERT (Lorenzen *et al.*, 1992) designs experiments and also performs an analysis after data collection, using appropriate SAS procedures. SPRINGEX and STATISTICAL NAVIGATOR (Raes, 1992) are commercially available expert systems which elicit information and guide the user to an appropriate technique. Of these systems, only DEXPERT can be applied to design studies as well as conduct the analyses.

In his review of statistical expert systems, Hand (1985) formulated important questions which should be considered before any particular system is designed. It is of course essential to ask oneself "for what will the system be used?". Hand indicated two alternative kinds of response to this question:

(i) For the design of studies and selection of statistical strategies.
(ii) For conducting the analysis.

Hand pointed out that the response has major implications for the optimal implementation of the system. We found Fortran to be a suitable language for implementation of rules when the analysis requires several executions of statistical packages. Widely different methods have been used for both types of systems. Completely different approaches may even be used to solve the same problem, as in the case of SPRINGEX and STATISTICAL NAVIGATOR (Raes, 1992). The fact that all these systems act in different ways once more indicates that no unique deterministic solution exists to ordinary statistical problems.

Limitations

The fact that statistical expert systems are designed according to different principles and act differently, reflects the basic truth that statisticians may have different ideas about the correct handling of a problem. In a study of strategies proposed by different statisticians for solving prescribed practical problems, Tung and Schuenmeyer (1991) found some common fundamental strategies, although many strategies were often unique to only one or a few statisticians. For example, statisticians engaged in economy frequently refer to other concepts than statisticians working in social sciences. This indicates that there may be no clearly defined limits when a general statistical problem is addressed. Van den Berg and Visser (1990) suggested that it may be preferable to focus on a single statistical discipline in the design of support systems. This also makes it more likely that Gillies' second condition will be met. Constructing a successful expert system will be much easier if one aims at a particular user group, not taking all potential users into account.

The only group of statistical techniques offered by Logistrule is logistic regression. The system does not provide any form of statistical consulting, which means that a user must know in advance that logistic regression is called for. It is important that decisions concerning a target group are made as early as the design stage. In Express, interfaces to users and to statistical packages are designed in a way which hopefully makes the system independent of the user group. Thus Express can serve an expert as well as a statistical novice. The system FOCUS (Prat *et al.*, 1992) was designed along similar principles, with tools provided for writing knowledge-based front end modules. The design of these modules determines the appearance of the final system.

Users belonging to the same target group may still possess different background knowledge about the methods applied. For this reason, user interfaces must be flexible with regard to speed and the amount of help provided. The limits to the knowledge incorporated in an expert system should be well defined, although the environment in which the system is constructed should impose as few limitations as possible. Large-scale systems dealing with huge areas of application can of course be useful, but the limits to what the system can handle must be set in advance. Furthermore, these limits must be presented to the user, with a warning if the system does not cover the situation the user has in mind. Following their attempt to construct an expert system for binomial experiments, Grüger and Ostermann (1986) concluded that well defined statistical problems are usually easy to handle, while the construction of large-scale expert systems was almost impossible. In some situations, incremental solutions may be advantageous, with a small system constructed first and major extensions added later. This approach is possible in sets of rules defined in Express.

Knowledge representation

The main purpose of our work with Logistrule was to test a representation in Express of the knowledge contained in the strategy for model building described by Hosmer and Lemeshow (1989). Considering the performance of Logistrule in the low birth weight example, the representation seems adequate, but a more detailed assessment is also required. The references to external software form an essential part of the knowledge base. Two distinct modules handle the interface to external programs. The commands appropriate to a specific package and the general statistical problem must be selected first. At the initial stage, these commands are often incomplete. Typically, information about variable and file names must be filled in. This is done automatically by the system. Second, the output file produced by the statistical package is scanned and essential information is extracted and written to the working memory, which can be accessed from the rules. The interface modules rely on a substantial amount of technical specifications to function properly. Most specifications must be stored in the code file before the analysis starts, but additional information is generated during execution. For example, the number of variables considered in a regression model may vary from one system execution to the next.

A typical If-Then-Else clause in Express has the following format:

```
Get value of slot A
If value of A not found then
   Make necessary preparations
   Find A
Else
   If A equals B then .....
End
```

Each rule number in Express refers to a separate Fortran subroutine, which may include one or several If-Then-Else clauses. Any other Fortran statements needed can of course be introduced. The Express rules do not correspond to standard rules used in ordinary production systems, but statistical strategies rarely fit into a representation by production rules (Aarseth & Heuch, 1996b). The slot value **A** referred to in the rule may be found in several different ways. One possibility is to use Fortran statements for a direct calculation inside the rule. Another possibility is to ask the user to decide its value. Very often, another rule must be activated to determine the slot value. To find the correct rule, the slot **A** must be connected to a rule number in the code file. The first time the system tries to read the value of **A** in the data base, this number is supplied instead, indicating which rule should be executed next. The rule number connected to **A** can in turn lead to the execution of another rule, but frequently the value of **A** can only be found by execution of an external package. The rule number read from the code file will then indicate where the information is stored which is used to locate and execute the correct package.

The data analytic situations handled by Express usually require several executions of external statistical packages. We have found it convenient to combine the use of rules written in Fortran with coding performed in a couple of system files. This facilitates easy communication between the rules and the packages. The two modules which form the interface to external software are not completely separated from the rules. This overlap makes it easier to carry out preparations needed before a package can be executed. In our experience, when commands are written for a particular package, a considerable amount of information about quantities handled in the rules must be passed on to the interface modules. This is achieved mainly through calls to utility Fortran subroutines in the Express library which write essential information to the working memory. When the module for execution of the package takes control, it reads predefined commands one-by-one from the code file and substitutes information from the data base in incomplete commands.

When results are extracted from the output produced by an external package, the situation is somewhat different. Depending on several characteristics of the particular analysis, the number of results extracted may vary from execution to execution. For example, the number of variables used and the number of levels for a particular variable may differ between two package executions that are identical in all other regards. Thus it is necessary for the extraction module to obtain information from the working memory that can be used to search for the correct information in the output file produced by the package. As in the process of completing commands, the system starts by considering predefined codes in the code file and then uses the data base to complete the information needed to perform the correct extractions. The values extracted are written to the data base for any later use by the rules. Thus, if the value of slot **A** has been found, this value is stored in the data base, and then the rule which tried to get the value of **A** in the first place can be restarted, and the chaining continues.

We have found that Fortran code, in combination with a dynamic data base and a static knowledge base (the code file), provides a suitable framework when the computations are based on existing software. It is especially important to have a convenient interface to such software when the rules incorporate complex procedures which are only available in major external packages. In Logistrule, external statistical software must be started many times during an analysis. If the results produced by such packages were not accessible, it would not be feasible to represent the statistical knowledge based on the particular numerical procedures. The Fortran subroutines evidently offer a flexible way of representing knowledge.

There are obvious advantages as well as disadvantages to this approach. Rules in a system without any interface to statistical packages may be simpler in some respects. When Irgens (1991) began the implementation of her set of rules, it soon became clear how difficult it was to give precise general instructions concerning the implementation of a knowledge base. The implementation process must deal with two different problems. First, a considerable amount of information must be specified by the knowledge engineer in the code file, and the data base must be initialized. This is done using the editor in Express, following well defined rules. The second, more ill-defined problem is the implementation of the Fortran code. It is hard to formulate guidelines about the writing of rules without losing the general flexibility inherent in Fortran. Construction of rules may appear to be very time-consuming, with almost too much freedom in the syntax, but there are also several advantages. First, being familiar with the language of the system undoubtedly makes it simpler to understand the syntax. Fortran is a wide-spread well known language. Many users will be able take advantage of their previous experience to utilize the freedom offered by Express at the rule specification stage. This freedom will of course make it possible to write the same rule in many different ways.

In an attempt to avoid too complicated rules, we have constructed a library of Fortran utility subroutines which has been particularly useful. Certain routines are essential in nearly all rules, while others will be used more rarely. The main purpose is communication, with the user or more often the knowledge base (the code file) and the data base. The freedom in the rule specification part makes it possible to represent knowledge in different ways. In our experience with the sets of rules available in Express, the construction of a statistical strategy map can form a convenient first step in the representation of knowledge.

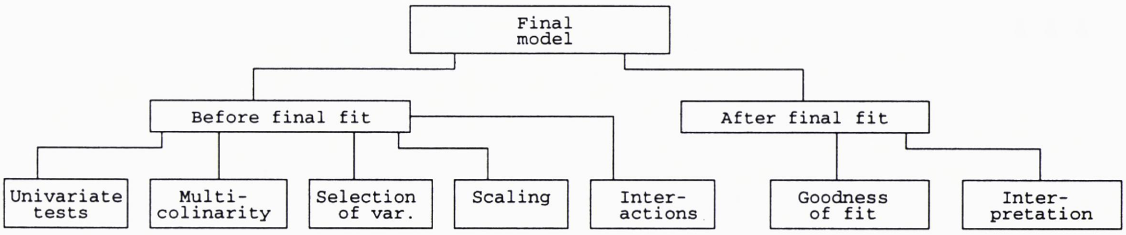


Figure 11 Simple strategy map for LOGISTRULE.

Figure 11 shows a crude strategy map for Logistrule. This initial strategy defines the main tasks that must be handled by the final system. Links must be established to the existing nodes (boxes in Figure 11) for further refinement. Again the flexibility of Express makes it possible to introduce such a refinement in different ways, but when a new node is added, we must take into account how it can be implemented in Express. Using this approach for developing a statistical strategy, we can start writing the first rules at an early stage. We will need a main rule for activating the different nodes which may be implemented before any decisions have been made on further extensions. It might seem reasonable that as much as possible of the strategy should be developed before any implementation starts. The experience gained during the implementation of Logistrule has shown that this is not always necessary, and a more convenient approach may be to finish one step before proceeding to the next. For example, the univariate part of Logistrule was implemented and tested before the strategy in the multivariate case was completed.

In this way it is also possible to make extensions to already existing sets of rules. Thus we can start implementing a set of rules within rather narrow, well defined limits. If we decide that the limits should be extended, it will be easy to obtain the new strategy by adding to the already existing one. This feature is also useful if we are going to adapt a set of rules to particular user groups. Constructing a common base system which is easy to refine, will make it possible to accomodate many user categories with little effort.

Statistical strategies

The next basic question is whether it is possible to construct satisfactory strategy maps for ordinary statistical problems. Despite the different preferences among statisticians, we can try construct strategies which may be of use in many situations. We will simply have to find methods which are generally acceptable and still are adapted to the situations handled. The completed system will only represent a particular point of view and should not be regarded as a statistical oracle which can always give absolutely correct answers. It is necessary to implement the strategy to be able to test its performance properly. Such testing can be performed in different ways, including simulation or evaluation by running through well known examples (Aarseth & Heuch, 1996c).

Certain parts of the strategy presented by Hosmer and Lemeshow (1989) were not described in sufficient detail to form a basis for our construction of rules in Logistrule. This was the case, for instance, for the grouping algorithm for an interval scaled independent variable which was non-linear in logit. The book by Hosmer and Lemeshow is no doubt well written, easy to understand and it covers the modelling process in an instructive manner. In a review, Scott (1991) stated that "*the great strength of the book is the careful, detailed description it gives of the whole modelling process*". The same impression was given by other reviews (Kupper, 1990; Iyer, 1991). This shows that even if a modelling process appears to be very well described, it needs careful examination before it can be implemented in a knowledge-based system.

Oldford and Peters (1986) emphasized that a statistical strategy should be well targeted, in the sense that the implementation should behave in the same way as the theoretical strategy proposed by the statisticians. Furthermore, a strategy is said to be complete if it is well targeted for all practical problems to which the strategy is supposed to be applicable. The completeness property is not often satisfied by presentations in textbooks because it would entail extensive descriptions of minor details in the analysis. The authors usually prefer giving references to other textbooks or papers. In implementing the strategy described by Hosmer and Lemeshow (1989), we thus found it essential to introduce certain extensions. For some problems, however, we are aware that we have not constructed complete rules. Logistrule does not, for example, include any mechanism for detecting and handling multicolinearity beyond examination of simple correlations. Problems involving missing data or few observations also provide examples of the incompleteness of the system. Thus although the strategy as implemented in Logistrule performs adequately in the case of the low birth data, the performance may be poor in other situations. However, following a relatively simple extension of the system, tests for completeness could be carried out.

If knowledge is implemented on the basis of a statistical strategy map, a detailed inspection of the map should reveal missing attributes. Thus the advantage of the strategy map not only lies in the simple implementation in Express but also in the potential for detecting incompleteness. Furthermore, presenting the user with a simple strategy map will give an impression of the limitations of the analysis. The final strategy map should be considerably more detailed than that shown in Figure 11.

As a preparation for defining a strategy for logistic regression, Irgens (1991) examined a number of papers on associations between potential risk factors and cancer incidence, published by International Journal of Cancer, in an attempt to detect patterns in the analysis. If it is regarded as necessary to extract underlying knowledge from many research papers, the construction of the statistical strategy map will require a large amount of work. Irgens (1991) found that several papers did not explain why particular statistical methods were selected. Sometimes it was even difficult to understand completely what kind of analysis was performed. However, certain frequently applied strategies could be used as a basis for a set of rules, although no commonly accepted strategy apparently existed for some parts of the analysis.

It is important in such situations to define definite limits for the system. One should not try to write rules which extend over a wide spectrum of strategies. Such rules are likely to be grossly incomplete and may easily be misused by inexperienced persons. Strategies should be selected which can serve a specific user grup in the best possible manner. If the system is intended for statisticians, it may be advisable to let the user participate actively in the selection of methods. In this case parts of the system may even remain incomplete, indicating that earlier practice does not provide any obvious solution to the problem. The system may give information about literature on the problem at hand, and the statistician can perform the remaining analysis relying on general experience. Most such problems will no doubt be solved more adequately by a statistician than by a knowledge-based system. It is therefore essential that a knowledge-based system aiming at statisticians should be integrated into their working environment.

The conditions formulated by Gillies (1991) for a successful solution by expert systems may be satisfied for many statistical problems. Data analysis needs systems which can take advantage of already existing software. Despite the amount of work required in the implementation, flexible systems with good interfaces to the user and existing packages should form suitable tools for constructing statistical knowledge-based systems. The most difficult problem is still to define the statistical strategy. Although it seems almost impossible to implement complete strategies, useful systems can be built at least when the potential user group is known in advance.

REFERENCES

Aarseth, J.H. (1996). The knowledge base of Logistrule, a system for model building in logistic regression. Technical Report, Department of Mathematics, University of Bergen.

Aarseth J.H. and Heuch I. (1996a). User's guide to Express: A tool for building knowledge-based systems for statistical data analysis. *Statistical Report* no. 27. Department of Mathematics, University of Bergen, Bergen.

Aarseth J.H. and Heuch I. (1996b). Interfaces in a knowledge-based statistical system, as exemplified by Express. *Statistical Report* no. 29. Department of Mathematics, University of Bergen, Bergen.

Aarseth J.H. and Heuch I. (1996c). Assessing uncertainty in knowledge-based systems for data analysis by simulation. *Statistical Report* no. 30. Department of Mathematics, University of Bergen, Bergen.

Bendel B. and Afifi A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association* **72**, 46-53.

Bhattacharjee G.P. (1970). The incomplete gamma integral. *Applied Statistics* **19**, 285-287.

Breslow N.E. and Day N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. International Agency for Research on Cancer, Lyon.

Brown M.B. (1990). *BMDP Statistical Software Manual*. University of California Press, Berkeley.

Carlsen F. and Heuch I. (1986). Express - An expert system utilizing standard statistical packages. In: *COMPSTAT. Proceedings in Computational Statistics*, de Antoni F, Lauro N and Rizzi A (eds.). Physica-Verlag, Heidelberg, 265-270.

Cochran W.G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics* **10**, 417-451.

Gillies A.C. (1991). *The Integration of Expert Systems into Mainstream Software*. Chapman & Hall, London.

Grüger J. and Ostermann R. (1986). Construction and integration of a statistical expert system for binomial experiments. *Statistical Software Newsletter* **3**, 124-128.

Hand D.J. (1985). Statistical expert systems: Necessary attributes. *Journal of Applied Statistics* **12**, 19-27.

Hand D.J. (1987). The application of expert systems in statistics. In: *Interactions in Artifical Intelligence and Statistical Methods*, Phelps B (ed.). Gower Technical Press, Aldershot, 3-17.

Heuch I., Aarseth J.H., Ottersen G. and Carlsen F. (1990). Adaption of Express to the IBM PC: A tool for building knowlegde-based statistical system using existing packages. In: *COMPSTAT Software Catalogue*. Dubrovnik, 13-14.

Hosmer D.W. and Lemeshow S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics* **A10**, 1043-1069.

Hosmer D.W. and Lemeshow S. (1989). *Applied Logistic Regression*. Wiley, New York.

Irgens Å. (1991). *A Knowledge-based System for Logistic Regression using Express*. Thesis for the cand. scient degree. Department of Mathematics, University of Bergen, Bergen (in Norwegian).

Iyer R. (1991). Book review of D.W. Hosmer and S. Lemeshow's "Applied logistic regression". *The Statistician* **40**, 458-458.

Jennings D.E. (1986). Outliers and residual distribution in logistic regression. *Journal of the American Statistical Association* **81**, 987-990.

Kay R. and Little S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* **74**, 495-501.

Kleinbaum D.D.G. (1994). *Logistic Regression*. Springer, New York.

Kupper L.L. (1991). Book review of D.W. Hosmer and S. Lemeshow's "Applied logistic regression". *Journal of the American Statistical Association* **85**, 901-901.

Lemeshow S. and Hosmer D.W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* **115**, 92-106.

Lorenzen T.L., Truss L.T., Spangler S.W., Corpus W.T. and Parker A.B. (1992). DEXPERT: an expert system for design of experiments. *Statistics and Computing* **2**, 47-54.

Mickey J. and Greenland S. (1989). The impact of confounder selection criteria on effect estimation. *Americal Journal of Epidemiology* **129**, 125-137.

Molenaar I.W. (1988). To exist or not to exist. A comment on statistical expert systems. *Statistical Software Newsletter* **14**, 127-130.

Nys M., Darius P. and Marasinghe M. (1992). An interactive window-based environment for experimental design. In: *COMPSTAT. Proceedings of the 10th Symposium on Computational Statistics*, vol. 2, Dodge Y. and Whittaker J. (eds.). Physica-Verlag, Heidelberg, 233-238.

Oldford R.W. and Peters S.C. (1986). Implementation and study of statistical strategy. In: *Artifical Intelligence and Statistics*, Gale W.A. (ed.). Addison-Wesley, Reading, Mass., 335-353.

Prat A., Catot J.M., Lores J., Fletcher P., Galmes J. and Sanjeevan K. (1992). A separable architecture for the construction of knowledge based front ends. *AICOM* **5**, 184-190.

Pregibon D. (1981). Logistic regression diagnostics. *Annals of Statistics* **9**, 705-724.

Raes J.F.M. (1992). Inside two commercially available statistical expert systems. *Statistics and Computing* **2**, 55-62.

Scott A.J. (1991). Book review of D.W. Hosmer and S. Lemeshow's "Applied logistic regression". *Biometrics* **47**, 1632-1633.

StatXact (1989). *StatXact. Statistical Software for Exact Nonparametric Inference. User Manual.* CYTEL Software, Cambridge, Mass.

Streitberg B. (1988). On the non-existence of expert systems. Critical remarks on artificial inteligence in statistics. *Statistical Software Newsletter* **14**, 55-62.

Tung S.T.Y. and Schuenmeyer J.H. (1991). An expert system for statistical consulting. *Journal of Applied Statistics* **18**, 35-47.

Van den Berg G.M. and Visser R.A. (1990). Knowledge modelling for statistical consultation systems. Two empirical studies. In: *COMPSTAT. Proceedings in Computational Statistics*, Momirovic K. and Mildner V (eds.). Physica-Verlag, Heidelberg, 75-80.

Wax Y. (1992). Collinearity diagnosis for a relative risk regression analysis: an application to assessment of diet-cancer relationship in epidemiological studies. *Statistics in Medicine* **11**, 1273-1287.

Appendix

Sample session

In this session we will use three of the independent variables of the low birth example to go through a sample session with Express and Logistrule. Of course this results in an insufficient analysis of the data set, but a considerable amount of time is needed for the complete set of variables. More detailed descriptions of other sample sessions with Express were given by Aarseth and Heuch (1996a).

1. START EXPRESS FROM THE DOS COMMAND LINE

Start Express from the DOS command line by typing

C:\EXPRESS> **EXPRESS**

To enter the menu for selection of set of rules **push any key**.

2. SELECT PROPER SET OF RULE

Use the arrow keys to **highlight** the item "5.Logistrule (Logistic regression for up to ten x-variables)" and carry out the selection by pressing **ENTER**.

3. ENTERING VARIABLES INTO THE DATA STORAGE

You have now reached the main menu of Express as displayed in Figure 2.2 in the User's Guide to Express (Aarseth and Heuch, 1996a). The first step needed is to read the data to be considered into the data storage of Express. The file C:\EXPRESS\HOSMER.DAT includes all variables in the low birth example. As described earlier, this data set includes information about 189 women and their new-born infants. The outcome variable is "low birth weight". The purpose of the study was to determine what in a woman's behaviour during pregnancy could alter the chances of delivering a baby of low birth weight. The dependent variable has been recoded into a dichotomous variable with the value 1 for low birth weight (less that 2500g) and the value 0 for normal birth weight. This variable, assigned the name LOW, occupies only the first position on the data file. The following list shows locations of the independent variables and names used in Express:

Age of the mother (AGE), 2-4
Weight in pounds at the last menstrual period (LWT), 5-8
Race (RACE), 9-10
Smoking status during pregnancy (SMOKE), 11-12
History of premature labor (PTL), 13-14
History of hypertension (HT), 15-16
Presence of uterine irritability (UI), 17-18
Number of physician visits during first trimester (FTV), 19-20

This sample session deals with the independent variables AGE, LWT and SMOKE only. In addition to the LOW variable, these must be read into the data storage separately. To start entering the variables, **highlight** the item "DATA" in the main menu and press **ENTER**. Then **highlight** "MANIPULATE DATA STORAGE" and select by pressing **ENTER**. It is now possible to list data in the storage by pressing F9 and selecting one of the variables included. (If this is the first time Express is used, the storage will be empty). Another alternative indicated is erasing the data storage (F7). We want to add new data so the key **F5** should be pressed.

We must first supply the name of the file containing the data. Thus write **c:\express\hosmer.dat**. Furthermore, as this file includes several variables, the response YES must be **highlighted** to the question "Does this file include other variables?". Again, carry out the actual selection by pressing **ENTER**. The next two fields to be filled in concern the leftmost and rightmost position occupied by the variable. The first variable (LOW) is located in position **1**, so this number is written in both fields. The questions concerning missing values should both be answered by NO. Again this is done by **highlighting** NO and then pressing **ENTER**. Express now reads the data from the external file into the data storage. Finally, a variable name must be given. The name can occupy up to 6 positions and must not include any blank spaces. Type **LOW**. The procedure described must then be repeated for the 3 independent variables. The only difference is that these occur in other positions on the file. Now try to list one of the variables by first pressing **F9**, **highlighting** one of the variables and pressing **ENTER**. To return to the main menu, press **ESC**.

4. SELECT VARIABLES TO BE ANALYSED

The next step is to select the variables to be analysed. Note that all instructions given so far deal with the data storage only. This storage is general and can be reached from any set of rules. The following specifications connect the data with the particular problem of logistic regression. Again **highlight** "DATA" and press **ENTER**. Now select the item "SELECT DATA FOR ANALYSIS" by pressing **ENTER**. It is now possible to select among the variables included in the data storage. First select the dependent variable by **highlighting** the variable LOW and pressing **ENTER**. Repeat this procedure for AGE, LWT and SMOKE. To end the selection of variables, press **ESC** and then **ENTER**. During a short interval, nothing happens on the screen while the system prepares internally for the analysis. Then **press any key** to return to the main menu.

5. SHOW AN EXPLANATION OF THIS SET OF RULES

Before the analysis is started, take a look at the explanation file for this set of rules. First select "FILES" in the main menu by **highlighting** and pressing **ENTER.** Then select the item "EXPLANATION OF RULES". Now the map shown in Figure 11 is displayed. Use arrow keys, PgUp and PgDn to move around in this explanation file. To exit, press **F10**.

6. LIST THE SLOTS FOR THIS SET OF RULES

Also take a look at the slots associated with the current set of rules. **Highlight** "SLOTS" in the main menu and press **ENTER** twice. A list of the different slots will appear on the screen. Use PgUp and PgDn to move inside in the list. If a particular slot is **highlighted** and selected by pressing **ENTER**, information about the slot value and how it was determined will appear on the screen. As the analysis has not yet been started, only a short message will appear to the effect that the slot has not yet been found. Now **press any key** (except the ESC key) to return to the list of slots. **Press F5** to list the values of all slots. Again we notice that no values have yet been found. To return to the main menu, **press ESC** twice.

7. SET SYSTEM VARIABLES

A last step before the analysis starts is to ensure that the system variables have been properly set. **Highlight** "SYSTEM" and press **ENTER.** Suppose that we want Express to give as much information as possible during the analysis. Then all system variables should be set to ON, except for the "DEBUG" variable which must be OFF. To change one of the settings, simply **highlight** the proper variable and press **ENTER**. **Press ESC** to leave this submenu.

8. EXECUTE THE ANALYSIS

We are now ready to start the analysis (chaining of rules). First **highlight** the option "RULES" and press **ENTER**. For this set of rules, two items are shown in the basic problem menu. We want a complete analysis for all variables selected. Thus **press 2** for the multivariate analysis. The analysis is started by pressing **ENTER**. Express will now use the upper window to indicate which rules are activated and which slots should be found. The lower window of the main screen is used to give additional explanations to various slots.

Follow the instructions in the bottom line during the analysis. If the system prompts for particular values, the ESC key may be used to indicate that the system should find these values without user interaction. In fact, certain slots in this set of rules should be assigned values by the user. This applies to slots indicating whether independent variables are nominal-scaled or not, and slots representing the type of an independent variable. All three independent variables can be regarded as interval scaled in the present situation. This may be indicated by first pressing **ENTER** and then writing the value **0** (for "No"). Variable types may be specified in the same way. For the variable AGE, the value **2** should be specified, for an adjustment variable. LWT should be defined as an ordinary variable by assigning the value **3**, while SMOKE should be regarded as the study variable by indicating the value **1**.

Each time the external package BMDP must be executed, it is possible to watch on the screen how commands are prepared. To continue the analysis, press **F10**. This key must also be pressed when plot has been displayed on the screen.

9. EXPLORE A SUMMARY OF THE ANALYSIS

After the conclusion has been presented, the main menu will once more be active. A summary of the analysis can be seen by displaying the log file on the screen. **Highlight** "FILES" and press **ENTER** twice. The log file is shown in the lower window of the main screen. To enlarge the window, press **F1**. The arrow keys and PgUp and PgDn can be used to move around in the file. Press **F10** to exit from the log file. Parts of the log file from this sample session are listed below. It is also possible to review the outout from the eternal packages executed during the session by selecting the option "SHOW RESULTS FROM PACKAGE". Both the log file and the file containing output from packages may be copied to external files specified by the user, by selecting the option "STORE FILE" from the submenu of "FILES".

Now take a new look at the lists of slots, by **highlighting** "SLOTS" and pressing **ENTER** twice. Also press **F5** to see the short summary of the slot values. To get a more comprehensive explanation, **highlight** the desired slot and press **ENTER**. To return to the list of slots, **press any key**. Finally, try to investigate the relationships between different slots. **Highlight** "RELATIONSHIPS", **highlight** the slot "Has dependent variable been prepared?" and press **ENTER**. The lower window of the main menu will now illustrate slot relations. To move around in the diagram, use F5 and ENTER. To exit to the main menu, press **ESC**.

10. END THE SESSION AND EXIT FROM EXPRESS

To end this sample session, **highlight** "EXIT" and press **ENTER**. If you want to leave temporarily, select "PAUSE". In this case all slot values will be stored for later use. If "QUIT" is selected, all results are erased.

The log file summarizing this sample session is listed below.

Sample session logfile

```
ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is No!
This slot is needed to determine: Summary of full model

ACTIVATING RULE
NO.:  17 (Main rule for univariate analysis ).
RULES REMAINING ON THE STACK:    16  17

ACTIVATING RULE
NO.:   6 (Rule for activating univariate analysis for a particular variable).
).
RULES REMAINING ON THE STACK:    16  17   6

Considering slot:  Summary slot for x1. The corresponding slot value has
not yet been determined.
The system will attempt to find its value.

ACTIVATING RULE
NO.:   2 (Rule for deciding the type of an x-variable ).
RULES REMAINING ON THE STACK:    16  17   6   2

ACTIVATING RULE
NO.:   1 (Creating a BMDP file with the dependent variable (coded as 0 and
1)  ).
RULES REMAINING ON THE STACK:    16  17   6   2   1

The system has reached the following conclusion
to the question 'Has dependent variable been prepared?'
The answer to this question is No!

Explanation of the slot: 'Has dependent variable been prepared?'
In this set of rules, the dependent variable must have the values:
            0 - Indicating failure.
            1 - Indicating success.
If this is not the case, the system recodes the dependent variable.
This is done by determining a cutpoint. Every value above the cut-
point will be coded as 1, the remaining as 0. When a BMDP
file has been generated, the analysis can proceed.
This slot is needed to determine: 'Is the univariate analysis complete?'

Considering slot:  Number of levels of dependent variable. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP 2D.
```

```
        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Determine number of levels for a variable'.
/INPUT FILE = 'C:\EXPRESS\SYSFIL\538.PXE'.
    VARIABLES = 1.
    FORMAT = FREE.
/VARIABLE NAMES = LOW.
/PRINT COUNT.
    LINESIZE = 80.
/END


In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          MAX                1.0000000
          MIN                0.0000000
          DISTINCT           2

ACTIVATING RULE
NO.:   1 (Creating a BMDP file with the dependent variable (coded as 0 and
1)  ).
RULES REMAINING ON THE STACK:    16  17   6   2   1

The system has reached the following conclusion
to the question 'Has dependent variable been prepared?'
The answer to this question is No!

Explanation of the slot: 'Has dependent variable been prepared?'
In this set of rules, the dependent variable must have the values:
            0 - Indicating failure.
            1 - Indicating success.
If this is not the case, the system recodes the dependent variable.
This is done by determining a cutpoint. Every value above the cut-
point will be coded as 1, the remaining as 0. When a BMDP
file has been generated, the analysis can proceed.
This slot is needed to determine: 'Is the univariate analysis complete?'

The external package has found:
Number of levels of dependent variable    = 2

Explanation of the slot: Number of levels of dependent variable
In order to decide if the dependent variable must be recoded,
we must know the number of distinct values of the variable.
If this value is greater than 2, recoding is nessesary.
This slot is needed to determine: 'Has dependent variable been prepared?'

The external package has found:
Minimum value of the dependent variable    = 0.000000

Explanation of the slot: Minimum value of the dependent variable
If the minimum value of the dependent variable is different from
zero, the variable must be recoded (unless the variable
has a single value only).
This slot is needed to determine: 'Has dependent variable been prepared?'
```

```
The external package has found:
Maximum value of the dependent variable    = 1.000000

Explanation of the slot: Maximum value of the dependent variable
If the maximum value of the dependent variable is different from
1, the variable must be recoded (unless the variable
has a single value only).
This slot is needed to determine: 'Has dependent variable been prepared?'

Considering slot:  A BMDP file containing the dependent var. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP DATA MANAGER.

        THE FOLLOWING COMMANDS WERE APPLIED:

READ SFILE = 'C:\EXPRESS\SYSFIL\538.PXE'.
     VNAMES = LOW.
     FORMAT = FREE.
     FILE = HELP.                  /
SAVE
   FILE = HELP.
   SFILE = 'C:\EXPRESS\LOGREG\DEP.FIL'.
   CODE = FILY.
   NEW.
   KEEP = LOW.                      /
FINISH                             /


In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          VARIABLES         1

ACTIVATING RULE
NO.:   1 (Creating a BMDP file with the dependent variable (coded as 0 and
1)  ).
RULES REMAINING ON THE STACK:    16  17   6   2   1

The system has reached the following conclusion
to the question 'Has dependent variable been prepared?'
The answer to this question is Yes!

Explanation of the slot: 'Has dependent variable been prepared?'
In this set of rules, the dependent variable must have the values:
            0 - Indicating failure.
            1 - Indicating success.
If this is not the case, the system recodes the dependent variable.
This is done by determining a cutpoint. Every value above the cut-
point will be coded as 1, the remaining as 0. When a BMDP
file has been generated, the analysis can proceed.
This slot is needed to determine: 'Is the univariate analysis complete?'
```

```
ACTIVATING RULE
NO.:   2 (Rule for deciding the type of an x-variable ).
RULES REMAINING ON THE STACK:    16  17   6   2
A BMDP file including the dependent variable and the independent
variable must be generated, for later use in the analysis.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP DATA MANAGER.

        THE FOLLOWING COMMANDS WERE APPLIED:

READ SFILE = 'C:\EXPRESS\LOGREG\DEP.FIL'.
     CODE = FILY.
     FILE = ENFIL.   /
READ SFILE = 'C:\EXPRESS\SYSFIL\539.PXE'.
     FORMAT = FREE.
     VNAMES = LOW, AGE.
     KEEP = AGE.
     FILE = TOFIL.   /
JOIN FILES = ENFIL, TOFIL.
     NEWFILE = NYFIL. /
SAVE FILE = NYFIL.
     SFILE = 'C:\EXPRESS\LOGREG\XEN.FIL'.
     CODE = FILAGE.
     NEW.            /
FINISH               /


In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          VARIABLES         1

ACTIVATING RULE
NO.:   2 (Rule for deciding the type of an x-variable ).
RULES REMAINING ON THE STACK:    16  17   6   2

Considering slot:  Number of levels for x1. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP 2D.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Determine number of levels for a variable'.
/INPUT FILE = 'C:\EXPRESS\SYSFIL\539.PXE'.
    VARIABLES = 2.
    FORMAT = FREE.
/VARIABLE NAMES = LOW, AGE.
          USE = AGE.
/PRINT COUNT.
    LINESIZE = 80.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          MAX               45.0000000
          MIN               14.0000000
          DISTINCT          24
          MEDIAN            23.0000000
          RANGE             31.0000000
          Q1                19.0000000
          Q3                26.0000000

ACTIVATING RULE
NO.:   2 (Rule for deciding the type of an x-variable ).
RULES REMAINING ON THE STACK:    16  17   6   2

The external package has found:
Number of levels for x1    = 24
This slot is needed to determine: 'Is x1 nominal-scaled?'

The user has reached the following conclusion
to the question 'Is x1 nominal-scaled?'
The answer to this question is No!
The system has not yet determined this slot value.

Explanation of the slot: 'Is x1 nominal-scaled?'
If no natural ordering of categories can be introduced,
the variable is nominal-scaled.

The user has decided:
Variable type of x1    = 2
The system has not yet determined this slot value.

Explanation of the slot: Variable type of x1
Different variable types are:
     1 - Study variable (exposure factor). Will always be
         included in the model.
     2 - Adjustment variable. Is considered biologically
         important. Will always be retained in the model.
     3 - Ordinary variable. Included in the model on
         statistical grounds only.

ACTIVATING RULE
NO.:   4 (Rule for activating analysis for an adjustment variable ).
RULES REMAINING ON THE STACK:    16  17   6   4

Considering slot:  Value for summary of significance for x1. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

ACTIVATING RULE
NO.:  13 (Rule for univariate analysis of a variable with many levels ).
RULES REMAINING ON THE STACK:    16  17   6   4  13

Considering slot:  Plot of observed risk for x1. The corresponding slot
value has not yet been determined.
The system will attempt to find its value.
```

```
ACTIVATING RULE
NO.:  11 (Rule for constructing plot of observed proportions ).
RULES REMAINING ON THE STACK:    16  17   6   4  13  11

Considering slot:  'Should x1 be grouped before plot?'. The corresponding
slot value has not yet been determined.
The system will attempt to find its value.

The external package has found:
Number of levels for x1     = 24
This slot is needed to determine: 'Should x1 be grouped before plot?'

The system has reached the following conclusion
to the question 'Should x1 be grouped before plot?'
The answer to this question is Yes!

Explanation of the slot: 'Should x1 be grouped before plot?'
Preparations are made for generating a plot of observed risk
for a particular variable. If the variable has more than 8
levels, a grouping of levels is performed with the quartiles
and median as cutpoints.
This slot is needed to determine: Plot of observed risk for x1
Executes univariate logistic regression, with the median and
the two quartiles as cutpoints. The output will also include
the plot of observed risks.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Plot of observed risks (with log.reg.)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\XEN.FIL'.
       CODE = FILAGE.
/TRANSFORM
        HELP_ = AGE.
        IF ( HELP_ LE 19.00000 ) THEN (NYVAR_ = 16.50000).
        IF ( HELP_ GT 19.00000 ) THEN (NYVAR_ = 21.00000).
        IF ( HELP_ GT 23.00000 ) THEN (NYVAR_ = 24.50000).
        IF ( HELP_ GT 26.00000 ) THEN (NYVAR_ = 35.50000).
        AGE = NYVAR_ .
/REGRESS DEPEND = LOW.
         CATEGORIAL = AGE.
         MODEL = AGE.
         START = OUT.
         MOVE = 1.
         CMOVE = 0.
         ENTER = 0.99.
         REMOVE = 1.00.
         DVAR = PART.
/PLOT SIZE = 60,15.
      XVAR = AGE.
      YVAR = OBSPROP.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          LIKELIHOOD       -117.336
          G LIKELIHO       -114.892
          CHI-SQUARE        4.887
          P-VALUE=          0.180
          COEFFICIEN        0.2877
          COEFFICIEN        0.4925
          COEFFICIEN       -0.5108
          ERROR             0.415
          ERROR             0.454
          ERROR             0.483
          COEF/SE           0.693
          COEF/SE           1.08
          COEF/SE          -1.06
          EXP(COEF)         1.33
          EXP(COEF)         1.64
          EXP(COEF)         0.600
          LOWER-BND         0.588
          LOWER-BND         0.668
          LOWER-BND         0.231
          UPPER-BND         3.02
          UPPER-BND         4.01
          UPPER-BND         1.56
          ENTER LIMI       A plot or table has been extracted.

ACTIVATING RULE
NO.:  11 (Rule for constructing plot of observed proportions ).
RULES REMAINING ON THE STACK:    16  17   6   4  13  11

ACTIVATING RULE
NO.:  13 (Rule for univariate analysis of a variable with many levels ).
RULES REMAINING ON THE STACK:    16  17   6   4  13

Considering slot:  p-value, test for trend, x1 continuous. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:
/PROB TITLE = 'Logistic regression with BMDP LR'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\XEN.FIL'.
       CODE = FILAGE.
/REGRESS DEPEND = LOW.
         INTERVAL = AGE.
         MODEL = AGE.
         START = OUT.
         MOVE = 1.
         CMOVE = 0.
         ENTER = 0.99.
         REMOVE = 1.00.
         DVAR = PART.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          LIKELIHOOD        -117.336
          LIKELIHOOD        -116.023
          CHI-SQUARE         2.626
          P-VALUE=           0.105
          COEFFICIEN        -0.05007
          ERROR              0.0316
          COEF/SE           -1.58
          EXP(COEF)          0.951
          LOWER-BND          0.894
          UPPER-BND          1.01
          CONSTANT           0.3587

ACTIVATING RULE
NO.:  13 (Rule for univariate analysis of a variable with many levels ).
RULES REMAINING ON THE STACK:    16  17   6   4  13

The external package has found:
p-value, test for trend, x1 continuous     = 0.1050000
This slot is needed to determine: Value for summary of significance for x1

Considering slot:  p-value, departure from linear mod., x1. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

Considering slot:  Log-likelihood with x1 (cont. grouped). The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Logistic regression with BMDP LR'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\XEN.FIL'.
       CODE = FILAGE.
/TRANSFORM
        HELP_ = AGE.
        IF ( HELP_ LE 19.00000 ) THEN (NYVAR_ = 16.50000).
        IF ( HELP_ GT 19.00000 ) THEN (NYVAR_ = 21.00000).
        IF ( HELP_ GT 23.00000 ) THEN (NYVAR_ = 24.50000).
        IF ( HELP_ GT 26.00000 ) THEN (NYVAR_ = 35.50000).
        AGE = NYVAR_ .
/REGRESS DEPEND = LOW.
         INTERVAL = AGE.
         MODEL = AGE.
         START = OUT.
         MOVE = 1.
         CMOVE = 0.
         ENTER = 0.99.
         REMOVE = 1.00.
         DVAR = PART.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          LIKELIHOOD       -117.336
          LIKELIHOOD       -116.508
          CHI-SQUARE        1.657
          P-VALUE=          0.198
          COEFFICIEN       -0.02947
          ERROR             0.0233
          COEF/SE          -1.27
          EXP(COEF)         0.971
          LOWER-BND         0.927
          UPPER-BND         1.02

ACTIVATING RULE
NO.:  13 (Rule for univariate analysis of a variable with many levels ).
RULES REMAINING ON THE STACK:    16  17   6   4  13


The external package has found:
p-value, test for trend, x1 continuous      = 0.1050000
This slot is needed to determine: Value for summary of significance for x1

Considering slot:  p-value, departure from linear mod., x1. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

The external package has found:
Log-likelihood with x1 (cont. grouped)      = -116.508000
This slot is needed to determine: Change in dev.cont-cat, x1(cont.grouped)

The external package has found:
Likelihood with x1 (category variable)      = -114.892000
This slot is needed to determine: Change in dev.cont-cat, x1(cont.grouped)

The system has found:
Change in dev.cont-cat, x1(cont.grouped)      = 3.232010
This slot is needed to determine: Change in dev.cont-cat, x1(cont.grouped)

The system has found:
Degrees of freedom, x1 cont-cat (cont.)       = 2.000000
This slot is needed to determine: p-value, departure from linear mod., x1

The system has found:
p-value, departure from linear mod., x1      = 0.1986909
This slot is needed to determine: Value for summary of significance for x1

ACTIVATING RULE
NO.:   4 (Rule for activating analysis for an adjustment variable ).
RULES REMAINING ON THE STACK:    16  17   6   4

The system has found:
Value for summary of significance for x1     = 2

Explanation of the slot: Value for summary of significance for x1
This slot may be assigned the values 1, 2 or 3, where 3
indicates that the variable in question is highly significant
```

```
while the value 1 is used for non-significant variables. In some
cases it may be natural to postpone the decision on significance
and this slot is then assigned the value 2.
This slot is needed to determine: 'Will x1 be considered in multiv. mod.?'

The system has reached the following conclusion
to the question 'Will x1 be considered in multiv. mod.?'
The answer to this question is Yes!

Explanation of the slot: 'Will x1 be considered in multiv. mod.?'
If the variable considered is either a study or an adjustment
variable, it will always be retained in the multivariate model.
An ordinary variable which is not significant in the univariate
model, will not be regarded as a candidate in the multivariate
analysis.
This slot is needed to determine: Number of var. in multivariate analysis

ACTIVATING RULE
NO.:   6 (Rule for activating univariate analysis for a particular variable).
).
RULES REMAINING ON THE STACK:    16  17   6


                    Summary slot for x1

Summary of univariate analysis

Name = AGE
Number of levels = 24
The variable is interval scaled
Adjustment variable

Table with Q1, median and Q3 as cutpoints

                |     1       2       3       4 |Total
        --------+-------------------------------+------
        0       |    36      36      22      36 |  130
                |                               |
        1       |    15      20      15       9 |   59
        --------+-------------------------------+------
        Total   |    51      56      37      45 |  189

Odds ratio         1.0000 1.3333 1.6363 0.6000
exp(coeff.)               1.3300 1.6400 0.6000
Coefficient               0.2877 0.4925 -0.510
St.error                  0.4150 0.4540 0.4830
Wald statistic            0.6930 1.0800 -1.060
95% CI, lower             0.5880 0.6680 0.2310
95% CI, upper             3.0200 4.0100 1.5600


LOGISTIC MODEL:

LOGIT = 0.3587000-0.05007000AGE
```

```
Coefficient = -0.05007000
St.Error    = 0.03160000
Wald stat.  = -1.580000
exp(coeff.) = 0.9510000
Lower       = 0.8940000
Upper       = 1.010000
p-value for test for trend = 0.1050000
p-value for departure from linear model=  0.1986909
Summary indicator of significance = 2
Should be considered in multivariate analysis
```

.

.

. Univariate analysis for LWT and SMOKE is omitted.

.

.

```
                    Summary slot for x2

Summary of univariate analysis

Name = LWT
Number of levels = 75
The variable is interval scaled
Ordinary variable

Table with Q1, median and Q3 as cutpoints

                |      1      2      3      4 |Total
        --------+-----------------------------+------
          0     |     27     33     35     35 |  130
                |                             |
          1     |     25     10     12     12 |   59
        --------+-----------------------------+------
        Total   |     52     43     47     47 |  189

Odds ratio       1.0000 0.3272 0.3702 0.3702
exp(coeff.)             0.3270 0.3700 0.3700
Coefficient             -1.117 -0.993 -0.993
St.error                0.4550 0.4350 0.4350
Wald statistic          -2.450 -2.290 -2.290
95% CI, lower           0.1330 0.1570 0.1570
95% CI, upper           0.8040 0.8730 0.8730


LOGISTIC MODEL:

LOGIT = 1.040000-0.01438000LWT

Coefficient = -0.01438000
St.Error    = 0.006210000
Wald stat.  = -2.320000
exp(coeff.) = 0.9860000
Lower       = 0.9740000
Upper       = 0.9980000
p-value for test for trend = 0.01300000
p-value for departure from linear model = 0.05334365
Summary indicator of significance = 2
Should be considered in multivariate analysis
```

```
                    Summary slot for x3

Summary of univariate analysis

Name = SMOKE
Number of levels = 2
The variable is interval scaled
Risk factor variable/study variable.

Table of observed frequencies

               |     1       2  |Total
        -------+----------------+------
         0     |    86      44  |  130
               |                |
         1     |    29      30  |   59
        -------+----------------+------
        Total  |   115      74  |  189

Odds ratio         1.0000 2.0219
exp(coeff.)               2.0200
Coefficient               0.7041
St.error                  0.3200
Wald statistic            2.2000
95% CI, lower             1.0800
95% CI, upper             3.8000

p-value for likelihood ratio test = 0.02740000
Summary indicator of significance = 3
Should be considered in multivariate analysis

ACTIVATING RULE
NO.:  17 (Main rule for univariate analysis ).
RULES REMAINING ON THE STACK:    16  17

The system has found:
Number of var. in multivariate analysis    = 3
This slot is needed to determine: 'Has variable selection been completed?'

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is No!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model
```

```
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP DATA MANAGER.

        THE FOLLOWING COMMANDS WERE APPLIED:

READ SFILE = 'C:\EXPRESS\LOGREG\XEN.FIL'.
     CODE = FILAGE. FILE = ENFIL. /
READ SFILE = 'C:\EXPRESS\LOGREG\XTO.FIL'.
     CODE = FILLWT. FILE = TOFIL. /
READ SFILE = 'C:\EXPRESS\LOGREG\XTRE.FIL'.
     CODE = FILSMOKE. FILE = TRFIL. /
JOIN FILES = ENFIL, TOFIL, TRFIL.
     KEY = LOW.
     NEWFILE = NYFIL. /
SAVE FILE = NYFIL.
     SFILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
     CODE = MULTI.
     KEEP = LOW ,AGE ,LWT ,SMOKE
          .
     NEW.  /
SAVE FILE = NYFIL.
     SFILE = 'C:\EXPRESS\LOGREG\DIVERSE.RES'.
     CODE = RES.
     KEEP = LOW ,AGE ,LWT ,SMOKE
          .
     NEW.  /
FINISH     /


In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          VARIABLES          1

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model
```

```
Considering slot:  'Are any variables strongly correlated?'. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

Considering slot:  Correlation matrix for independent var.. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP 8D.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Correlations of independent variables'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/VARIABLE USE = 2 TO  4 .
/PRINT LINESIZE = 80.
/END

In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          MATRIX           A plot or table has been extracted.

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model

Considering slot:  'Are any variables strongly correlated?'. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is No!
This slot is needed to determine: Summary of full model
```

```
ACTIVATING RULE
NO.:    7 (Main rule for variable selection in multivariate analysis ).
RULES REMAINING ON THE STACK:     16    7

Considering slot:  Table with Wald step 1. The corresponding slot value has
not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Multivariate logistic regression '.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/REGRESS DEPEND = LOW.
         INTERVAL = AGE , LWT , SMOKE.
         MODEL = AGE , LWT , SMOKE.
/END


In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          G LIKELIHO        -111.418
          AGE               -0.03743
          AGE               -1.14
          LWT               -0.01239
          LWT               -2.01
          SMOKE              0.6696
          SMOKE              2.05

ACTIVATING RULE
NO.:    7 (Main rule for variable selection in multivariate analysis ).
RULES REMAINING ON THE STACK:     16    7

Considering slot:  Table with Wald step 1. The corresponding slot value has
not yet been determined.
The system will attempt to find its value.

The system has found:
Number of variables that can be removed     = 0
This slot is needed to determine: Final table with Wald statistic and coef


                    Final table with Wald statistic and coef

  VARIABLE   |    TYPE    | WALD STATISIC    |  COEFFICIENT   |  REMOVABLE
-------------+------------+------------------+----------------+---------------
AGE          |Adjustment| -1.140000        | -0.03743000    | No
LWT          |Ordinary  | -2.010000        | -0.01239000    | No
SMOKE        |Study     | 2.050000         | 0.6696000      | No

                    Log-likelihood = -111.418000
```

```
ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is No!
This slot is needed to determine: Summary of full model

ACTIVATING RULE
NO.:  20 (Main rule for rescaling ).
RULES REMAINING ON THE STACK:    16  20

Considering slot:  'Is x1 correctly scaled?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The system has reached the following conclusion
to the question 'Should x1 be treated as categorical?'
The answer to this question is No!
This slot is needed to determine: 'Has scaling been completed?'

ACTIVATING RULE
NO.:  21 (Rule rescaling the model ).
RULES REMAINING ON THE STACK:    16  20  21

Considering slot:  'Is x1 linear in logit?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The external package has found:
Number of levels for x1    = 24
This slot is needed to determine: 'Is x1 linear in logit?'
```

```
ACTIVATING RULE
NO.:  22 (Rule checking for linearity in logit ).
RULES REMAINING ON THE STACK:    16  20  21  22

The external package has found:
Minimum value of x1    = 14.000000
This slot is needed to determine: Wald statistic for x1*ln(x1)

Considering slot:  Wald statistic for x1*ln(x1). The corresponding slot
value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Multivariate logistic regression with x*ln(x)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM IF ( AGE GT 0 ) THEN XLNVAR_ = AGE*LN(AGE).
           IF ( AGE EQ 0 ) THEN XLNVAR_ = 0.
/REGRESS DEPEND = LOW.  ITER = 40.
         INTERVAL= AGE,LWT,SMOKE
, XLNVAR_.
         MODEL = AGE,LWT,SMOKE
, XLNVAR_.
/END
/PROB TITLE = 'Multivariate logistic regression (continuous)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM HELP_ = AGE.
           IF ( HELP_ LE 19.00000 ) THEN (NYVAR_ = 16.50000).
           IF ( HELP_ GT 19.00000 ) THEN (NYVAR_ = 21.00000).
           IF ( HELP_ GT 23.00000 ) THEN (NYVAR_ = 24.50000).
           IF ( HELP_ GT 26.00000 ) THEN (NYVAR_ = 35.50000).
           AGE = NYVAR_ .
/REGRESS DEPEND = LOW.
         INTERVAL = AGE,LWT,SMOKE
         .
         MODEL = AGE,LWT,SMOKE
         .
/END
/PROB TITLE = 'Multivariate logistic regression (categorical)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM HELP_ = AGE.
           IF ( HELP_ LE 19.00000 ) THEN (NYVAR_ = 16.50000).
           IF ( HELP_ GT 19.00000 ) THEN (NYVAR_ = 21.00000).
           IF ( HELP_ GT 23.00000 ) THEN (NYVAR_ = 24.50000).
           IF ( HELP_ GT 26.00000 ) THEN (NYVAR_ = 35.50000).
           AGE = NYVAR_ .
/REGRESS DEPEND = LOW.
         CATEGORI= AGE
         .
         MODEL = AGE,LWT,SMOKE
         .
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          LNVAR_               0.000
          G LIKELIHO        -111.714
          G LIKELIHO        -109.854

ACTIVATING RULE
NO.:  22 (Rule checking for linearity in logit ).
RULES REMAINING ON THE STACK:    16  20  21  22

The external package has found:
Wald statistic for x1*ln(x1)     = 0.000000
This slot is needed to determine: p-value for x1 linear in logit

Considering slot:  p-value for x1 linear in logit. The corresponding slot
value has not yet been determined.
The system will attempt to find its value.

The external package has found:
Multivar. log-likelihood with x1 cont.     = -111.714000
This slot is needed to determine: Change in deviance (x1 cont. - x1 cat.)

The external package has found:
Multivar. log-likelihood with x1 cat.     = -109.854000
This slot is needed to determine: Change in deviance (x1 cont. - x1 cat.)

The system has found:
Change in deviance (x1 cont. - x1 cat.)     = 3.720001
This slot is needed to determine: p-value for x1 linear in logit

The system has found:
DF for test for linearity in x1     = 2
This slot is needed to determine: p-value for x1 linear in logit

The system has found:
p-value for x1 linear in logit     = 0.1556726
This slot is needed to determine: 'Is x1 linear in logit?'

ACTIVATING RULE
NO.:  21 (Rule rescaling the model ).
RULES REMAINING ON THE STACK:    16  20  21

The system has reached the following conclusion
to the question 'Is x1 linear in logit?'
The answer to this question is Yes!
This slot is needed to determine: 'Is x1 correctly scaled?'

ACTIVATING RULE
NO.:  20 (Main rule for rescaling ).
RULES REMAINING ON THE STACK:    16  20

The system has reached the following conclusion
to the question 'Is x1 correctly scaled?'
The answer to this question is Yes!
This slot is needed to determine: 'Has scaling been completed?'
```

```
Considering slot:  'Is x2 correctly scaled?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The system has reached the following conclusion
to the question 'Should x2 be treated as categorical?'
The answer to this question is No!
This slot is needed to determine: 'Has scaling been completed?'

ACTIVATING RULE
NO.:  21 (Rule rescaling the model ).
RULES REMAINING ON THE STACK:    16  20  21

Considering slot:  'Is x2 linear in logit?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The external package has found:
Number of levels for x2    = 75
This slot is needed to determine: 'Is x2 linear in logit?'

ACTIVATING RULE
NO.:  22 (Rule checking for linearity in logit ).
RULES REMAINING ON THE STACK:    16  20  21  22

The external package has found:
Minimum value of x2    = 80.000000
This slot is needed to determine: Wald statistic for x2*ln(x2)

Considering slot:  Wald statistic for x2*ln(x2). The corresponding slot
value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Multivariate logistic regression with x*ln(x)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM IF ( LWT GT 0 ) THEN XLNVAR_ = LWT*LN(LWT).
           IF ( LWT EQ 0 ) THEN XLNVAR_ = 0.
/REGRESS DEPEND = LOW.  ITER = 40.
         INTERVAL= AGE,LWT,SMOKE
, XLNVAR_.
         MODEL = AGE,LWT,SMOKE
, XLNVAR_.
/END
/PROB TITLE = 'Multivariate logistic regression (continuous)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM HELP_ = LWT.
           IF ( HELP_ LE 110.0000 ) THEN (NYVAR_ = 95.00000).
           IF ( HELP_ GT 110.0000 ) THEN (NYVAR_ = 115.5000).
           IF ( HELP_ GT 121.0000 ) THEN (NYVAR_ = 130.7500).
           IF ( HELP_ GT 140.5000 ) THEN (NYVAR_ = 195.2500).
           LWT = NYVAR_ .
```

```
/REGRESS DEPEND = LOW.
         INTERVAL = AGE,LWT,SMOKE
         .
         MODEL = AGE,LWT,SMOKE
         .
/END
/PROB TITLE = 'Multivariate logistic regression (categorical)'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/TRANSFORM HELP_ = LWT.
           IF ( HELP_ LE 110.0000 ) THEN (NYVAR_ = 95.00000).
           IF ( HELP_ GT 110.0000 ) THEN (NYVAR_ = 115.5000).
           IF ( HELP_ GT 121.0000 ) THEN (NYVAR_ = 130.7500).
           IF ( HELP_ GT 140.5000 ) THEN (NYVAR_ = 195.2500).
           LWT = NYVAR_ .
/REGRESS DEPEND = LOW.
         CATEGORI= LWT
         .
         MODEL = AGE,LWT,SMOKE
         .
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          LNVAR_                 0.000
          G LIKELIHO          -112.642
          G LIKELIHO          -110.133

ACTIVATING RULE
NO.:  22 (Rule checking for linearity in logit ).
RULES REMAINING ON THE STACK:    16  20  21  22

The external package has found:
Wald statistic for x2*ln(x2)    = 0.000000
This slot is needed to determine: p-value for x2 linear in logit

Considering slot:  p-value for x2 linear in logit. The corresponding slot
value has not yet been determined.
The system will attempt to find its value.

The external package has found:
Multivar. log-likelihood with x2 cont.    = -112.642000
This slot is needed to determine: Change in deviance (x2 cont. - x2 cat.)

The external package has found:
Multivar. log-likelihood with x2 cat.    = -110.133000
This slot is needed to determine: Change in deviance (x2 cont. - x2 cat.)

The system has found:
Change in deviance (x2 cont. - x2 cat.)    = 5.017990
This slot is needed to determine: p-value for x2 linear in logit

The system has found:
DF for test for linearity in x2    = 2
This slot is needed to determine: p-value for x2 linear in logit
```

```
The system has found:
p-value for x2 linear in logit    = 0.08134997
This slot is needed to determine: 'Is x2 linear in logit?'

ACTIVATING RULE
NO.:  21 (Rule rescaling the model ).
RULES REMAINING ON THE STACK:    16  20  21

The system has reached the following conclusion
to the question 'Is x2 linear in logit?'
The answer to this question is No!
This slot is needed to determine: 'Is x2 correctly scaled?'

The system has reached the following conclusion
to the question 'Should x2 be treated as categorical?'
The answer to this question is Yes!
This slot is needed to determine: 'Is x2 correctly scaled?'

The system has decided that the variable must be treated as categorical.
The variable is therefore grouped, with cutpoints Q1, median and Q3.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP DATA MANAGER.

        THE FOLLOWING COMMANDS WERE APPLIED:

READ SFILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
     CODE = MULTI.
     FILE = MULTFIL. /
TRANSFORM FILE = MULTFIL.
     HELP_ = LWT.
     IF ( HELP_ LT 110.0000 ) THEN (NYVAR_ = 95.00000).
     IF ( HELP_ GE 110.0000 ) THEN (NYVAR_ = 115.5000).
     IF ( HELP_ GE 121.0000 ) THEN (NYVAR_ = 130.7500).
     IF ( HELP_ GE 140.5000 ) THEN (NYVAR_ = 195.2500).
     LWT = NYVAR_ . /
SAVE FILE = MULTFIL.
     SFILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
     CODE = MULTI.
     DELETE = HELP_ , NYVAR_ .
     NEW. /
FINISH /

In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          VARIABLES         1

ACTIVATING RULE
NO.:  20 (Main rule for rescaling ).
RULES REMAINING ON THE STACK:    16  20

The system has reached the following conclusion
to the question 'Is x2 correctly scaled?'
The answer to this question is Yes!
This slot is needed to determine: 'Has scaling been completed?'
```

```
Considering slot:  'Is x3 correctly scaled?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The system has reached the following conclusion
to the question 'Should x3 be treated as categorical?'
The answer to this question is No!
This slot is needed to determine: 'Has scaling been completed?'

ACTIVATING RULE
NO.:  21 (Rule rescaling the model ).
RULES REMAINING ON THE STACK:    16  20  21

Considering slot:  'Is x3 linear in logit?'. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The external package has found:
Number of levels for x3    = 2
This slot is needed to determine: 'Is x3 linear in logit?'

The system has reached the following conclusion
to the question 'Is x3 linear in logit?'
The answer to this question is Yes!
This slot is needed to determine: Number of levels for x3

ACTIVATING RULE
NO.:  20 (Main rule for rescaling ).
RULES REMAINING ON THE STACK:    16  20

The system has reached the following conclusion
to the question 'Is x3 correctly scaled?'
The answer to this question is Yes!
This slot is needed to determine: 'Has scaling been completed?'

Considering slot:  Table with Wald statistics after scaling. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Multivariate logistic regression '.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/REGRESS DEPEND = LOW.
         CATEGORI =   LWT   .
         MODEL = AGE , LWT , SMOKE.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          G LIKELIHO        -110.145
          AGE               -0.04240
          AGE               -1.27
          LWT               -1.026
          LWT               -2.25
          (2)               -0.9872
          (2)               -2.16
          (3)               -0.9580
          (3)               -2.04
          SMOKE              0.6102
          SMOKE              1.85

ACTIVATING RULE
NO.:  20 (Main rule for rescaling ).
RULES REMAINING ON THE STACK:    16  20

Considering slot:  Table with Wald statistics after scaling. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.


                    Table with Wald statistics after scaling

  VARIABLE   |    TYPE    | WALD STATISIC    |  COEFFICIENT   |  REMOVABLE
-------------|------------|------------------|----------------|------------
AGE          |Adjustment| -1.270000        | -0.04240000    | No
LWT          |Ordinary  | (1)-2.250000     | (1)-1.026000   |
             |          | (2)-2.160000     | (2)-0.9872000|
             |          | (3)-2.040000     | (3)-0.9580000| No
SMOKE        |Study     | 1.850000         | 0.6102000      | No
                    Log-likelihood = -110.145000

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model
```

```
The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are interactions included?'
The answer to this question is No!
This slot is needed to determine: Summary of full model

ACTIVATING RULE
NO.:  24 (Rule detecting significant interactions ).
RULES REMAINING ON THE STACK:    16  24

Considering slot:  Overview of interactions. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

Considering slot:  Log-likelihood with study var.*x1. The corresponding
slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/REGRESS DEPEND = LOW.
         CATEGORI =   LWT
 .
         MODEL = AGE,LWT,SMOKE
 , SMOKE*AGE.
/PRINT LEVEL = BRIEF. CASE = 0. NO CORR.
/END
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/REGRESS DEPEND = LOW.
         CATEGORI =   LWT
 .
         MODEL = AGE,LWT,SMOKE
 , SMOKE*LWT.
/PRINT LEVEL = BRIEF. CASE = 0. NO CORR.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          G LIKELIHO       -109.754
          P-VALUE=           0.007
          G LIKELIHO       -106.696
          P-VALUE=           0.012

ACTIVATING RULE
NO.:  24 (Rule detecting significant interactions ).
RULES REMAINING ON THE STACK:    16  24

Considering slot:  Overview of interactions. The corresponding slot value
has not yet been determined.
The system will attempt to find its value.

The external package has found:
Log-likelihood with study var.*x1     = -109.754000
This slot is needed to determine: Overview of interactions

The system has reached the following conclusion
to the question 'Is interaction study var*x2 signific.?'
The answer to this question is Yes!
This slot is needed to determine: Overview of interactions


                    Overview of interactions

  Interaction    |  Log-likelihood    |       G      |   df   |  p-value
_________________|____________________|______________|________|______________
 Main effects    | -110.145000        |              |        |
 SMOKE*AGE       | -109.754000        | 0.7819977    | 1      | 0.3765309
 SMOKE*LWT       | -106.696000        | 6.897995     | 3      | 0.07522106

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model
```

```
The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are interactions included?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has final model been fitted?'
The answer to this question is No!
This slot is needed to determine: Summary of the interpretation

ACTIVATING RULE
NO.:  25 (Rule which fits the final model ).
RULES REMAINING ON THE STACK:    16  25

Considering slot:  Summary of full model. The corresponding slot value has
not yet been determined.
The system will attempt to find its value.

Considering slot:  Log-likelihood for full model. The corresponding slot
value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP LR.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Logistic regression for full model'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\MULTI.FIL'.
       CODE = MULTI.
/REGRESS DEPEND = LOW. ITERATION = 15.
         CATEGORI =   LWT
 .
         MODEL = AGE,LWT,SMOKE
         ,LWT*SMOKE
          .
/PRINT CELL=USED.
       CASE = 3.
/SAVE FILE = 'C:\EXPRESS\LOGREG\GOODNESS.FIL'.
      CODE = GOODNESS.
      CONT = CELL.
      NEW.
/END
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          PATTERNS          105
          AGE               14.0000
          AGE               45.0000
          AGE               23.2116
          SMOKE             0.0000
          SMOKE             1.0000
          SMOKE             0.3915
          LWT               41
          LWT               50
          LWT               51
          LWT               47
          G LIKELIHO       -106.696
          2*O*LN(O/E        0.012
          HOSMER-LEM        0.895
          95% C.I.          ****
          AGE              -0.03676
          AGE              -1.07
          AGE               0.964
          LWT              -2.106
          LWT              -3.15
          LWT               0.122
          (2)              -1.558
          (2)              -2.45
          (2)               0.211
          (3)              -1.899
          (3)              -2.80
          (3)               0.150
          SMOKE            -0.6337
          SMOKE            -0.992
          SMOKE             0.531
          *                 2.230
          *                 2.37
          (2)               1.075
          (2)               1.17
          (3)               1.876
          (3)               1.98
          CONSTANT          1.185
          CONSTANT          1.38
          CONSTANT          3.27
          CELL FREQ         0.11
          LESS THAN         209

ACTIVATING RULE
NO.:  25 (Rule which fits the final model ).
RULES REMAINING ON THE STACK:    16  25

Considering slot:  Summary of full model. The corresponding slot value has
not yet been determined.
The system will attempt to find its value.

The external package has found:
Log-likelihood for full model    = -106.696000
This slot is needed to determine: Summary of full model
```

```
                    Summary of full model

 VARIABLE    |   TYPE     | WALD STATISTIC |  COEFFICIENT   |  EXP(COEFF)
-------------|------------|----------------|----------------|-------------
AGE          |Adjustment|  -1.070000      | -0.03676000    | 0.9640000
LWT          |Ordinary   |  (1)-3.150000   | (1)-2.106000   | 0.1220000
             |           |  (2)-2.450000   | (2)-1.558000   | 0.2110000
             |           |  (3)-2.800000   | (3)-1.899000   | 0.1500000
SMOKE        |Study      |  -0.9920000     | -0.6337000     | 0.5310000
SMOKE*LWT    |           |  (1)2.230000    | (1)2.370000    |
             |           |  (2)1.075000    | (2)1.170000    |
             |           |  (3)1.876000    | (3)1.980000    |
Constant     |           |  1.380000       | 1.185000       | 3.270000

             Log-likelihood for full model = -106.696000

ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are interactions included?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has final model been fitted?'
The answer to this question is Yes!
This slot is needed to determine: Summary of the interpretation
```

```
The system has reached the following conclusion
to the question 'Has goodness of fit been assessed?'
The answer to this question is No!
This slot is needed to determine: Summary of the interpretation

ACTIVATING RULE
NO.:  26 (Rule carrying out goodness of fit test ).
RULES REMAINING ON THE STACK:    16  26

Considering slot:  'Is summary goodness of fit poor?'. The corresponding
slot value has not yet been determined.
The system will attempt to find its value.

The external package has found:
Number of covariate patterns    = 105
This slot is needed to determine: p-value overall goodness of fit

The external package has found:
Goodness of fit, Pearson    = 0.01200000
This slot is needed to determine: p-value overall goodness of fit

The external package has found:
Goodness of fit, Hosmer & Lemeshow    = 0.8950000
This slot is needed to determine: p-value overall goodness of fit

The system has found:
p-value overall goodness of fit    = 0.8950000
This slot is needed to determine: 'Is summary goodness of fit poor?'

The system has reached the following conclusion
to the question 'Is summary goodness of fit poor?'
The answer to this question is No!
This slot is needed to determine: 'Should new analys. be done due to GOF?'

The system has reached the following conclusion
to the question 'Should new analys. be done due to GOF?'
The answer to this question is No!

Explanation of the slot: 'Should new analys. be done due to GOF?'
If the goodness of fit is poor, you may decide to restart the
multivariate analysis. This time the limits for including an
independent variable in the model will be less restrictive.
This slot is needed to determine: Summary of the interpretation

Considering slot:  Table with influential covar. patterns. The
corresponding slot value has not yet been determined.
The system will attempt to find its value.

ACTIVATING RULE
NO.:  27 (Rule detecting influential points in the final model ).
RULES REMAINING ON THE STACK:    16  26  27

The system has reached the following conclusion
to the question 'Has data file been generated?'
The answer to this question is No!
This slot is needed to determine: Table with influential covar. patterns
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
```

```
The following package is executed: BMDP 1D.

        THE FOLLOWING COMMANDS WERE APPLIED:

/PROB TITLE = 'Express executing BMDP 1D'.
/INPUT FILE = 'C:\EXPRESS\LOGREG\GOODNESS.FIL'.
       CODE = GOODNESS.
/SAVE FILE = 'C:\EXPRESS\LOGREG\GOODNESS.TO'.
      CODE = GOODTO.
      NEW.
      DELETE = OBSPROP,SEPRED,STDRESID,LOGODDS,DEVIANCE.
/END



In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS         VALUES EXTRACTED

          VARIABLES         1

ACTIVATING RULE
NO.:  27 (Rule detecting influential points in the final model ).
RULES REMAINING ON THE STACK:    16  26  27

The system has reached the following conclusion
to the question 'Has data file been generated?'
The answer to this question is Yes!
This slot is needed to determine: Table with influential covar. patterns

Considering slot:  Plot: est.success * change in coeff.. The corresponding
slot value has not yet been determined.
The system will attempt to find its value.
Please note: Some packages will during execution destroy
      the screen produced by EXPRESS, but this screen will
      be restored when the execution is complete.
The following package is executed: BMDP DATA MANAGER.

        THE FOLLOWING COMMANDS WERE APPLIED:

READ SFILE = 'C:\EXPRESS\LOGREG\GOODNESS.TO'.
     CODE = GOODTO.
     FILE = FILEN.  /
SORT KEY = INFLUENCE.
     ORDER = D.  /
TRANSFORM DELTAX2 = (NRMRESID**2)/(1 - HATDIAG).
          MJ = SUCCESS + FAILURE.
          IF ( SUCCESS GT 0 ) THEN
             (HJ1 = SUCCESS*(LN(SUCCESS/(MJ*PREDPROB))).)
          ELSE
             (HJ1 = 0 .).
          ALLSUC = MJ - SUCCESS .
          IF ( ALLSUC GT 0 ) THEN
             (HJ2 = ALLSUC*(LN( ALLSUC/(MJ*(1 - PREDPROB)))).)
          ELSE
             (HJ2 = 0 .).
          HJ = HJ1 + HJ2.
          DHJELP = SQRT(2*HJ).
          HJELPTO = SUCCESS - MJ*PREDPROB.
          IF ( HJELPTO GE 0 ) THEN (DEVI = DHJELP .)
          ELSE (DEVI = -DHJELP .).
          DELTAD = (DEVI**2)/(1 - HATDIAG).  /
PLOT XVAR = PREDPROB,PREDPROB,PREDPROB.
     YVAR = INFLUENCE,DELTAX2,DELTAD.
     SIZE = 60,25. /
TRANSFORM NEWFILE = NYFIL.
          IF (INFLUENCE LT 0.8) AND (DELTAX2 LT 4) AND (DELTAD LT 4)THEN
             (USE = -1).
          IF ( INFLUENCE EQ XMIS ) THEN ( USE = -1). /
SAVE FILE = NYFIL.
     SFILE = 'C:\EXPRESS\LOGREG\DIVERSE.DAT'.
     NEW.
     DELETE = ALLSUC,HJ,HJ1,HJ2,HJELPTO,DEVI,DHJELP,SUCCESS,MJ,PREDPROB,
              NRMRESID,HATDIAG,INFLUENC,DELTAX2,DELTAD,FAILURE.
     FORMAT = '(10F12.6)'. /
SAVE FILE = NYFIL.
     SFILE = 'C:\EXPRESS\LOGREG\DIVERSE.INF'.
     NEW.
    KEEP = SUCCESS,MJ,PREDPROB,NRMRESID,HATDIAG,INFLUENC,DELTAX2,DELTAD.
     FORMAT = '(8F7.3)'. /
FINISH /
```

```
In the output file produced by the execution of the external package,
EXPRESS has used the following search keys:

          SEARCH KEYS          VALUES EXTRACTED

          .+                A plot or table has been extracted.
          .+                A plot or table has been extracted.
          .+                A plot or table has been extracted.

ACTIVATING RULE
NO.:  27 (Rule detecting influential points in the final model ).
RULES REMAINING ON THE STACK:    16  26  27

The system has reached the following conclusion
to the question 'Has data file been generated?'
The answer to this question is Yes!
This slot is needed to determine: Table with influential covar. patterns

ACTIVATING RULE
NO.:  26 (Rule carrying out goodness of fit test ).
RULES REMAINING ON THE STACK:    16  26

The system has reached the following conclusion
to the question 'Is summary goodness of fit poor?'
The answer to this question is No!
This slot is needed to determine: 'Should new analys. be done due to GOF?'

The system has reached the following conclusion
to the question 'Should new analys. be done due to GOF?'
The answer to this question is No!
```

Explanation of the slot: 'Should new analys. be done due to GOF?'
If the goodness of fit is poor, you may decide to restart the multivariate analysis. This time the limits for including an independent variable in the model will be less restrictive.
This slot is needed to determine: Summary of the interpretation

Table with influential covar. patterns

Table: Covariate patterns

AGE	LWT	SMOKE
18.000	95.000	1.0000
19.000	195.25	1.0000
25.000	95.000	0.0000
24.000	130.75	0.0000
15.000	115.50	0.0000

Table continued : Diagnostic statistics

y	m	π	r	h	ΔBETA	ΔX**2	ΔD
0.000	5.000	0.472	-2.116	0.277	2.379	6.197	8.851
0.000	4.000	0.458	-1.838	0.272	1.735	4.639	6.727
4.000	4.000	0.566	1.751	0.222	1.124	3.940	5.850
3.000	4.000	0.222	2.543	0.127	1.078	7.408	5.773
2.000	2.000	0.187	2.953	0.087	0.910	9.552	7.356

The system has found:
Number of influential patterns = 5
This slot is needed to determine: Table with influential covar. patterns

The user has reached the following conclusion
to the question 'Should obs be rem. and new anal. done?'
The answer to this question is No!
The system has not yet determined this slot value.

Explanation of the slot: 'Should obs be rem. and new anal. done?'
The system has identified influential observations. You can now decide whether these should be removed and the multivariate analysis restarted.

ACTIVATING RULE
NO.: 16 (Main rule for complete analysis).
RULES REMAINING ON THE STACK: 16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

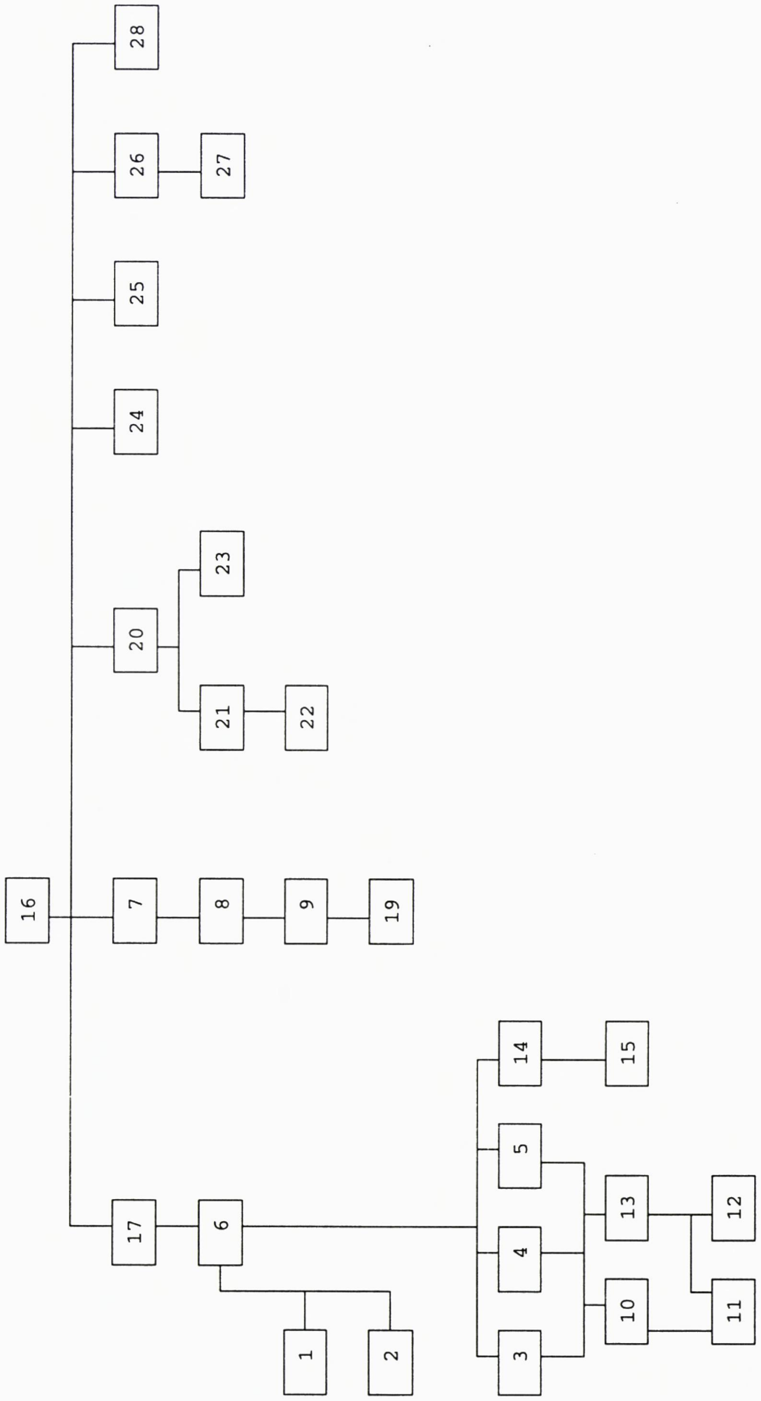
The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file, including all variables selected in the univariate anaysis, must be generated.
This slot is needed to determine: Summary of full model

```
The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are interactions included?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has final model been fitted?'
The answer to this question is Yes!
This slot is needed to determine: Summary of the interpretation

The system has reached the following conclusion
to the question 'Has goodness of fit been assessed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of the interpretation

The system has reached the following conclusion
to the question 'Has interpretation been carried out?'
The answer to this question is No!

ACTIVATING RULE
NO.:  28 (Rule interpreting the final results ).
RULES REMAINING ON THE STACK:    16  28

Considering slot:  Summary of the interpretation. The corresponding slot
value has not yet been determined.
The system will attempt to find its value.


                  Summary of the interpretation

Odds ratio for variables not included in interactions:

     VARIABLE    |       ODDS RATIO
____________________________________________
AGE              | 0.9640000

Odds ratio for the study variable:


          LWT
       _________________________________
SMOKE  | Gr1.     Gr2.     Gr3.     Gr4.  |
_______|__________________________________|
MIN-AVR|   .780   1.868   1.189   1.626  |
MIN-MAX|   .531   4.935   1.555   3.464  |
Adjusting for the variable(s):
```

```
ACTIVATING RULE
NO.:  16 (Main rule for complete analysis ).
RULES REMAINING ON THE STACK:    16

The system has reached the following conclusion
to the question 'Is the univariate analysis complete?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has multivariate file been completed?'
The answer to this question is Yes!

Explanation of the slot: 'Has multivariate file been completed?'
Before the analysis can proceed, a multivariate BMDP file,
including all variables selected in the univariate anaysis,
must be generated.
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are any variables strongly correlated?'
The answer to this question is No!
This slot is needed to determine: Final table with Wald statistic and coef

The system has reached the following conclusion
to the question 'Has variable selection been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Has scaling been completed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model

The system has reached the following conclusion
to the question 'Are interactions included?'
The answer to this question is Yes!
This slot is needed to determine: Summary of full model
The system has reached the following conclusion
to the question 'Has final model been fitted?'
The answer to this question is Yes!
This slot is needed to determine: Summary of the interpretation

The system has reached the following conclusion
to the question 'Has goodness of fit been assessed?'
The answer to this question is Yes!
This slot is needed to determine: Summary of the interpretation

The system has reached the following conclusion
to the question 'Has interpretation been carried out?'
The answer to this question is Yes!
```

Simple strategy map



Overview of different rules in LOGISTRULE. If to rules are conected thay may call eachother.

Description of rules

Simple explanations are given below of the different rules included in Logistrule. The rules make extensive use of the Fortran utility routines provided by Express. Because of restrictions on the amount of memory available, the main program in Logistrule has been divided into six main subunits, which are used to execute the chaining of rules. These programs obtain a rule number as input, and the proper rule is called. Each such main program has the same structure as any other main program in a set of rules constructed using Express. The actual source code for the rules is given in the technical report by Aarseth (1996).

The numbering of the separate rules was partly determined during the development of Logistrule and does not correspond to any logical order. The chaining always starts with rule number 16.

Subroutine name: **REG1**
Rule number: **1**

Description: This rule makes certain that the dependent variable only has two levels and that these are the values zero and one. If the variable includes more than two levels, the user can specify a cut-point to be used to separate the groups.

Structure:

```
Get number of levels for the dependent variable
Get minimum and maximum
        If only 1 level then
                Exit
        else if 2 levels which are correctly coded
                Continue
        else
                Decide cut-point
                Execute package that recodes the variable
        end
```

Subroutine name: **REG2**
Rule number: **2**

Description: This rule first performs initial preparation in connection with a selected independent variable. This includes creating a data file and warning the user if the number of observations is not identical to that for the dependent variable. It is decided whether the variable is nominal or interval scaled. Variable type (study, adjustment or ordinary variable) is selected. Depending on type, this rule indicates which rule should be executed next.

Structure:

```
If dependent variable has not been prepared then
        Goto rule number 1
end
If file with dependent and independent variable has not been generated then
        Start BMDP to create file
end
Get number of levels for independent variable
If the number of observations is not identical to that
                                for the dependent variable then
        Give warning
end
Get decision whether variable is nominal or interval scaled
Get type of variable and return proper rule number to be execute next
```

Subroutine name: **REG3**
Rule number: **3**

Description: This is the main rule for univariate analysis of the selected study variable. This rule merely activates the rule needed to perform the correct analysis. There are two alternative rules, one for analysis of nominal variables or variables with 8 or fewer levels (rule number 10), and another one for analysis of interval scaled variables with more than 8 levels (rule number 13). At the end of the univariate analysis, when this rule is restarted, it will execute necessary statements to ensure that the study variable is included in the multivariate model.

Subroutine name: **REG4**
Rule number: **4**

Description: In exactly the same way as rule number 3, this rule decides which one among the two alternative rules 10 and 13 should be executed. However, in this case the variable considered is an adjustment variable.

Subroutine name: **REG5**
Rule number: **5**

Description: Again exactly as rule number 3, this time for an ordinary variable. When control is returned to this rule at the end of the univariate analysis, the decision whether the variable should be included in the multivariate analysis is taken considering the tests performed during the intermediate analysis.

Subroutine name: **REG6**
Rule number: **6**

Description: Main rule for univariate analysis. This rule first decides which independent variable the analysis should be performed on. It tries to present a summary of the analysis. If this summary has not yet been found, the proper rule will be started (rule number 2). At the end of the analysis, the rule will decide whether grouping of the independent variable is necessary. This decision is based on the value of the statistical significance slot and the number of levels for the variable.

Structure:

```
Get number of independent variable to be analysed
Get summary slot for univariate analysis
If not found then
        Goto rule number 2
end
If grouping necessary then
        Goto rule number 14
end
```

Subroutine name: **REG7**
Rule number: **7**

Description: Main rule for variable selection. This rule fits a multivariate model and produces a summary table of the fit, including variable names, Wald statistics and coefficients from the multivariate model fitted. The number of non-significant variables is found and a new rule (rule number 8) will be started to attempt to remove ordinary non-significant variables. When control has been retrieved, a new fit of the reduced model is carried out, and again non-significant variables may appear.

Structure:

```
Repeat
Get summary of the fitted model
If not found then
        Fit the model
        Make summary and count number of removable variables
        Goto rule number 8
end
Until every variable is significant
```

Subroutine name: **REG8**
Rule number: **8**

Description: This rule first executes rule number 9 which finds the final variables that can be removed from the model. After retrieving control, this rule actually discards the variables from the analysis.

Structure:

```
Get the number of non-significant variables that can be removed from the model
If not found then
        Goto rule number 9
else if some variables should be removed then
        Remove variables
end
```

Subroutine name: **REG9**
Rule number: **9**

Description: This rule compares the coefficient of the study variable in the full model with the coefficient in the reduced model. A large change in the coefficient should be taken as an indication of confounding, and not all non-significant independent variables should be removed from the model. If there is more than one potential confounder, rule number 19 is executed to make the selection.

Structure:

```
Get the change in the coefficient(s) for the study variable
If any confounders then
        If more than one candidate then
                Goto rule number 19
        else
                No variables must be removed
        end
else
        All non-significant variables may be removed
end
```

Subroutine name: **REG10**
Rule number: **10**

Description: This rule takes care of the univariate analysis for independent variables on a nominal scale, and for variables in general with the number of levels less than 8. The first action is to make a plot of the observed risks for the different levels. Second, it is decided whether a contingency table analysis should be performed or if logistic regression should be applied to fit a univariate model. After the proper BMDP program has executed, the rule will assign the correct value to the statistical summary slot.

Structure:

```
Get plot of observed risks
Get type of analysis (contingency table or logistic regression)
If contingency table then
        Use BMDP 4F to conduct the proper analysis
        Decide value of the significance summary slot
else
        Use BMDP LR to fit a univariate logistic model
        Decide value of the significance summary slot
end
```

Subroutine name: **REG11**
Rule number: **11**

Description: The main purpose of this rule is to generate a plot of observed risks for different levels of a variable. This plot will be displayed during the univariate analysis. If the independent variable has more than 8 levels, the quartiles and the median will be used as cut-points, subdividing the range into 4 groups.

Structure:

```
Get the number of levels for independent variable in question
If number of levels is greater than 8 then
        Make plot after grouping with BMDP LR
else
        Make plot with BMDP LR
end
```

Subroutine name: **REG12**
Rule number: **12**

Description: The purpose of this rule is to find *p*-values in the univariate situation for an independent variable with less than 8 levels, provided that the variable should be analysed with logistic regression. The model will include the independent variable on an interval scale. In addition to the *p*-value for the test of trend, a *p*-value is calculated for departure from a linear model.

Structure:

```
Get deviance difference for models with and without the independent variable
If not found then
        Fit models with BMDP LR
end
Get deviance for model containing the variable as categorical
Get deviance for model containing the variable as interval
Calculate deviance difference
Calculate p-value for test for departure from a linear model
```

Subroutine name: **REG13**
Rule number: **13**

Description: The rule for conducting the univariate analysis for an independent variable with more than 8 levels. It first invokes rule number 11 to get a plot of observed risks. Second, BMDP LR is used to fit a model which makes it possible to compute a *p*-value for trend. The last step is to find a *p*-value for departure from linearity. During this step, the independent variable is divided into 4 groups, using quartiles and median as cut-points. Two models are fitted, first one including the variable as interval scaled and then a model with the variable included as categorical.

Structure:
Get plot of observed risks
Get *p*-value for test for trend
Get deviance for model containing the grouped variable as interval variable
Get deviance for model containing the grouped variable as category variable
Calculate *p*-value
Decide value of summary slot for statistical significance

Subroutine name: **REG14**
Rule number: **14**

Description: This rule only decides if an independent variable should be grouped before the multivariate analysis starts. If the summary significance slot has the values 1 or 2 and the variable is not nominal-scaled, grouping is activated (through rule number 15).

Subroutine name: **REG15**
Rule number: **15**

Description: If it has been decided by rule number 14 that grouping should be carried out, this rule executes the grouping. The first action is to decide if the first group should be left unchanged. Then the rule proposes new cut-points. If the user is not satisfied with these, he/she may specify their own cut-points. Finally, the variable must be recoded, and for this purpose the BMDP DATA MANAGER is used.

Structure:
Decide if the first group shall be unchanged
Compute new cut-points
Let the user get the possibility to override the grouping proposed by the rule
Make necessary changes in the multivariate file using BMDP DM

Subroutine name: **REG16**
Rule number: **16**

Description: This is the primary rule in Logistrule, which can invoke the main rules needed to complete the analysis. This rule is activated first in any application of Logistrule where the purpose is a full analysis.

Structure:
Activate the univariate analysis (rule number 17)
Make BMDP file containing significant variables from the univariate analysis
Check correlation among independent variables
Activate variable selection (rule number 7)
Activate necessary rescaling of independent variables (rule number 20)
Activate rule for inclusion of interactions (rule number 24)
Activate rule for fitting final model (rule number 25)
Activate goodness of fit rule (rule number 26)
Activate interpretation of fitted model (rule number 28)

Subroutine name: **REG17**
Rule number: **17**

Description: This rule is used when we want to carry out univariate analysis step by step for all independent variables. It is simply a loop which activates the main rule for univariate analysis each time (rule number 6) with a new independent variable as input. This loop will continue to start rule number 6 until all independent variables have been analysed.

Subroutine name: **REG19**
Rule number: **19**

Description: Additional rule for selection of variables. This rule is in charge when the decision is to be made concerning variables removed from the model when a confounding effect has been demonstrated. This rule is activated if rule number 9 detects several independent variables which can be confounders. The rule must decide which variables should be retained in the model and which should be removed. This is done by reentering the non-significant variables one-by-one. This is repeated until the difference between the coefficients of the study variable from the full model and the model fitted by this rule is reduced to a value indicating that no confounders are excluded.

Structure:

```
Repeat
        Include one of the non-significant variables
        Find difference between coefficients for the study variable
Until coefficient change is small
```

Subroutine name: **REG20**
Rule number: **20**

Description: This rule goes through all independent variables in search of variables that may need to be rescaled. Rescaling is carried out if an interval scaled variable is found to be non-linear in logit. This rule does not itself test for non-linearity or perform the necessary scaling, but it activates the rules that can handle these tasks (rule number 21). When the rule has finished the loop through all variables, it will fit the model after scaling and construct a summary table.

Structure:

```
Repeat
        Activate rule for scaling (rule number 21)
Until every variable in the model checked
Fit model with rescaled variables
Make summary of new model
```

Subroutine name: **REG21**
Rule number: **21**

Description: This rule performs necessary recoding on the BMDP file used in the analysis if the variable considered is non-linear in logit. Before any scaling is done, this rule will try to determine whether the variable is linear in logit. If the variable has only two levels, the answer is affirmative, otherwise rule number 22 must be executed before the answer can be given. If the variable considered must be recoded, there are two possibilities. First, if the number of levels is less than 8, rule number 23 is executed to perform the grouping, otherwise this rule handles the task itself.

Structure:

```
Get is variable linear in logit?
If not found then
        If variable has only two levels then
                Is linear in logit
        else
                Goto rule number 22
        end
end
If not linear in logit then
        If number of levels is less than 8 then
                Goto rule number 23
        else
                start BMDP DM to recode the variable
        end
end
```

Subroutine name: **REG22**
Rule number: **22**

Description: Rule that fits the appropriate models to decide whether a variable is linear in logit. First, if there are no negative values, the term $x\ln x$ is included in the model. If the coefficient of this term is significant, non-linearity is indicated. In addition we compare two models with the variable in question included as category and interval variables, respectively. These two tests together determine if the variable should be regarded as linear in logit or not.

Structure:

```
If no values below zero then
        Decide if $x\ln x$ is significant
end
Get $p$-value for test for departure from linear model
Decide if variable is linear in logit
```

Subroutine name: **REG23**
Rule number: **23**

Description: If rule number 21 is not able to make the necessary grouping, this rule will be activated. This rule is almost identical to rule number 15, but it includes small changes necessary for the multivariate situation.

Subroutine name: **REG24**
Rule number: **24**

Description: Rule that fits up to 9 different model, each model including a specified interaction. If any of the interactions are significant, they will be included in the complete model. The purpose of this rule is to decide which interactions to include.

Structure:

```
Get are there any significant interactions?
If not found then
        fit necessary models
        Use $p$-values for statistical significance to decide if interactions shall be included
end
```

Subroutine name: **REG25**
Rule number: **25**

Description: This rule simply performs the final fit of the model, with the rescaled variables and any significant interactions included. A table summarizing the model is presented.

Subroutine name: **REG26**
Rule number: **26**

Description: After the final model has been fitted, goodness of fit is assessed. This rule starts with a presentation of global goodness of fit statistics. If these indicate a poor fit, the user may decide that the analysis shall be restarted. The second check is identification of influential observations. Observations identified as those with the strongest influence on the coefficients in the final model, are presented to the user. On the basis of these values, the user may decide that the corresponding observations should be removed and the multivariate analysis restarted.

Structure:

```
Get global goodness of fit
If poor fit then
        Ask user if analysis shall be redone
end
Get influential observations
If not found then
        Goto rule number 27
end
Ask user if observations shall be removed and analysis redone
```

Subroutine name: **REG27**
Rule number: **27**

Description: This rule first calculates the statistics (using BMDP) needed for identification of influential observations. It then generates a table with the covariate patterns selected.

Subroutine name: **REG28**
Rule number: **28**

Description: This rule takes care of the interpretation. This first involves calculation of odds ratios for independent variables not included in any interactions. Second, odds ratios are calculated for the study variable (which is always included in any significant interactions).

Structure:

```
Get odds ratio for variables not included in any interactions
Get odds ratio for the study variable
Present the interpretation
```