A CONSULTATION SYSTEM FOR STATISTICAL ANALYSIS

September 1992

Atsuhiro Hayashi

A CONSULTATION SYSTEM FOR STATISTICAL ANALYSIS

September 1992

Atsuhiro Hayashi

Department of Mathematics
Kawasaki Medical School

Contents

1. Introduction and Summary

The development and popularization of personal computers have had a strong influence on the field of statistics during the last ten years. The analysis of data which was once carried out only on main frame computers can now even at home with a personal computer.

Today every practicing statistician uses a computer and some statistical programs. Some statisticians use terminals accessing large main frames which run classical software packages such as SPSSX, BMDP or SAS. Most of this software is also available for the personal computers they have on their desks.

The range of types of software and hardware in use is considerable. This easier access to computational facilities has led to non-statisticians often making use of statistical software — whether or not they understand the theory underlying the methods they use. This increased user base has resulted in a strong commercial interest in the statistical software market. There has, in fact, been a statistical software boom.

But, many users who have data sets and purposes for analyses are worried about how to select statistical methods and combine them, or their possible misuse of methods, because they lack knowledge of and experience with statistical analysis.

In the field of Artificial Intelligence(AI), there are some representational models of knowledge. The semantic network model is an effective model for the hierarchical knowledge, and the knowledge may be easily modified.

To deal with the above mention problems, we have developed a consultation system for statistical analysis based on hypertool. This system advises to the user one appropriate statistical method based on our knowledge base linking of the knowledge and programs of statistics.

The system derives a statistical method to fit the user's purpose by means of a dialogue between the user and this system. The system is designed for data analysts and students without statistical knowledge or experience.

In this paper, we describe a statistical consultation system that provides assistance to non-statisticians. There are many kinds of statistical software programs, but most of them require some knowledge of and experience with statistics. Therefore, the development of a consultation system that can pro-

vide knowledge of data analyses is eagerly desired by many users.

The semantic network is one representational model of knowledge. This model is effective for hierarchical knowledge and is easy to modify. The hypertool has the same structure as the semantic network model.

We developed the consultation system based on hypertool. This system is expandable and has the advantage of being able to add/modify statistical knowledge.

In Chapter 2, for preparation, we surveyed former statistical consultation systems. There are many consultation systems for statistics in the world. Each system has its characteristic features. Therefore we have summarized these systems.

Chapter 3 deals with the nature of statistical analysis. In the case of ordinary data analysis, an analyst uses many statistical methods and techniques based on the experience in his mind. We observed his analysis processes, since the knowledge of these works is important in building a consultation system.

In the field of AI, there have many kinds of studies about acquirement, storage and use of knowledge to imitate human judgment. In Chapter 4, we describe some knowledge representational techniques. Especially, the semantic network model is an effective model for the hierarchical knowledge, and the knowledge may be easily modified. Then, we discuss this model in detail.

We developed a new statistical software program which we named Seto/B that makes good use of personal computers. It includes more than 50 statistical methods, from introductory to advanced, with graphical outputs and 8 powerful utilities for treating the data file. In Chapter 5, we explain the structure and features of Seto/B, because Seto/B is one important part of our consultation system.

To create a knowledge data base with respect to selection of a method, we investigated the relations between the purposes of analyses, type of data and methods of statistical analyses. We also constructed a decision rule as a tree structure which we call a 'statistical strategy map'. We discuss how we create this map in Chapter 6.

The structure of the strategy map is similar to the semantic network model. We assigned 'nodes' of the semantic network to question, dictionary and pro-

gram, and 'links' to selection, referring and execution, respectively. There are some computer tools for dealing with the semantic network model. One of this software is called 'hypertool' which refers to text files and programs mutually with links. We constructed a consultation system using this software, known as the 'Statistical Consultation System based on Hypertool(SCSH)'. In Chapter 7, we describe the structure of this system and illustrate it with an example of the process of analysis. (A part of the statistical knowledge in SCSH is shown in the appendix.)

Evaluation of the system is treated in Chapter 8. The system has been used experimentally by some people. As a whole, this is a convenient and useful system. We discuss some of the features of our system and compare it with other consultation systems.

Finally in Chapter 9, we present our conclusions concerning our consultation system. The system possesses flexibility with regard to the addition, modification and improvement of knowledge using 'hypertool'. This is important for a statistical consultation system. We have confirmed that SCSH provides effective support for statistical data analysis. But it is not enough to only display information and process statistical results with some rules. The statistical knowledge of one research field and data analysis should be synthesized. Further experience in using this system in actual fields of research should make it more effective.

2. Survey of Statistical Knowledge System

Many kinds of statistical software programs are available, but most of them require the knowledge and experience of statistics. In such a circumstance beginners of data analysis usually meet troubles when they start an analysis without any advisors.

Over against these situations, there are some consultation systems for statistics in the world. These systems incorporate knowledge and experience of statistics. Each system has the feature of realization technique. At the starting point of the study, we survey these systems.

2.1 REX

The pioneer of the statistical expert system is Gale and Pregibon (1982). They noticed the relation of knowledge base and statistics. Gale (1986) introduced the first expert system of statistics called REX (Regression EXpert).

REX encodes enough knowledge to do a simple linear regression analysis safely. Its methods are to systematically check assumptions underlying the use of regression analysis. If an assumption violation is detected, it determines a correction or signals an unexpected problem. REX is coded using expert system techniques as an interface to the S Statistical System (Becker and Chambers (1984)). The inference engine interprets the statistical knowledge, and interacts with the user and the statistical system. It runs in a multiple windows environment, one for dialogue, one for graphics, one for interpretations of tests and other for various information. And REX makes a report file of the executing analysis. It explains the process of analysis performed by REX with values, formulas and graphs. Therefore, the user can trace and understand his own analysis.

After REX, he would plan a new system 'Student' (1986). It is designed to allow a professional statistician to construct a knowledge-based consultation system. This system grows its knowledge in the process of selecting methods, analyzing working examples and answering questions. Student is front-end of S Statistical System, too. REX refers fixed statistical knowledge, however Student has the improvement facility of knowledge. There are a lot of statistical knowledge and each is flexible to modify. We cannot collect all of them at the

first, so the improvement facility is very important and useful.

2.2 S/EXP

Ishibashi and Takeda (1990) discussed the expert system based on S Statistical System named S/EXP. The aim of this system is quite similar to the REX. The major differences are that S/EXP expects a beginner's usage, and that it provides a check list for statistical tests and statistical knowledge in a hierarchical structure in order to understand easily for him.

The knowledge in the system is described using 'IF THEN' production rules. Each rule is composed of the condition and the action of statistics. If a user does not agree with the rule, he can modify the threshold value in the condition.

To assist user interface, it supports the multiple windows environment on engineering work station. Each window displays a statistical checking list, a result of pre/post-testing, a result of analysis and so on. The user confirms these windows mutually, and advances his analysis. S/EXP has help facility, so the system shows the explanation of the result and reason of the testings. At the present this system supports only regression like REX, but they have a plan to expand the functions and knowledge of the system to other analyses.

In their paper, they showed the process of data analysis and the knowledge and experience necessary in such process. Also they discussed the possibilities to incorporate them in an expert system. After this, they classified into two categories from the viewpoint of easiness of incorporation in a system. It is a good reference to construct these consultation systems.

2.3 RASS

Nakano, Yamamoto and Okada (1991) discussed RASS (Regression Analysis Supporting System). It is specified to regression analysis, too. The reason that above 3 systems treated only regression analysis is that this analysis is well known and well researched. And it is most frequently to apply to real field. So, these systems treated only this analysis at the first point.

The one of features is that RASS based on the technique of Object Oriented Programming (OOP). OOP is a new concept of computer programming

paradigm. Its basic ideas are 'Object' and 'Message sending'. The former is the group of programs and connected data, and the latter is the arguments between programs. Each Object is activated only when it receives a Message. After that, the Object judges the actual job related to the Massage, and it operates to the data of itself. Another concept of OOP is 'Inheritance'. It is the hierarchical structure of 'Object', and the lower Objects (SubClass) hold the characteristic of upper Objects (SuperClass). This feature is useful to describe the relation of master and servant of Objects. If the Object cannot execute the job itself, then the Object delivers to its upper Objects.

The user of RASS drives to send a Message to an Object. This situation is same as a researcher analysis a data by a statistical package. He commands to the package by the name of the suitable procedure.

This system has a semi-auto mode. If user has no idea of the next analysis, then he requests a recommendation of an operation and follows it. It leads to the correct result of regression analysis.

RASS is programmed in Prolog as inference engine, in C as numerical execution and graphics, and in X-Windows as user interface. Prolog is a typical computer language for artificial intelligence(AI). And its facility of 'backtracking' is effective to find a specified knowledge in the hierarchical structure. X-Windows offers the system of multiple windows on many kinds of work station. It is the standard environment, so it is easy to install other computers.

2.4 The Statistical Consultant

Robert S. (1987) published a shareware system named 'The Statistical Consultant'. This system distributed on bulletin board system (BBS) or PC-SIG (i.e., CD-ROM service of BBS). It is an authorized implementation of 'A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, Second Edition' (Frank, et al.(1981)).

The Statistical Consultant is an expert system designed to assist the user in selecting the appropriate statistical test for his problem. The system will ask a series of questions, starting with, 'how many variables do you have?'. The responses to questions leads to the identification of a particular statistical technique. Most questions are phrased for yes/no responses. In these cases

one needs only type 'y' or 'n'. Occasionally other responses may be required. The system shows the suitable method of statistics with procedure name of statistical packages (Osiris, SPSS and SAS).

This system suggests not only statistical methods but also references, so user can know some books to learn in detail. And they prepared some glossary of terms used by the consultant in text file, then user can look up an unknown word in the suggestion.

The Statistical Consultant is only the suggestion of statistical method with procedure name of package, not execute analysis. But this system requires a small personal computer with MS-DOS, so users refer in many fields, for example in laboratory, in classroom or beside mainframe terminal.

2.5 Other Systems in COMPSTAT '90

In 1990, COMPSTAT (the congress of the International Association for Statistical Computing) was held in Dubrovnik, Yugoslavia. There were some publications about statistical expert.

Van den Berg and Visser (1990) discussed the design of computerized support in statistics. They noticed that the individual differences between the ideas of experts on statistical consultation and on the application of analysis methods are so large. They held two empirical studies, and got some results. In the first exploratory investigation, they asked twenty expert statisticians with which analysis methods they were familiar and how they chose between these methods. From this study they gained knowledge about the concepts and reasoning that are applied during statistical consultation. In the next study, they examined the similarities and differences the same statisticians perceive among analysis methods. This produced a representation of methods, called a 'method catalog', for each expert. Subsequently, the method catalogs of all experts have been compared. The comparison of the results of the two studies leads to the consultation that the representation of the problem during statistical consultation differs from the representation of the solution in terms of the analysis method to be applied.

This research is important to construct the knowledge base in computer system. The consultation system is the design to simulate the actual human

expert, so, we must learn the structure of knowledge database.

Gebhardt (1990) introduced EXPLORA. This system cannot suggest any statistical method or analysis. The goal of EXPLORA is to analyze large data sets and to extract interesting results displaying them in a textual form. This technique includes the machine learning. We will apply this technique to collect the statistical knowledge in the dialogue between an expert and an analyst.

There are some expert systems to treat the specified application regions. Young Tung (1990) made an expert system. This system is designed to help analysts in the pharmaceutical field validate an assay via model fitting and residual analysis. The tool of this system is the hypertext and computer packages. Darius, Duchateau and Nys (1990) introduced DAEDALUS. This system treats the statistical management of experimental data. The statistical analysis engine of this system is SAS and the inference engine of statistical strategy is TAXSY. The rule bases described by TAXSY are handled SAS datasets. Dorda, Froeschl and Grossmann (1990) introduced WAMASTEX. This system focuses at the statistical needs of clinical physicians working and researching in hospital departments. WAMASTEX is entirely integrated into also the SAS package by that system's macro language facilities. They validate this system by comparison between the judgement of system and experts in some empirical studies.

3. Nature of Statistics

3.1 Process of a Data Analysis

We show the process of data analysis and the technique, as the knowledge from our experience in such process. Also we discuss the possibilities to incorporate them in a system.

In the case of usual data analysis, we perform the following steps (Figure 3.1 shows the flow of these process). It is important to consider the building of an expert system.

1) Data input :

The analysts bring the values of their observation from various fields. They input these data from keyboard and makes some files in computers. Someone orders these tasks to the punch-in service company.

2) Data check :

The raw data includes many errors by the mistake of the observer, the miss-punch, the careless-miss and so on. We check these mistakes using some tools. And if we find it, we go back to the original data sheets and conform it. In this process the frequency table is the powerful tool. It can get the counts of each item, and we can check logical mistakes or some errors. Sometime we find unexpected value as 'outlier'.

Methods : Frequency table, Histogram, Cross tabulation.

3) Simple summary for each variable :

We grasp the whole outline of each variable at the first. We want to know minimum and maximum value, range, mean, median, shape of distribution and so on. These informations tell us the characteristics of the data.

Methods : Basic statistics, Cross tabulation, Frequency table, Histogram, Exploratory Data Analysis (EDA).

4) Relation between two variables :

We grasp the simple relation between two variables. In lower dimension such as two, we can display the data form on graphical outputs. It is easy to catch the characteristics of the data.

Methods : Simple regression line, Scattergram, Scatter plot matrix.

5) Modeling :

We consider the theory in the previous study of this research fields, and build the structure of the data as the statistical model. This step is the cooperated procedure with statisticians and analysts.

6) Data selection :

We select appropriate data for our purpose. We pick up some variables or cases from whole data set to fitting the statistical models.

Methods : Variable selection, Case selection.

7) Analysis :

We decide appropriate statistical analysis according to our purpose and data type. Usually, we calculate this procedure using some statistical program package (SPP), for example SAS, SPSS, BMDP, S or Seto/B.

Methods : Multivariate analysis, Exploratory Data Analysis (EDA), Statistical methods.

8) Interpretation of results :

We interpret the results based on background, and consider the relation between the statistical result and the actual world. If we find any mistrust, we check the stabilities of results or the sensitivity of data.

9) Data transformation :

We check the satisfaction of statistical assumption. If it is violated, we eliminate problems by the transformation or deletion of cases. And after these conversions, we go back to forward step.

Methods : Power transformation, Delete outlier case.

We analyze data in these steps backward and forward mutually. For each step we need operations to give commands to the statistical software. Therefore, we need an expert system for each process.

The knowledge and experience possible to incorporate in system may include decisions based on values of statistics, and advice to users expressed in brief sentences. The decisions are written in knowledge expressions such as 'IF THEN' production rules. But all decisions should not be done automatically

by the system. It is better that the system gives only cautions and information helpful for the user to decide. The reasons are that the decision usually has a relationship with background and that it is dangerous to decide using only the knowledge easy to incorporate. From the above discussion, it is appropriate for the system to indicate the following information.

1) Results of decisions by the system.
2) Information to understand the results.
3) Information to conclude the results.
4) Information to decide in the case that the system cannot make any decision.

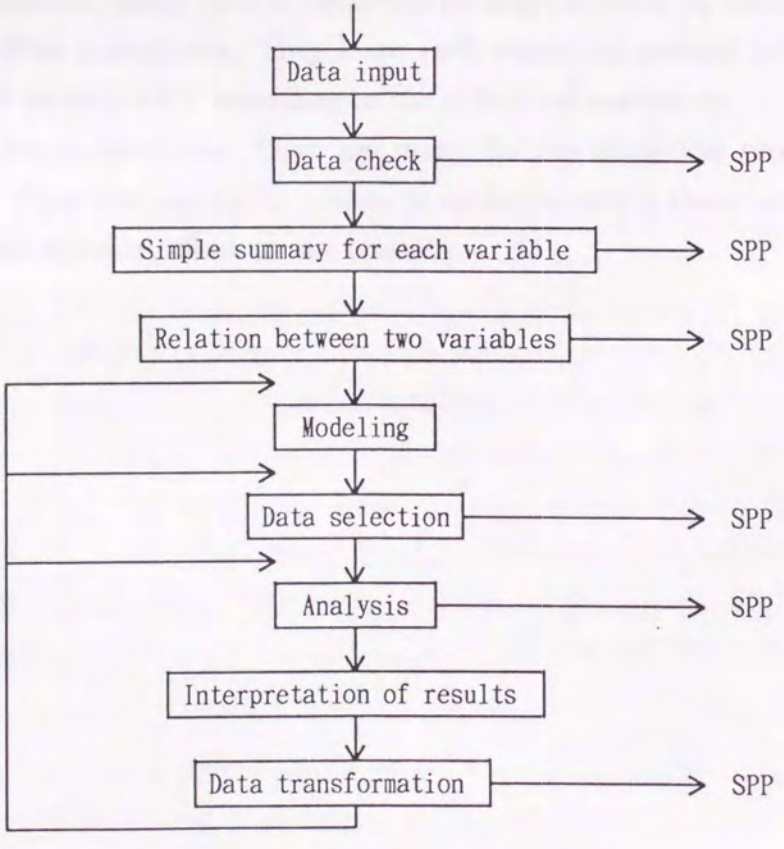

Figure 3.1. Process of a data analysis

3.2 Relation between Statistical Theory and Computer

The computer power is growing rapidly, we can execute many calculations or complicated analysis, for example exploratory data analysis (EDA), interactive dynamic graphics, simulation and so on.

The statisticians develop some statistical theories in their research. Ten years ago, if they want to practice a theory on the computer, they have written it's programs in FORTRAN or BASIC language. It is the area of computer science, another work from statistics, so it is difficult for the pure statisticians.

But now, some statistical software packages have the effective facility called matrix language like SAS/IML or S. The expression of formulas in these systems is similar to mathematics one, not necessary to use DO-loop in FORTRAN. It is easy to program for statisticians, so they can confirm their theories on computer easily. And the analysts can apply the new theory to their data.

At the present, many users of statistical packages demand an expert system for these software programs. They know each statistical method not so well, they cannot distinguish it according to the statistical conditions.

In the above discussion, there are many various statistical programs in the world. Therefore, an expert system is useful to switch these methods or programs like a police officer on the crossing.

4. Knowledge Representation

4.1 Knowledge Representational Models

The research of Artificial Intelligence(AI) aims to simulate the actual human behaviors on the computer. These action bases on the knowledge database. There are roughly two regions of study. One is the store of knowledge and the other is the derivation one from all. The efficiency of knowledge usage depends upon the structure of knowledge representation. The expert system is the product of AI study. It includes the knowledge of human experts, and offers them according to the demands of non-experts.

There are many models for the knowledge representation, and each of them has the intrinsic feature. We introduce the basic concept of some knowledge representational models.

1) Production rule model

This model is composed of 'IF-THEN' rules. The execution based on it is the repetition of the selection and the evaluation of rules. Each action occurs the reference of knowledge database called 'pattern matching'. Therefore, the knowledge increasing brings to the dilation of inference speed. But these rules are easy to describe and to modify them. If the size of system is small, this model is effective. And there are some improved manner for the big system.

2) Frame base model

This model was developed to describe the psychological phenomenon of human remembrance or recognition. The 'frame' is the representative structure of a conceptual objects. One frame includes some 'slot'. Each slot has one information about the frame and points other frame by a link. The lower frame represents more detail information of an object. A set of whole frame with slots has one system of knowledge. Because of the feature of this model is the rich representability and flexibility, this model is widely practical uses.

3) Semantic network model

This model is an effective model for the hierarchical knowledge. It composes with nodes and links, and the former represents for conceptual objects, the latter for relations between nodes. And one whole network has one system of knowledge. We discuss the semantic network model fully in the next section.

4) Logic base model

This model is composed of 'atomic formula' that described a fact between two objects. A set of atomic formula is called 'predicate'. The form of atomic formula is similar to natural language. And this model represents the strict definition and can derive the strict inference.

4.2 Semantic Network Model

In Artificial Intelligence (AI), there are many kinds of studies about acquirement, storing and deriving the knowledge to simulate the human judgment.

We introduce some models for the knowledge representation in the previous chapter. Quillian (1969) developed the semantic network model to apply the understanding means of language, called 'Teachable Language Comprehender'. This study aims to describe the remembrance of long term in the psychological region. He used the network for the relation between word concepts.

The semantic network model is an effective model for the hierarchical knowledge. This model composes with nodes and links, the former for objects, the latter for relations between nodes. The upper node inherits the integrated concept belonging to it. And one whole network has one system of knowledge.

There are some applications based on this model. SCHOLAR (Carbonell (1970)) is a CAI (Computer Assisted Instruction) system about geography of South America. It can converse with computer and student interactively using natural language. CASNET (Weiss, Kulikowski, Amarel and Safir (1978)) is a diagnostic system of glaucoma. They use the causal network with nodes for the state of disease, link for causal of it. TORUS (Rousspoulos and Mylopoulos (1975)) is a database management system of understanding meanings. It links concept and actual event.

If we design a system based on semantic network model, it is righteous to make good use of the advantage of it. We think that the more important point is to grasp the defect, and we build up the system without it. Because each representational model has both merits and demerits.

The semantic network model is simple to understand, and flexible to edit-

ing knowledge. The feature of this model is below (Figure 4.1 shows an example of semantic network).

1) the semantic network composes with nodes and links.
2) one node expresses an object.
3) one link has a semantic label that expresses the relation of both side nodes.
4) one series of links between two nodes represents the relation of two nodes.
5) Advantage
   5a) available for definite relation.
   5b) hierarchical structure.
   5c) easy to understand.
   5d) easy to modify and to add new nodes and links.
6) Defect
   6a) difficult to use in complex problems.
   6b) impossible to denote all relation at the first point.
   6c) much retrieval time in a large network.
   6d) propriety of result of reasoning.

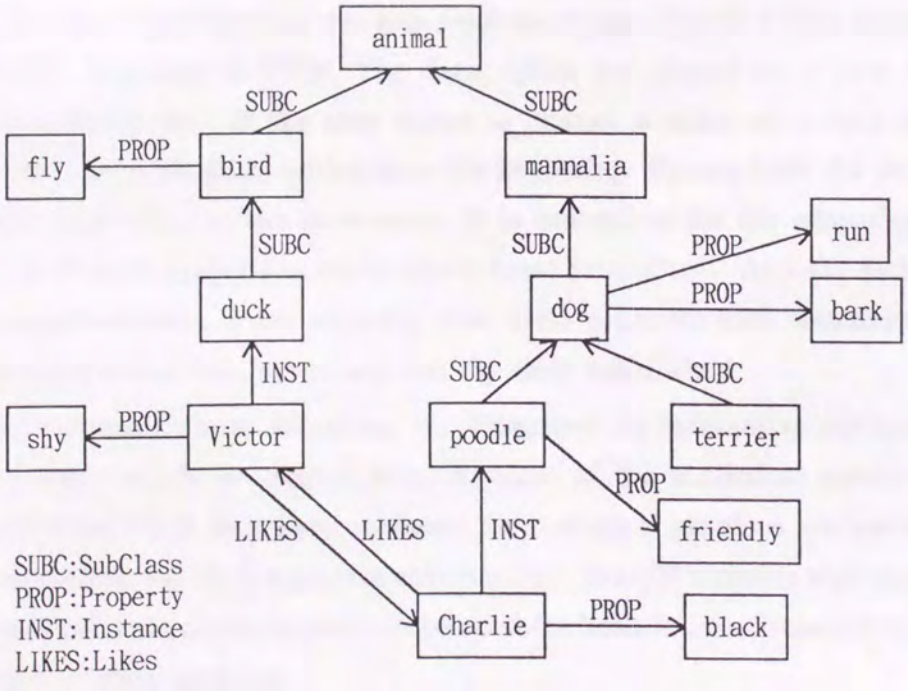

Figure 4.1. A example of semantic network

5. Seto/B — Statistical Analysis Software

We published a series of books about statistical methods titled 'Handbook of Statistical Analysis with Programs for Personal Computers (PHB)'. These books treat wide statistical methods with each BASIC language program.

After these publishing, many readers requested the programs of data file handling version. Therefore, we developed a new statistical software program which we named Seto/B that makes good use of personal computers. It includes more than 50 statistical methods, from introductory to advanced, with graphical outputs and 8 powerful utilities for treating the data file.

5.1 Design of Seto/B

The series of PHB integrates the theories, programs of BASIC language and analysis examples concerning with many statistical analysis methods. The aims of these books are the diffusion and the reference of statistical methods in many applicable fields. The readers can learn not only statistical theories but also numerical experiments using included programs. Sometimes, programmers of statistics refer to them, and researches use to analyze the data sets of their study.

The way of getting data sets to a program is used READ-DATA statement of BASIC language in PHB. The data values are placed on a part of the program statement. If the user wants to change a value or a data set, he modifies it by a program editor from his keyboard. He can look the program and the data value at the same time. It is convenient for the education, but the user of data analysts is worry about these procedures. And the facility of data transformation is not support, then users calculate each transformation for example logarithm, power and etc. by their calculators.

So, we solve above situation, we developed an interactive environment for statistics which we named Seto/B. Most of the statistical methods are brought from PHB, but some methods for statistical graphics are made. We improved them for the interactive environment. Seto/B includes wide methods of statistical data analysis, and can be used for both the confirmatory and the exploratory data analyses.

And this system handles the data files in computers, the analysts can apply

one data file to many statistical methods with no change. We developed some getting routines of data from files newly. They have some facilities of the error recovery, the history of used file name and so on.

The features of Seto/B are as follows:

(1)Menu driven system

There are two methods to control programs. One is the command driven system, and the other is the menu driven system. We intend to progress analyses with interactive style. The former needs its manual in hand. Therefore, we took the menu driven system. The users select one method from menu, can execute their statistical analysis quickly and be productive right from the start interactively.

(2)Portability

Many personal computers involve BASIC language interpreter. Therefore, we can use BASIC easily. All parts of Seto/B is described by Microsoft BASIC. And if users want to use Seto/B on a personal computer, they may modify only the module that depends on its hardware.

(3)Construction with common modules

Seto/B is constructed with many common modules – each module has one facility – such as graphical output, matrices computation, computation of probability distributions, data input/output, etc. Then the design gains more maintenablility.

(4)Extensibility

Seto/B is a set of complete programs. If users have the knowledge of statistics and BASIC language, they can modify and extend methods to Seto/B easily, because of source programs are released for users.

(5)Support of graphic functions

It is very useful for understanding the results of analyses to display by not only numerical values but graphs. Many personal computers have graphic functions on the display device. Seto/B makes good use of this function. Especially, the module of the scatter plot has many display options, so various styles of the scatter plot are available. And the graphic module uses basic commands to draw graphic outputs, then it is easy to modify for the plotting device.

(6)Data management utilities

We often need the functions of data management in analyses process, for example data entry, modify, transformation, selection and so on. Seto/B provides several utilities for treating data file. Adding the system has the some converters from/to system file from application programs (dBASE, MULTIPLAN, Lotus, etc.), users bring in data from these systems.

The development of Seto/B is performed on the personal computers PC-9801 series (NEC). But we use only primitive commands of BASIC language, it is not difficult to implement into other computers.

5.2 System Structure of Seto/B

Seto/B composes of many statistical method programs and a supervisor. Each statistical method program is a complete program. It includes some routines dealing with data input/output, computation of statistics, graphic output and so on. The supervisor controls the analytical process.

In this system, the supervisor is called in the first from the user, and it displays available statistical methods. The second we select one of them, then the supervisor calls its program and the control is turned over to it. The third the called one manages all of the statistical methods. The last the program calls back the supervisor. This procedure is progressed by interactive steps.

5.3 Statistical Methods

More than 50 statistical methods in Seto/B are classified into the following 5 groups. Most of the statistical methods are brought from PHB, but some methods are developed newly. This system includes from introductory to advanced methods with graphical outputs, therefore, Seto/B is applicable to manifold data analysis fields. (Available statistical methods and utilities in Seto/B are shown in table 5.1.)

(1) Simple statistics
(2) Graphical methods
(3) Estimation and testing methods
(4) Non-parametric methods
(5) Multivariate statistical analysis

5.4 Data Management Utilities

We prepared 8 utilities for treating data files, so you can enter and access data easily. The facilities of utilities are as follows.

(1) Data entry and editing
(2) Case and variable merge of system file
(3) Transpose system file
(4) Data transformation and selection
(5) Converter to system file from data statement
(6) Converter to system file from text file
(7) Converter to system file from other system file
(8) Converter from system file to other system file

Seto/B interfaces with standard ASCII text file, dBASE II/III, MULTIPLAN, Lotus 1-2-3, etc., then you can use Seto/B with other application programs mutually. For example, an analyst manages the data of his experiment in a data-base software. If he wants to summarize them by some statistical methods, then he brings the data to Seto/B using above converter, and analyzes some methods. After this, he gets the results not only on printer output but also in text file of computer. He can process it to a report, easily.

In particular, the 'Data transformation and selection' utility has some convenient commands to change values using functional transformation, to select observations and to create new variables from existing ones. These commands like BASIC language. And users can use the special variables called 'system variable' and 'global variable'. These are useful for calculating lags between observations and storing temporary values. Or one of system variable treats the missing values in our system, also users can define missing observations.

These utilities are powerful for the management of statistical data sets, therefore, users carry out various statistical analyses with some variable transformation without other programs.

5.5 Discussion

Seto/B is an interactive statistical software for personal computer including comprehensive statistical methods. This is easy and convenient to use, and designed not only for beginners, but also for senior statisticians. Users need little knowledge of computers. If they have the knowledge of statistics and BASIC, they can modify and extend methods to Seto/B easily, because of source programs are exhibited for users.

And we have been offering the information of Seto/B (new facilities, corrections, bugs, etc.) by news letters and some computer networks (Bulletin Board Systems(BBS)) for the supports after release.

Table 5.1. The functions and contents of Seto/B

(0) Main menu

Seto/B – Statistical software for personal computer –

- Simple satistics
- Graphical methods
- Test and estimation
- Nonparametric test
- Multivariate analysis
- Utilities
- Environment variables set
- MS-DOS Command
- End (return to BASIC system)

Table 5.1. (Continued)

(1) Simple statistics

BSTAT	Basic statistics
BOX	Box and whisker chart
SREG	Scattergram and regression line
RCOR	Spearman's and Kendall's rank correlations
SMOOTH	Smoothing
CROSS	Crosstabulation
CROSS3D	Crosstabulation with 3D_block chart
AGGREG	Aggregation
SCAT	Scatter plot
SCATMAT	Scatter plot matrix
MSCAT	Multiple Scatter plot

(2) Graphical methods

SCAT	Scatter plot
SCATMAT	Scatter plot matrix
MSCAT	Multiple Scatter plot
CROSS3D	Crosstabulation with 3D_block chart
RADAR	Radar chart
ANDRE	Andrews' plot
CONSTE	Constellation graph
MFACE	Face graph
LVECT	Linked vector graph
PQnq	Probability plot (Normal)
PQmul	Probability plot (Multivariate normal)

Table 5.1. (Continued)

(3) Estimation and testing methods

ESTEST	Estimation and testing for mean and variance
TMEAND	Test for equality between two means
TFIT	Test of Fit
TINDEP	Test of Independence in a Contingency Table

(4) Nonparametric test

WILCOX	Wilcoxon Test
MANNWH	Mann-Whitney Test
MEDIAN	Median Test
WAERDE	van der Waerden Test
SIGN	Sign Test
SWILCO	Wilcoxon Signed Rank Test
SPEARM	Spearman's Rank Correlation Test
KENDAL	Kendall's Rank Correlation Test
KRUSK	Kruskal-Wallis Test
JONCKH	Jonckheere Test
K-MULT	k-Multiple Chart Test
FRIEDM	Friedman Test
PAGE	Page Test
CONCOR	Concordance Test

Table 5.1. (Continued)

(5) Multivariate statistical analysis

MREG1	Multiple regression analysis
MREG	Multiple regression analysis with variable selection & regression diagnostics
DISC12	Linear discriminant analysis & quadratic discriminant analysis (Two groups)
DISC34	Linear discriminant analysis with variable selection & canonical discriminant analysis (Several groups)
PCA	Principal component analysis
CANCOR	Canonical correlation analysis
FACTOR	Factor analysis
CLUST	Cluster analysis
ASSOC	Association measures of contingency table
QUANT1	Hayashi's first method of quantification
QUANT2	Hayashi's second method of quantification
QUANT3	Hayashi's third method of quantification
QUANT4	Hayashi's fourth method of quantification
PCOORD	Principal co-ordinates analysis
BIPLOT	Biplot

(6) Utilities

Input/Edit	Data entry and editing
Merge	Case & variable merge of system file
Transpose	Transpose system file
Transformation/Selection	Data transformation and selection
Convert 1	Converter to system file from data statement
Convert 2	Converter to system file from text file
Import	Converter to system file from other system file
Export	Converter from system file to other system file

6. A New Consultation System for Statistical Analysis

6.1 Statistical Strategy Map

When a statistician analyses a dataset, he advances it using his techniques of data analysis. We observe these analysis process, and we find that the statistician traces the analysis techniques based on the experience in his mind. There are composed of many conditions and methods about statistics mutually. He judged each conditions or states of the dataset, and reached to select one appropriate statistical method finally. These conditions and methods are placed like stems and leaves of a tree in his knowledge. The 'turning-point' of tree is the condition and the 'leaf' is the statistical method.

From these discussions, we think of an idea that we can management the statistical knowledge on the hierarchical structure. We collect the relation between conditions and methods about statistical analysis in our experience, and place this knowledge on the hierarchical structure as the statistical knowledge base.

To create the knowledge base with respect to the selection of a statistical method, we investigated relations between purposes of analyses, type of data and methods of statistical analyses from the viewpoint of Table 6.1.

Table 6.1. Purposes of analyses and type of data

Purposes of analysis :

- Estimation / Test
- Graphical representation
- Prediction
- Explanation
- Creating an index value
- Visualizing a relation between items
- Classification
- Latent structures

Type of data :

- Continuous variable
- Discrete variable (Item category)
- One variable / Two variables / Many variables

There are some classifications of statistical analysis methods. We classify many methods into the purposes of analysis at the first point. The reason using this classification is the same process of statistical consultation in our experience. The statisticians derive from the clients that the background of experiments, the purpose of analysis, the using method of previous researches and so on. Many statisticians grasp the methods of analysis by the purpose, then they ask 'What do you want to do on your dataset ?' in the first question. The following questions include the type of data and the field of research. Therefore, this classification is the very natural both for statisticians and for beginners.

The feature of this classification is that the height of the created tree is low and the tree has wide stems. This means that the tree is composed of many short stems. The condition of methods is placed on each edge of stems and users judge it each time. So the short stems are the fewer number of judgements to reach the appropriate method. If the number of turning-points is same, the tree with the wide and short stems is more suitable for the searching time than the thin and long one.

The knowledge of statistics compose many definite relations, for example normality assumption, type of data assumption and so on. And each component has connections with a hierarchical structure mutually.

We constructed the knowledge base like a tree structure named 'statistical strategy map' (Figure 6.1 is a part of it.). The user starts the top of node and he traces one of items or conditions in the map. Most statistical methods have some assumptions, so we recommend one method with some programs about checking assumptions. Finally, he gets one appropriate statistical method with some checking items.

We require the flexibility of the statistical strategy map. Because, the statistical technique is growing rapidly. If an improved method appears, we change quickly the strategy map of the related segment. Another reason of flexibility is that statisticians has each process of statistical analysis on one same data. Therefore, one of important points to construct a statistical strategy is easy to modify.

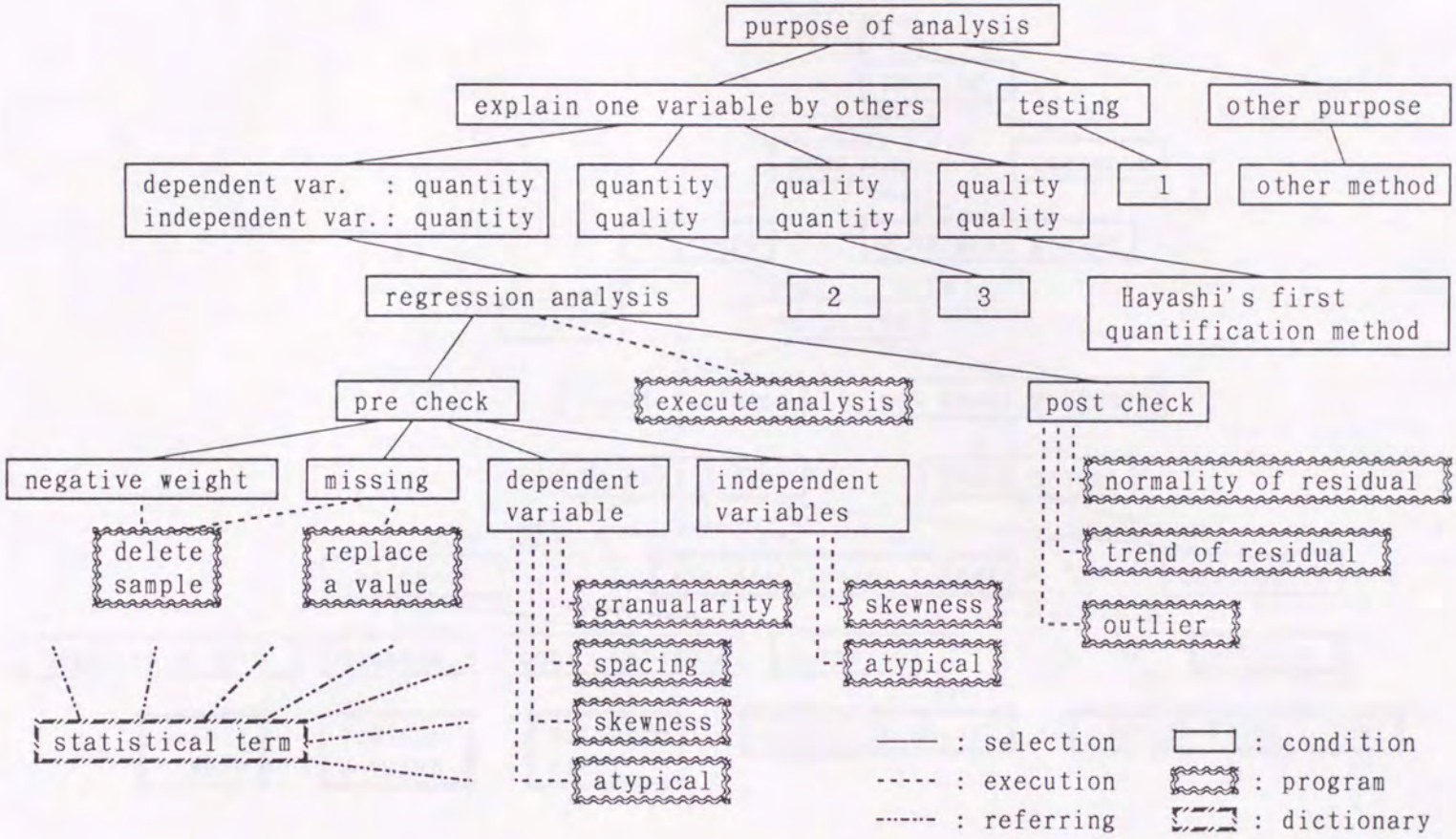

Figure 6.1. A part of the statistical strategy map

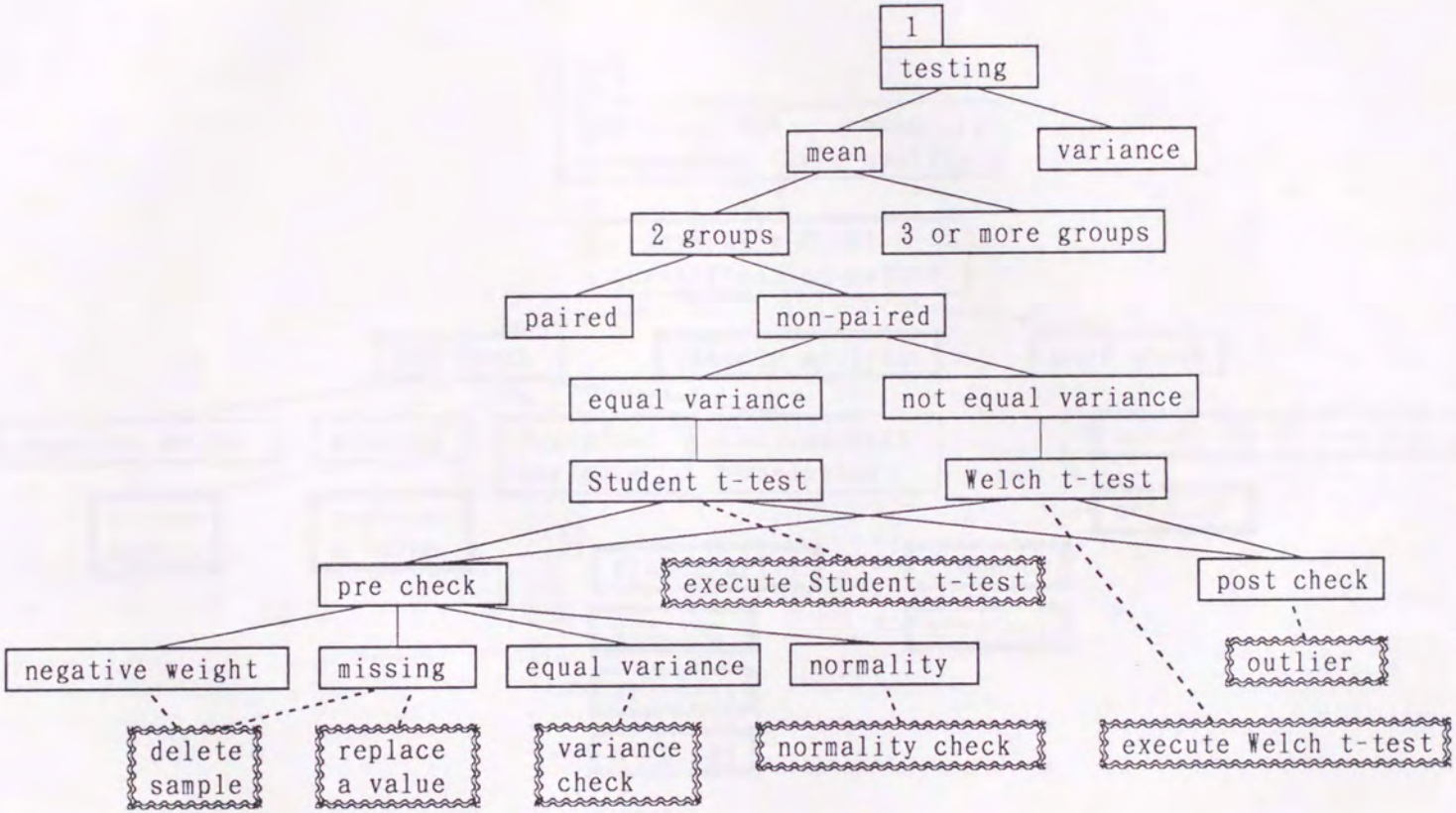

Figure 6.1. (Continued)

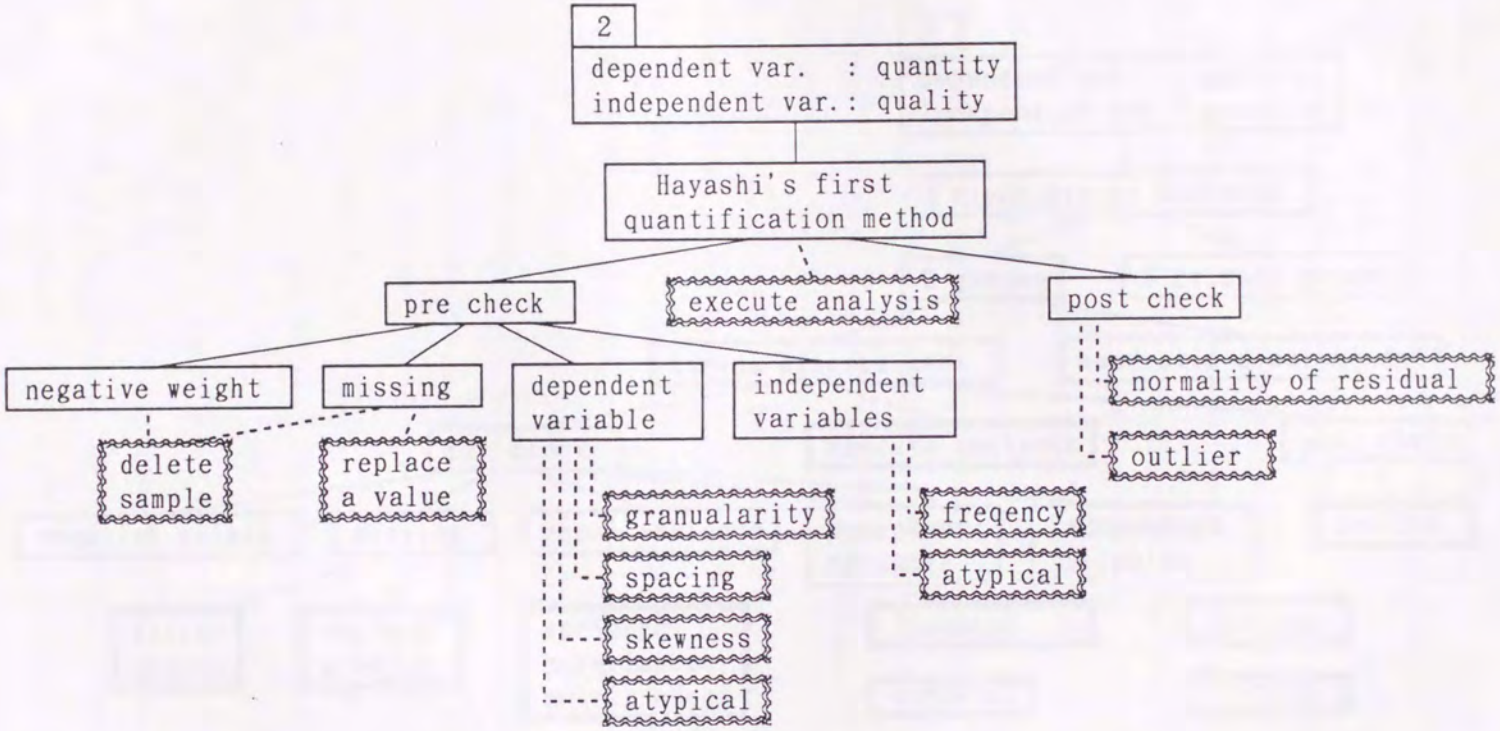

Figure 6.1. (Continued)

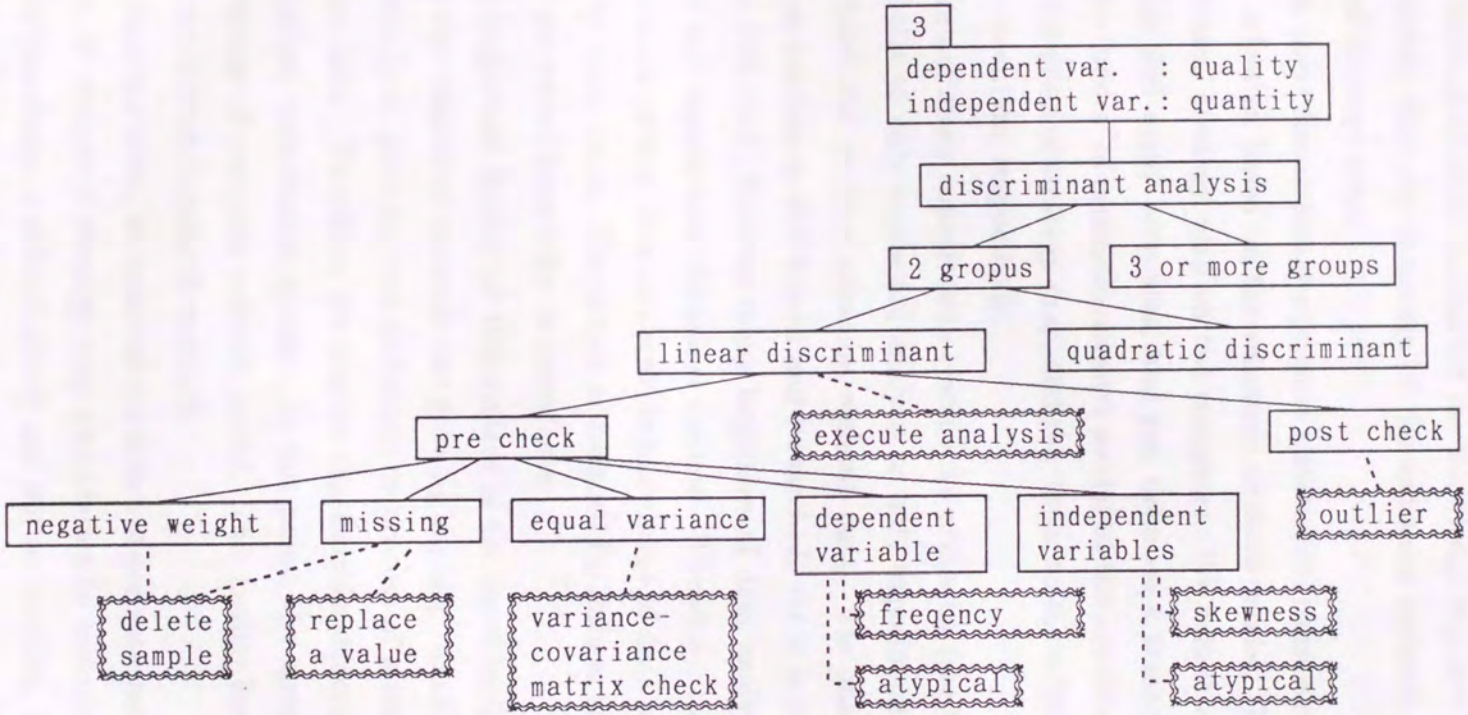

Figure 6.1. (Continued)

6.2 Apply to a Statistical Consultation System based on Semantic Network Model

We introduced some models for the knowledge representation in Chapter 4, and notice that the structure of the semantic network is similar to the statistical strategy map.

From these discussions, we propose a statistical consultation system of the method selection based on the semantic network model. This work realizes the statistical strategy map on the computer. We collect a lot of statistical knowledge and many programs, and put them on a semantic network. We assign the 'nodes' of semantic network model to the condition, the dictionary and the program of strategy map, and the 'links' to the selection, the referring and the execution, respectively.

There are many consultation systems for statistics (see Chapter 2). Some of them aim to only regression analysis or only experimental data. Nothing treats about the method selection without ours. The main reason is that regression analysis is well known and researched. Or it is most frequently to apply to real field. However many beginners of data analysts cannot decide or select one appropriate statistical method. Therefore, the development of a consultation system that can provide knowledge of data analyses is eagerly desired by many users. The system is designed for data analysts and students without statistical knowledge or experience.

The important facility of this system is the modifiability of knowledge. Because the statistical methods are growing rapidly. Another reason is that each statistician have his own technique or process of statistical analysis on one same data. Therefore, we require the flexible structure of strategy in the statistical consultation system. In this point, the strategy map inherits the advantage of semantic network model. This facility keeps the system to include the newest statistical methods.

On the other hand, we removed the defect of semantic network model. The structure of statistical strategy map contributes to exclude the defect. The statistical knowledge is defined clearly and not so complex. The retrieval time depends on the length of links, and our tree structure branches the wide stem, not so tall. Because we separate the purpose of analysis at the nearest point

of the root. It means that many methods are classified into many categories at the first point. Using this classification, the strategy map expands widely, and the retrieval time of the system is short.

The other unique point is that our system can apply to various fields, because it can manage wide and deep range of statistical methods. The reason that these general systems are not exist is necessary to build using the resource of not only knowledge but also programs of statistics. We have many opportunities of data analysis and developed Seto/B. Therefore, we are suitable to design this consultation system.

The data analysts answers the question from the system, traces one path of strategy map and finally reaches one appropriate statistical method with some checking items. The system derives a statistical method to fit the user's purpose by means of a dialogue between the user and this system. He can advance an analysis safely in statistics.

7. Implementation and Example

7.1 Implementation

There are many tools to treat the semantic network. One of these software tools called 'hypertool' refers text files and programs mutually with links.

We developed a consultation system for selecting statistical methods. This system linked the knowledge and the programs of statistics on hypertool, named 'SCSH(Statistical Consultation System based on Hypertool)'. We used 'HyperLink' that is one of hypertool on MS-DOS distributed by MaxThink Inc. with many support utilities.

We constructed SCSH with following steps.

1) Collecting statistical knowledge

   We survey statistical methods, assumptions and conditions.

2) Making strategy map

   We take relations among methods and assumptions as the statistical strategy map.

3) Making text files of each knowledge as text file

   We write the documentation of knowledge, branching condition, presentations, and so on.

4) Making statistical programs

   We developed a statistical software – Seto/B, so we have many statistical programs and techniques about these software programs. We improve these resources to the system.

5) Construction of consultation system on HyperLink

   We link among documents and programs above the strategy map.

After these process, we construct SCSH. Only step 4 may be a little difficulty for the non-programmer, but other steps are a good process for the integration of the statistical knowledge.

If a user finds an additional knowledge, he can add it to the system following above process. Then the new knowledge is available and he can improve our system. Unfortunately, you cannot implement your knowledge, you contact us to request our assist.

The system will ask to users a series of questions, starting with 'What purpose of analysis do you have ?'. Responses to question lead to the partic-

ular technique with some checking problems. Most of questions are phrased for yes/no responses or selecting one item from them. Finally the system recommends an appropriate analysis method or recommends a next process. At present, the system has more than 60 knowledge and about 30 methods of statistics.

Other feature of HyperLink is the style of distribution freely with some limitation. We buy HyperLink and make an application on HyperLink. In this case, we may distribute our application with HyperLink. But received user must not make another application on this HyperLink, he can use HyperLink as only engine of our application. These terms are very useful for our purpose. Because it is impossible to collect all knowledge at the start point, so we should get many users of our system in the various applied fields. We distribute SCSH to many researchers and hope some feedbacks of statistical knowledge.

7.2 Usage

The system is easy to use. Because some questions from the system are required file name or variable name as key-input, but most of questions are yes/no responses or selecting one item among them. These sequential questions and answers lead to the particular technique with some checking problems.

SCSH uses some environment variables of MS-DOS. The user sets these variables to a value of his system. The reason of this setting is that we predetermined our system to the multiple drive and directory system. The volume of all files in statistical knowledge, programs and dictionary is very large, so the combined management makes troubles. We separate each group in each drive or directory. Table 7.1 shows them and our values for example.

Table 7.1. Environment variables of SCSH

STATAIPROG=a:\	Drive and path of statistical programs. We place these program in root directory of drive A:.
DICTIONARY=\stat.dic	Drive and path of help dictionary. It explains technical terms and methods. Many users are set 'stat.dic' (we prepared),

	if you use personal dictionary, please set it.
EDITOR=e:\mi	Drive and path of Editor.
	Set your favorite editor.
	We set Mifes(MEGASOFT Corp.) in drive E:.
N88BAS=h:\	Drive and path of N88BASIC interpreter(NEC).
	Some programs require this interpreter.

After environment setting, You change SCSH directory by 'cd(chdir)' command and hit 'SCSH' to start our system. Also, we prepare a sample batch file to set environments and execute our system.

7.3 Example

We illustrate an example using 'Cost of construct of nuclear power plants' data from Cox and Snell(1981, page 81). The user has the data set and the purpose of analysis *i.e.*, 'Cost(C) explained by Date construction permit issued(D), Power plant net capacity(S) and Partial Turnkey plant(PT)'.

At the first display, we show the usage, getting file name and entrance the system(Fig.7.1). Users can move a cursor at only '⟨ ⟩'mark, and select one of these places. Next the system asks the user's purpose of the analysis(Fig.7.2). There are general purposes (8 items at present) based on wide methods from testing one variable to multivariate analysis. In this example user selects first item. For this purpose, we prepare 4 combinations of data type(Fig.7.3). In this case all of variables (independent and dependent) are quantity variables, so, select first item again.

Then the system recommends multiple regression analysis as an appropriate method with some checking problems(Fig.7.4). We think that the one group of a statistical method and some checkings leads to the right analysis based on statistics. So, before the execution of regression analysis, we check the skewness of variables. It is one of pre-checkings(Fig.7.5). The variable C skews to left, we recommend to transform by logarithmic function(Fig.7.6). If user wants to apply this transformation, the system transforms the data and assigns a new variable named Log_C (user named). After these pre-checkings, we executed regression analysis(Fig.7.7). The system outputs not only regression equation but also some significant tests. Many users finished at this point

without post-checkings of assumption. But we hope to take these process, so our system prepares these items, for example, normality of residual(Fig.7.8) and outlier(Fig.7.9).

This example shows that our system consults not only a statistical method (multiple regression analysis) but also pre/post-checkings (skewness, normality, outlier, etc.).

Other feature of the system is that users can execute statistical methods only with no checking. Because many checkings are troublesome operations for statisticians, but beginners need many advices and checkings. So, users select with or without checking adjusting the level of his statistical knowledge.

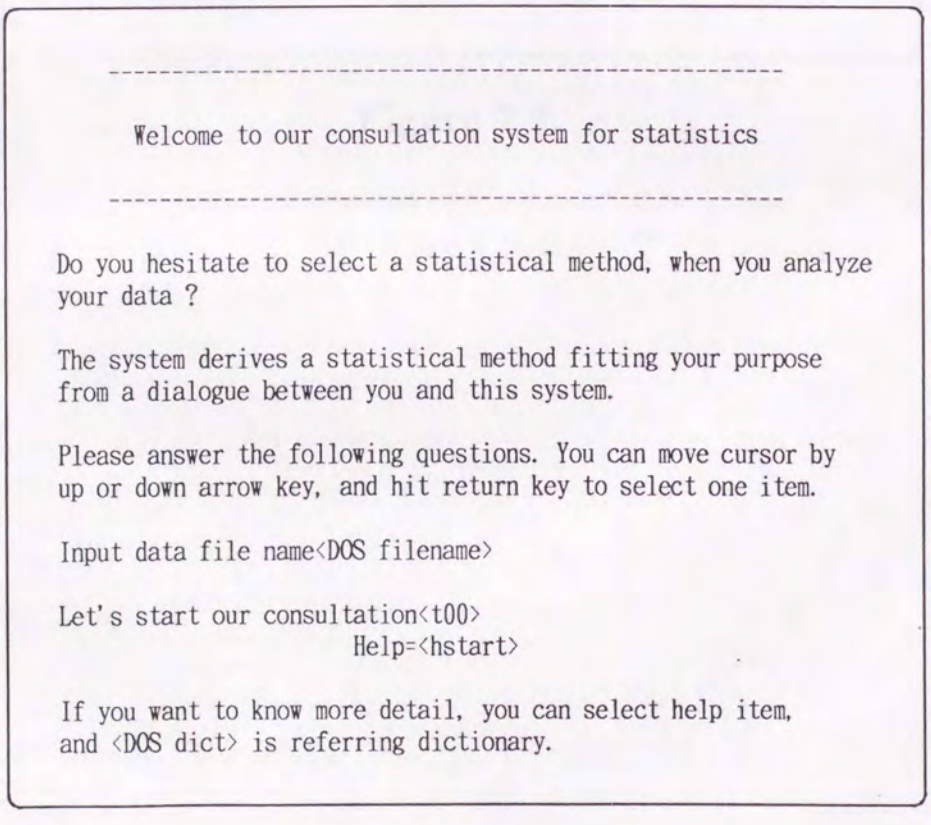

Figure 7.1.

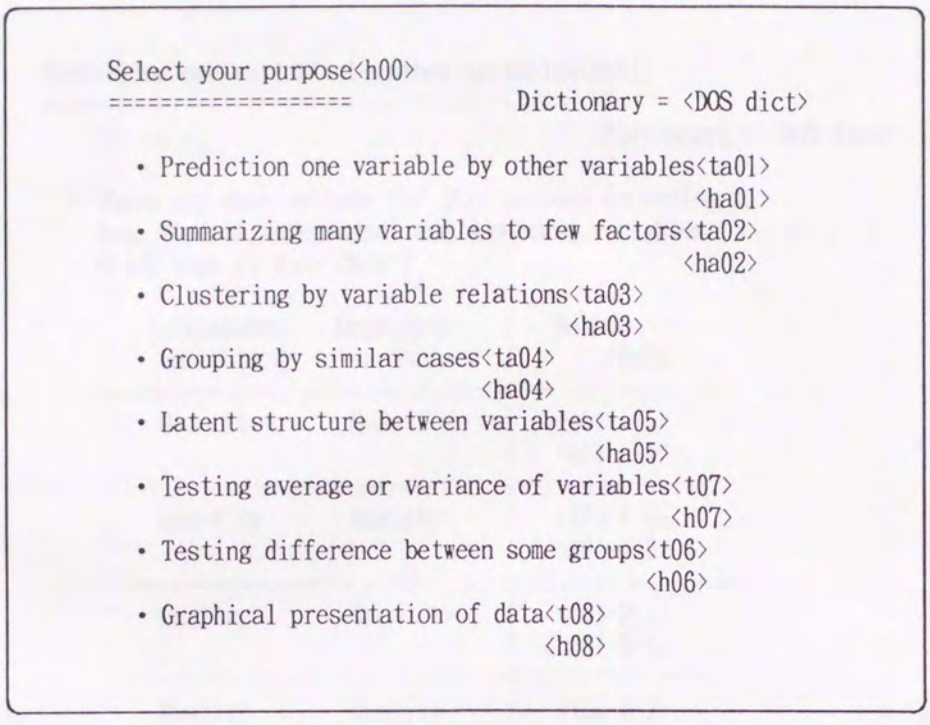
Select your purpose<h00>
Dictionary = <DOS dict>
• Prediction one variable by other variables<ta01>
<ha01>
• Summarizing many variables to few factors<ta02>
<ha02>
• Clustering by variable relations<ta03>
<ha03>
• Grouping by similar cases<ta04>
<ha04>
• Latent structure between variables<ta05>
<ha05>
• Testing average or variance of variables<t07>
<h07>
• Testing difference between some groups<t06>
<h06>
• Graphical presentation of data<t08>
<h08>

Figure 7.2.

Prediction one variable by other variables<ha01>

Dictionary = <DOS dict>

- There are some methods for this purpose according to data type of independent and dependent variables
- Which type is your data ?

Independent variable	Dependent variable	Method /Help
Quantity	Quantity	<t01-1-1> <h01-1-1>
Quantity	Quality	<t01-1-2> <h01-1-2>
Quality	Quantity	<t01-2-1> <h01-2-1>
Quality	Quality	<t01-2-2> <h01-2-2>

Figure 7.3.

```
Appropriate method
==================

    We recommend you to use multiple regression analysis.

    But you must check the following assumptions for this analysis.

     • Pre checkings<tregl-1>
                    <hregl-1>

     • Execution of regression analysis<DOS mreg>
                                       <hregl-2>

     • Post checkings<tregl-3>
                     <hregl-3>
```

Figure 7.4.

```
Checkings before regression analysis<hregl-1>
==================================              Dictionary = <DOS dict>

  You must check the following conditions before regression analysis.

    • Negative weight<DOS nweight>
                     <hnweight>
    • Missing values<DOS miss>
                    <hmiss>
    • Checking list of the independent variable<tregl-1-.3>
                                               <hregl-1-.3>
    • Checking list of the dependent variable<tregl-1-.4>
                                             <hregl-1-.4>
    • Multi-collinearity<DOS multico>
                        <hmultico>
```

Figure 7.5.

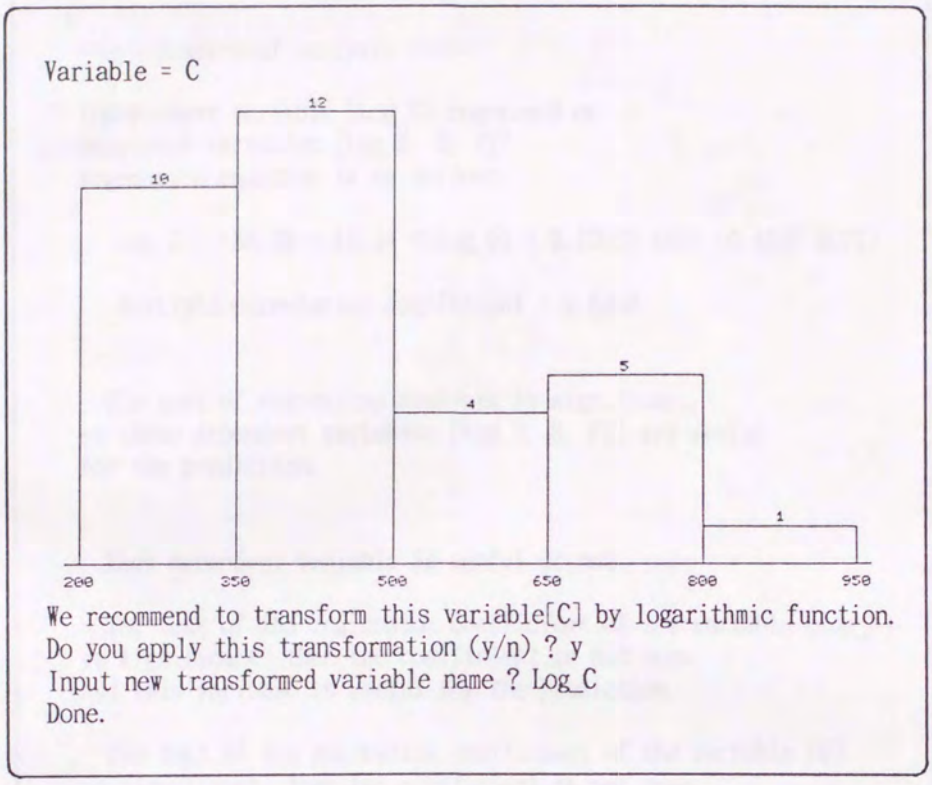
Variable = C
12
10
5
4
1
200
350
500
650
800
950
We recommend to transform this variable[C] by logarithmic function.
Do you apply this transformation (y/n) ? y
Input new transformed variable name ? Log_C
Done.

Figure 7.6.

```
===== Summary of analysis =====

Independent variable [Log_C] regressed on
dependent variables [Log_D, S, PT].
Regression equation is as follows.

   Log_C = -38.05 + 10.34 *(Log_D) + 0.00063 *(S) -0.4608 *(PT)

   Multiple correlation coefficient = 0.8268


  The test of regression analysis is significant,
so these dependent variables [Log_D, S, PT] are useful
for the prediction.


  Each dependent variable is useful or not.

  The test of the regression coefficient of the variable [log_D]
is significant, then the coefficient is not zero
and this variable is useful for the prediction.

  The test of the regression coefficient of the variable [S]
is significant, then the coefficient is not zero
and this variable is useful for the prediction.

  The test of the regression coefficient of the variable [PT]
is significant, then the coefficient is not zero
and this variable is useful for the prediction.
```

Figure 7.7.

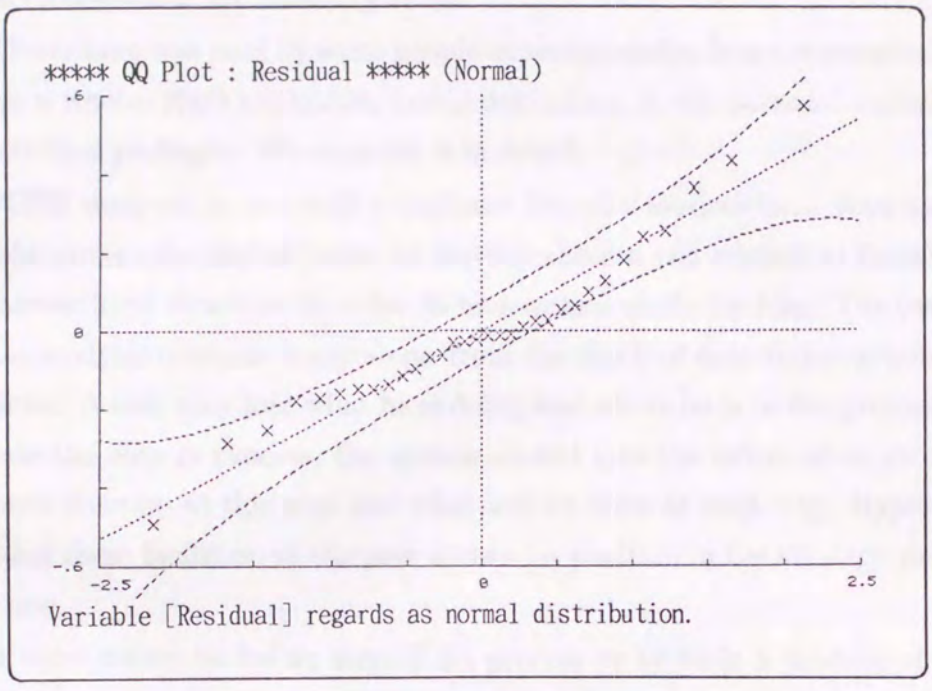

Figure 7.8.

```
Checking influential observations<hinflue>
================================          Dictionary = <DOS dict>

Leverage (Threshold value = .25)
  Following case number(s) are influential observations.
                                              (value of Leverage)
    26 ( .338494 )

DFFITS (Threshold value = .707107)
  Following case number(s) are influential observations.
                                              (value of DFFITS)
    17 ( 1.73923 )

Cook's distance (Threshold value = .807517)
  Influential observation could not find.

There are some influential observations.
Sometime these cases bring other result.
Please check these cases in the original data set
and pay attention to the result.
```

Figure 7.9.

8. Evaluation of the System

The system was used by some people experimentally. It is a convenient system as a whole. Each evaluation varies depending on the personal experience in statistical packages. We examine it in detail.

SCSH suggests to not only a beginner but also statisticians. And that it provides some selection of items for statistical tests and statistical knowledge in a hierarchical structure in order to understand easily for him. The process of data analysis contains many steps, from the check of data to the evaluation of results. A user may lose what he is doing and where he is in the process. To indicate the step in process, the system should give the information on what has been done up to this step and what will be done at next step. HyperLink provided these facilities, so the user knows his position in the strategy map at any time.

If users return to before step of his process or he finds a mistake of item selection, he can display his analysis path, and move back to a position by cursor quickly.

This system is driven by the user, so he can bring to incorrect results by the mistake of item selection. We hope that users have not the spite. Someone thinks that it is a weak point of this system, but nobody makes a mistake intentionally and none of them gets benefit. You don't worry about it.

Using HyperLink, our system has the flexibility of knowledge. It is important for the consultation system. Especially statistical technique is growing rapidly, or statisticians have each process of statistical analysis on one same data. So, one of important points to construct a statistical strategy is easy to modify. Our system realizes this nature.

And each knowledge was management some text files in the system. It is a simple structure of computer files. So, anyone makes new text files by an editor, and add to the knowledge easily.

This system consults a process and statistical methods with some suggestion, but if user wants to execute a method directly without pre and post testing. Because each checking is muggy for the well-known user and some user who views a result anyway conditions. So, this bypass process is provided for experts (statisticians).

As a related study there are some expert systems in the world. The first is REX (Regression EXpert) discussed in Gale(1986). It guides the analysis process by testing assumptions of regression, suggesting possible transformation when assumptions are violated, and justifying of regressions when requested.

The second is S/EXP by Ishibashi and Takeda(1990). The system incorporates knowledge and experience of statistics, and is constructed on S system (Becker and Chambers(1984)). Users can see parameters, numerical results and related graphs at a time through multiple windows on screen.

The last is RASS(Regression Analysis Supporting System) by Nakano, Yamamoto and Okada(1991). This system uses object oriented technique programmed in Plorog language. The statistical knowledge in the system stores a hierarchical structure of knowledge class object.

The aims of SCSH are quite similar to ones of these expert systems, but there are some differences among them.

The major differences are that these 3 systems treated only regression analysis. The main reason is that regression analysis is well known and researched. And it is most frequently to apply to real field.

The second is flexibility with regard to the addition, modification and improvement of knowledge using 'hypertool'. These 3 systems are closed systems of knowledge, and difficult to modify it. There are included statistical knowledge in their system. We manage statistical knowledge, method program and inference engine, respectively. Then we reform each module. This facility keeps the system to include the newest statistical methods.

Of course, the reverse of this flexibility exists. We suppose that the statisticians construct the knowledge base. Originally, a consultation system simulates human judgment based on storing information. The system cannot check the reliability of the knowledge. If an incorrect knowledge includes in the knowledge base, it is possible to lead to some wrong suggestions. Also in our system, the ignorance of statistics can modify the knowledge base, however it is dangerous to intrude the false knowledge into it. Therefore, the modification of knowledge base should be restricted to some researchers of statistics.

And the third is environment of computers. You seem to that it is not so important, but it is influence to learning the knowledge. Many data analysts

use the small size computer like a personal computer not a main frame. Above mentioned systems run on work station and our system on personal computer. So, SCSH gets more large size of users than other systems. The User requests various technique in the various field, so many users lead much statistical knowledge. We can get many kinds of knowledge by many users.

There are some points to improve the system. We wish to develop a system that covers as many statistical procedures as possible. At least it is impossible to collect all knowledge at the start point. We will grow up knowledge in our experience together. In the sense the present system is at an intermediate point of development.

After incorporating the statistical term dictionary and knowledge of the system, we plan to support 'Why-facility'. At present, our system suggests statistical methods and checking assumption. But users cannot ask the reason of suggestions. We prepare the online help of statistical methods for this purpose. However, it is not solve his question perfectly. Then we schedule to make up this facility.

9. Conclusion

Many kinds of statistical programs are available, but most of them require the knowledge and experience of statistics. In such a circumstance a beginner of statistical analysis usually meets troubles when he starts an analysis without any advisors.

Over against these situations, we have realized a statistical consultation system based on hypertool named SCSH. This system assists users to select the statistical method with some additional informations or checking items about suggested analysis. It incorporates statistical knowledge and programs through our experience.

To prepare this system, we proposed the statistical strategy map that is a tree structure with a lot of statistical knowledge and many programs. The user starts the top of node, traces one of items in each node, and finally he reaches one appropriate statistical method with some checking items. On the other hand, we notice that the structure of the semantic network model is similar to the statistical strategy map. This is a knowledge representational model researched in the Artificial Intelligence (AI). It is effective for the hierarchical structure and easy to modify.

We combine these products, and implement strategy map on the Hyper-Link. It is a computer software tool of the realizing semantic network. SCSH refers the knowledge and the programs of statistics mutually with links. The user answers the question from the system sequentially, and traces one path of strategy map. Most of all statistical methods have some assumptions, so SCSH recommends one method with some programs about checking assumptions. He can advance an analysis safely in statistics.

Other feature of SCSH is that users can execute statistical methods only with no checking. Because many checkings are troublesome operations for statisticians, but beginners need many advices and checkings. So, users select with or without checking adjusting the level of his statistical knowledge. We prepare these two processes of analysis, SCSH is available not only for beginners but also for statisticians.

The most important feature of this system is extensibility and has the advantage of adding/modifying statistical knowledge. Because, the statistical

technique is growing rapidly. Another reason is that statisticians has each process of statistical analysis on one same data. These characteristics inherit from the semantic network model.

We made SCSH on our experimental basis. In general, the consultation system cannot have all knowledge at the start point. So, we wish to improve the system using actual datasets or experiments. In the sense the present system is at an intermediate point of development. We always grow up the strategy map and feed back to the users.

One method of the collecting knowledge is the taking dialogue between SCSH and users. We derive unimplemented or new statistical knowledge from them, and improve strategy map. In these considerations, we will release this system to many users, and we want to have many opportunities of getting knowledge. The contract of HyperLink that the distribution freely with some limitation contributes to our research.

After improvement of the statistical term dictionary and knowledge of the system, we plan to support 'Why-facility'. At present, this system suggests statistical methods and checking assumption. But users cannot ask the reason of suggestions. We prepare the online help of statistical methods for this purpose. However, it is not solve his question perfectly. Then we schedule to make up this facility.

Acknowledgements

I am deeply grateful to Professor T. Tarumi, Okayama University, for his hearty guidance and continuous encouragement in various ways. I also wish to express my gratitude to Professor K. Wakimoto and Professor Y. Tanaka, Okayama University, for their comments and many valuable suggestions. Furthermore, I bow my sincere thanks to all members of Okayama Statisticians Group for their helpful suggestions. I would like to thank the staff in Kawasaki Medical School for the assistance of the research environments.

References

[01] Afifi,A.A., and Clark,V.(1990), Computer-aided multivariate analysis 2nd ed., Van Nostrand Reinhold Company.

[02] Becker,R.A.(1984), S An Interactive Environment for Data Analysis and Graphics, Wadsworth.

[03] Becker,R.A., and Chambers,J.M.(1984), The S System: Language: A Programming Environment for Data Analysis and Graphics, Pacific Grove, Calif.: Wadsworth & Brooks/Cole., (渋谷政昭、柴田里程訳 (1988)、Sシステム、共立出版)。

[04] Becker,R.A., Chambers,J.M., and Wilks,A.R.(1988), The New S Language: A Programming Environment for Data Analysis and Graphics, Pacific Grove, Calif.: Wadsworth & Brooks/Cole., (渋谷政昭、柴田里程訳 (1991)、S言語、共立出版)。

[05] Blum,R.L.(1982), Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical data base: the RX project, Computers and Biomedical Research 15, PP 164-187.

[06] Brooking,A.G.(1986), The Analysis Phase in Development of Knowledge-Based Systems, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 321-334.

[07] Chatterjee,S., and Price,B.(1977), Regression Analysis by Example, John Wiley & Sons.

[08] Christopher,F.C.(1989), Artificial Intelligence and Turbo C, Multiscience Press, Inc., (岩谷宏訳 (1990)、Ｔｕｒｂｏ　Ｃによる人工知能、工学社)。

[09] Cox,D.R., and Snell,E.J.(1981), Applied Statistics : Principles and Examples, Chapman and Hall.

[10] Daniel,C., and Wood,F.S.(1980), Fitting Equations to Data: Computer Analysis of Multifactor Data (Second Edition), New York: John Wiley & Sons.

[11] Darius,P.L., Duchateau,L., and Nys,M.(1990), A Knowledge-Based Environment for the Statistical Management of Experimental Data, *COMPSTAT '90*, Short Communications, Physica-Verlag. PP 63-64.

[12] Dorda,W., Froeschl,K.A., and Grossmann,W.(1990), WAMASTEX – Heuristic Guidance for Statistical Analysis, *COMPSTAT '90*, Physica-Verlag. PP 93-98.

[13] DuMouchel,W.(1987), Experience with a statistics advisor for industrial scientists and engineers, Bulletin of the International Statistical Institute, Proceedings of the 46th session, Vol, LII, Book 4, PP 369-386.

[14] Frank,M.A., Laura K., Terrence N.D., Patrick M.O., and Willard L.R. (1981), A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, Second Edition, Institute for Social Research, The University of Michigan.

[15] Gale,W.A., and Pregibon,D.(1982), An expert system for regression analysis, Proceedings of the 14th Symposium on the Interface, Springer-Verlag, PP 110-117.

[16] Gale,W.A., and Pregibon,D.(1984), Constructing an expert system for data analysis by working examples, *COMPSTAT '84*, Physica-Verlag, PP 227-236.

[17] Gale,W.A.(1986), REX Review, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 173-227.

[18] Gale,W.A.(1986), Student Phase 1 – A Report on Work in Progress, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 239-265.

[19] Gebhardt,F.(1990), An Expert System Strategy for Selecting Interesting Results, *COMPSTAT '90*, Physica-Verlag, PP 81-85.

[20] Hand,D.J.(1986), Patterns in Statistical Strategy, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 355-387.

[21] Hayashi,A., Wakimoto,K., Tanaka,Y., and Tarumi,T.(1986), Some Remarks for the handbook of Statistical Data Analysis, Abstracts of The Fourth Korea and Japan Joint Conference of Statistics, PP 115-119.

[22] Hayashi,A., and Tarumi,T.(1988), Seto/B — Statistical software for personal computers, The proceeding of the fifth Japan and Korea Joint Conference of Statistics, PP 9-11.

[23] Hayashi,A., and Tarumi,T.(1988), Star/B — Statistical software for personal computers, *COMPSTAT '88*, Short Communications, Physica-Verlag Heidelberg, PP 105-106.

[24] Hayashi,A., and Tarumi,T.(1992), A Consultation System for Statistical Analysis on Hypertool, *COMPSTAT '92*, Physica-Verlag Heidelberg, (in press).

[25] Hayes-Roth,F., Waterman,D.A., and Lenat,D.B.(eds.) (1983), Building Expert Systems, Reading, Mass. :Addison-Wesley., (AIUEO 訳 (1985) 、エキスパート・システム、産業図書)。

[26] Ishibashi,Y., and Takeda, K.(1990), An expert system for data analysis S/EXP and its effectiveness, Journal of the Japanese Society of Computational Statistics, Vol.3, Number 1, PP 61-67.

[27] Levine,R.I., Drang,D.E., and Edelson B.(1986), A Comprehensive Guide to AI and EXPERT SYSTEM, McGraw-Hill Inc., (越田一郎、中川裕志、森辰則訳 (1988) 、ＡＩとエキスパートシステム、マグロウヒル)。

[28] Nelder,J.A.(1991), GLIMPSE: A Knowledge-Based Front End for GLIM, In Buja,A. and Tukey,P.A.(ed.), Computing and Graphics in Statistics, Springer-Verlag, PP 125-131.

[29] Oldford,R.W., and Peters,S.C.(1988), DINDE: Towards more sophisticated software environments for statistics, SIAM J. Sci. Stat. Comput. 9, PP 191-221.

[30] Oldford,R.W., and Peters,S.C.(1986), Implementation and Study of Statistical Strategy, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 335-353.

[31] Pregibon,D.(1986), A DIY Guide to Statistical Strategy, In Gale,W.A. (ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 389-399.

[32] Robert,S.(1987), The Statistical Consultant, A Shareware in PC-SIG, No.949.

[33] Snell,E.J.(1987), Applied Statistics : A Handbook of BMDP Analysis, Chapman and Hall.

[34] Streitberg,B.(1988), On the nonexistence of expert systems : Critical remarks on artificial intelligence in statistics (with discussions), Statistical Software Newsletter 14, PP 55-74.

[35] Tarumi,T., and Hayashi,A.(1989), Consultation system to select a method in multivariate analysis, Bulletin of the International Statistical Institute, Contributed Papers, Book 2, PP 381-382.

[36] Thisted,R.A.(1986), Representing Statistical Knowledge for Expert Data Analysis Systems, In Gale,W.A.(ed.), Artificial Intelligence and Statistics, Addison-Wesley, Reading Mass., PP 267-284.

[37] Tierney,L.(1990), LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics, New York: John Wiley & Sons.

[38] Van den Berg, G.M., and Visser, R.A.(1990), Knowledge Modelling for Statistical Consultation System; Two Empirical Studies, *COMPSTAT '90*, Physica-Verlag. PP 75-85.

[39] Wittkowski,K.M.(1990), Statistical Knowledge-Based Systems – Critical Remarks and Requirements for Approval, *COMPSTAT '90*, Physica-Verlag, PP 49-56.

[40] Wolstenholme,D.E., and Nelder,J.A.(1986), A front end for GLIM, In Haux,R.(ed.), Expert Systems in Statistics, Stuttgart: Gustav Fisher, PP 155-177.

[41] Young Tung,S.T.(1990), An Expert System for Validation Procedures, *COMPSTAT '90*, Short Communications, Physica-Verlag. PP 93-98.

[42] 上野春樹、石塚満 (1987)、知識の表現と利用、オーム社。

[43] 宇田川拓雄 (1992)、統計データ解析チューターシステムの開発、平成3年度科学研究費補助金研究成果報告書。

[44] 武田啓子、石橋雄一 (1987)、データ解析システムへの要件、日本計算機統計学会シンポジウム講演予稿集、PP 5-6。

[45] 武田啓子 (1990)、データ解析エキスパートシステムの可能性とその試作、日本計算機統計学会大会論文集、PP 5-8。

[46] 田中豊、垂水共之、脇本和昌 (1984)、パソコン統計解析ハンドブック第II巻多変量解析編、共立出版。

[47] 田中豊、垂水 共之 (1986)、パソコン統計解析ハンドブック第III巻実験計画法編、共立出版。

[48] 垂水共之、林篤裕、田中豊、脇本和昌 (1987)、パソコン統計解析ハンドブックとＳｔａｒ／Ｂ、日本行動計量学会発表論文抄録集、PP A3-13-1–A3-13-4。

[49] 垂水共之、林篤裕 (1988)、Ｓｅｔｏ／Ｂ — パソコン統計解析ソフトウェア、共立出版。

[50] 中野純司、山本由和、岡田雅史 (1991)、知識ベース重回帰分析支援システム、応用統計学、Vol.20, No.1. PP 11-23。

[51] 林篤裕 (1986)、パソコン用統計プログラムの開発について、「データ解析ソフトウェア・システムの研究開発」科研シンポジウム配布資料。

[52] 林篤裕、垂水共之 (1987)、Ｓｔａｒ／Ｂ — その機能と特徴 —、日本計算機統計学会大会論文集、PP 9-13。

[53] 林篤裕 (1987)、統計プログラムパッケージのあり方とＳｔａｒ／Ｂ、川崎医学会誌一般教養篇、第１３号、PP 49-57。

[54] 林篤裕、垂水共之 (1988)、データ解析における手法選択のコンサルテーションシステムについて、分類とその応用に関する研究会、研究報告予稿集、PP 50-51。

[55] 林篤裕、垂水共之 (1989)、蓄積を目的とした統計データの記述とその利用、日本行動計量学会発表論文抄録集、PP 147-148。

[56] 林篤裕、垂水共之 (1990)、データ解析における手法選択のコンサルテーションシステムシステム、統計エキスパートシステム研究会配布資料。

[57] 林篤裕、垂水共之、柳貴久男 (1991)、ハイパーツールを用いた統計解析コンサルテーションシステムについて、日本統計学会講演報告集、PP 23-25。

[58] 林篤裕、垂水共之、柳貴久男 (1991)、ハイパーツールを用いた統計解析コンサルテーションシステムの構築、日本計算機統計学会シンポジウム論文集、PP 63-66。

[59] 本位田真一、市川照久 (1989)、エキスパートシステム基礎技術、オーム社。

[60] MaxThink Inc. (1989), User's Manual of HyperLink, (ウェーブトレイン、ＨｙｐｅｒＬｉｎｋユーザーズマニュアル)。

[61] 脇本和昌、垂水共之、田中豊 (1984)、パソコン統計解析ハンドブック第Ｉ巻基礎統計編、共立出版。

[62] 脇本和昌、田中豊、垂水共之 (1992)、パソコン統計解析ハンドブック第VI巻グラフィックス編、共立出版。

A. Appendix (Statistical Knowledge)

We show a part of our statistical knowledge in the system of Japanese version. Each knowledge is managed some text files in the system. It is a simple structure of computer files. So, anyone makes new text files by an editor as knowledge, and add to the system easily. The file name is bracketed ('[ ]' symbol) in the following list.

Users must prepare related statistical programs, dictionary of statistical terms and a kind of hypertool (we use HyperLink) except this knowledge.

List A.1. A part of the Statistical Knowledge

```
=====[start0.txt]=====

        ------------------------------------------------------

        統計解析手法選択のコンサルテーションシステムにようこそ！

        ------------------------------------------------------

        統計解析を行おうとした時、どの手法を使ったら良いのか、悩む事は
        ありませんか？

        本システムは、あなたと対話する事によって、解析目的から、統計解析の
        手法をアドバイスするものです。

        これから、いくつかの質問をします。
        「↑」「↓」キーを使って適当と思われる項目にカーソルを移動させ、
        「→」キーか、「リターン」キーを押して下さい。

        始めるには［→］キーを押してください。
                                                〈CD ai〉
```

List A.1. (Continued)

```
=====[start.txt]=====

                    --------------------------------

                          では、始めましょう

                    --------------------------------

      まず、解析を行いたいファイル名を教えて下さい。〈DOS filename〉

      その後、解析目的を聞いていきます。      〈t00〉
                                          Help=〈hstart〉

      詳しい説明が見たいときには、各説明の後ろにあるヘルプを呼び出して下さい。

      用語の解説は〈DOS dict〉の所を選択して下さい。


=====[hstart]=====

このシステムの進め方（システムとの対話方法）
========================================          用語の解説 = 〈DOS dict〉

      ・あなたの持っているデータに適した解析手法をアドバイスします。
        このためには、データを観測した、もしくは調査した人が、
        どの様な目的でデータを採取したかを教えてもらう必要があります。

      ・また、データの形式やｘｘについても質問しますので、
        あなたのデータに合った項目を選んで下さい。

      ・目的や項目を選択する場合には、これらの後ろにある、
        「＜＞」（ギュメと呼ぶ）の所にカーソルを移動させて
        「リターン」キーを押して下さい。（「→」キーも同じ働きをします。）

      ・簡単な説明（ヘルプ）を項目毎に用意しましたので、参考にして下さい。
        ギュメの中の頭文字が「ｈ」のものが、ヘルプです。

      ・間違ったり、ヘルプを見終わったりして、
        前に戻りたいときは、「←」キーを押して下さい。

      ・複数の画面が用意されている場合がありますが、その場合には、
        「up」「down」「↑」「↓」キーで上下に移動する事ができます。

      ・画面右上の「dict」は用語集を起動するものです。わからない用語を検索。

      ・そこで、一般的に考えられる解析目的を列挙しました。
        これらの中から、あなたの考えに合った目的を探して下さい。
```

List A.1. (Continued)

```
=====[t00]=====

解析の目的を教えて下さい<h00>
========================                    用語の解説 = <DOS dict>

  ・ある項目を他の項目で説明したり、今後の変化を予測したい<ta01>
                                                    <ha01>
  ・複数個の項目をより少ない項目で代表（要約）させて説明したい<ta02>
                                                        <ha02>
  ・ものや項目間の関係に基づいて分類したい<ta03>
                                              <ha03>
  ・似たもの同志をまとめて幾つかの群に分類したい<ta04>
                                                  <ha04>
  ・項目間の関係を説明する隠れた（表面に現れない）構造を知りたい<ta05>
                                                        <ha05>
  ・平均値や分散を検定したい<t07>
                                <h07>
  ・幾つかの群の間に差があるかを比較したい<t06>
                                              <h06>
  ・データの概観を眺めたい<t08>
                              <h08>


=====[h00]=====

解析の目的を教えて下さい
========================

    [説明]

      ・あなたの持っているデータに適した解析手法をアドバイスします。

      ・このためには、データを観測した、もしくは調査した人が、
        どの様な目的でデータを採取したかを知る必要があります。

      ・そこで、一般的に考えられる解析目的を列挙しました。
        これらの中から、あなたの考えに合った目的を探して下さい。
```

List A.1. (Continued)

```
=====[ta01]=====

ある項目を他の項目で説明したり、今後の変化を予測したい<ha01>
==================================================== 用語の解説 = <DOS dict>

  ・ある項目を他の項目で説明する時に使われる解析手法は、
    幾つか用意されています。
  ・「説明される項目」と「説明する項目」の測定タイプの組み合わせによって
    そのデータに適した解析手法が異なります。
  ・あなたが持っているデータに合致するタイプを以下の表から選んで下さい。

    説明される項目  説明する項目  |  手法／説明
   ------------------------------------------
         量的          量的      |  <t01-1-1>
                                 |  <h01-1-1>
   ------------------------------------------
         量的          質的      |  <t01-1-2>
                                 |  <h01-1-2>
   ------------------------------------------
         質的          量的      |  <t01-2-1>
                                 |  <h01-2-1>
   ------------------------------------------
         質的          質的      |  <t01-2-2>
                                 |  <h01-2-2>
   ------------------------------------------


=====[ha01]=====

ある項目を他の項目で説明したり、今後の変化を予測したい
==================================================== 用語の解説 = <DOS dict>

  ・ある項目を他の項目で説明する場合、
    前者を「説明される項目」とか「従属変数」と、
    後者を「説明する項目」とか「独立変数」と呼ぶ。
    また、「説明する項目」について、新しいデータが得られたら、
    それを基に、「説明される項目」を予測する事もできる。

  ・データのタイプには、定量的に測定された「量的」データと
    定性的に測定された「質的」データがある。
    「量的」データの例としては、「体重」「血圧」「テストの点数」が
    「質的」データの例としては、「性別」「職業」「症状の程度」が挙げられる。

  ・例１：「ボール投げの飛距離」を、基礎的な体力を示す「握力」
    「身長」「体重」で説明しようとした場合、
    「ボール投げの飛距離」が「説明される項目」であり、
    「握力」「身長」「体重」が「説明する項目」である。
    またこの場合は、「説明される項目」「説明する項目」共に、
    「量的」に観測されたデータである。

  ・例２：「テレビ番組の視聴率」を、「番組の内容」「放送時間帯」
    で説明しようとした場合、
    「テレビ番組の視聴率」が「説明される項目」であり、
    「番組の内容」「放送時間帯」が「説明する項目」である。
    またこの場合は、「説明される項目」が「量的」に観測された
    「説明する項目」が「質的」に観測されたデータである。
```

List A.1. (Continued)

```
=====[ta02]=====

複数個の項目をより少ない項目で代表（要約）させて説明したい<ha02>
=========================================================          <DOS dict>

    外的基準がある<ta02-1>
                  <ha02-1>

    外的基準がない<ta02-2>
                  <ha02-2>


=====[ta03]=====

ものや項目間の関係に基づいて分類したい<ha03>
====================================          用語の解説 = <DOS dict>

    ものや項目間の関係が
        類似度で測定されている<t03-1>
                              <ha3-1>

        類似度で測定されていない<t03-2>
                              <ha3-1>


=====[ta04]=====

適当と思われる手法
==================

    クラスター分析を行うのが適当だと思います

    しかし、その前にデータのチェックを行う事をお薦めします。
        クラスター分析を行う前のチェック項目<tcst1-1>
                                            <hcst1-1>

    クラスター分析を行う<DOS clust>
                        <hcst1-2>

    また、クラスター分析を行った後にチェックする事柄もあります。
        クラスター分析を行った後のチェック項目<tcst1-3>
                                            <hcst1-3>
```

List A.1. (Continued)

```
=====[ta05]=====

適当と思われる手法
==================

    因子分析を行うのが適当だと思います

    しかし、その前にデータのチェックを行う事をお薦めします。
        因子分析を行う前のチェック項目<tfac1-1>
                                      <hfac1-1>

    因子分析を行う<DOS factor>
                  <hfac1-2>

    また、因子分析を行った後にチェックする事柄もあります。
        因子分析を行った後のチェック項目<tfac1-3>
                                        <hfac1-3>

=====[t01-1-1]=====

適当と思われる手法
==================

    回帰分析を行うのが適当だと思います

    しかし、その前にデータのチェックを行う事をお薦めします。
        回帰分析を行う前のチェック項目<treg1-1>
                                      <hreg1-1>

    回帰分析を行う<DOS mreg>
                  <hreg1-2>

    また、回帰分析を行った後にチェックする事柄もあります。
        回帰分析を行った後のチェック項目<treg1-3>
                                        <hreg1-3>
```

List A.1. (Continued)

```
=====[h01-1-1]=====

説明される項目：量的
説明する項目　：量的
====================                    用語の解説 = 〈DOS dict〉

  [説明]

    ・説明される項目も説明する項目も共に量的に測定されている場合。

    ・例えば、「ボール投げの飛距離」を、基礎的な体力を示す「握力」
      「身長」「体重」でどの程度説明できるか解析するような場合である。

    ・この例の場合、説明される項目は、「ボール投げの飛距離」であり、
                    説明する項目は、「握力」「身長」「体重」である。


=====[t01-1-2]=====

適当と思われる手法
==================

    数量化１類を行うのが適当だと思います


    しかし、その前にデータのチェックを行う事をお薦めします。
      数量化１類を行う前のチェック項目〈tqnt1-1〉
                                      〈hqnt1-1〉


    数量化１類を行う〈DOS quant1〉
                    〈hqnt1-2〉


    また、数量化１類を行った後にチェックする事柄もあります。
      数量化１類を行った後のチェック項目〈tqnt1-3〉
                                      〈hqnt1-3〉
```

List A.1. (Continued)

```
=====[h01-1-2]=====

説明される項目：量的
説明する項目　：質的
====================                    用語の解説 = 〈DOS dict〉

  [説明]

    ・説明される項目が量的に、説明する項目が質的に測定されている場合。

    ・例えば、「テレビ番組の視聴率」を、それぞれ４分類された「番組の内容」と
      「時間帯」でどの程度説明できるかを解析するような場合である。

    ・この例の場合、説明される項目は、「テレビ番組の視聴率」であり、
                    説明する項目は、「番組の内容」「時間帯」である。

    ・この様なデータの場合、説明される項目の事を、「外的基準」、
                          説明する項目の事を、「要因」と呼ぶ。


=====[t01-2-1]=====

適当と思われる手法
==================

    判別分析を行うのが適当だと思います


    しかし、その前にデータのチェックを行う事をお薦めします。
        判別分析を行う前のチェック項目〈tdsc12-1〉
                                        〈hdsc12-1〉


    判別分析を行う〈DOS disc12〉
                  〈hdsc12-2〉


    また、判別分析を行った後にチェックする事柄もあります。
        判別分析を行った後のチェック項目〈tdsc12-3〉
                                        〈hdsc12-3〉
```

List A.1. (Continued)

```
=====[h01-2-1]=====

説明される項目：質的
説明する項目　：量的
====================
```

用語の解説 = 〈DOS dict〉

［説明］

・説明される項目が質的に、説明する項目が量的に測定されている場合。

・例えば、３種類に分類された「あやめ科の植物」を、「がくの長さ」「がくの幅」「花弁の長さ」「花弁の幅」でどの程度説明できるかを解析するような場合である。

・この例の場合、説明される項目は、「あやめ科の植物」であり、
説明する項目は、「がくの長さ」「がくの幅」
「花弁の長さ」「花弁の幅」である。

```
=====[t01-2-2]=====

適当と思われる手法
==================
```

数量化２類を行うのが適当だと思います

しかし、その前にデータのチェックを行う事をお薦めします。
数量化２類を行う前のチェック項目〈tqnt2-1〉
〈hqnt2-1〉

数量化２類を行う〈DOS quant2〉
〈hqnt2-2〉

また、数量化２類を行った後にチェックする事柄もあります。
数量化２類を行った後のチェック項目〈tqnt2-3〉
〈hqnt2-3〉

List A.1. (Continued)

```
=====[h01-2-2]=====

説明される項目：質的
説明する項目　：質的
====================                    用語の解説 = <DOS dict>
```

［説明］

・説明される項目も説明する項目も共に質的に測定されている場合。

・例えば、３つに分類された「テレビ番組の視聴率」を、
それぞれ４分類された「番組の内容」と「時間帯」で
どの程度説明できるかを解析するような場合である。

・この例の場合、説明される項目は、「テレビ番組の視聴率」であり、
説明する項目は、「番組の内容」「時間帯」である。

・この様なデータの場合、説明される項目の事を、「外的基準」、
説明する項目の事を、「要因」と呼ぶ。

```
=====[ta02-1]=====

外的基準がある<ha02-1>
==============                          用語の解説 = <DOS dict>
```

・外的基準をを他の項目で説明する時に使われる解析手法は、
幾つか用意されています。
・「外的基準」と「説明する項目」の測定タイプの組み合わせによって
そのデータに適した解析手法が異なります。
・あなたが持っているデータに合致するタイプを以下の表から選んで下さい。

外的基準	説明する項目	手法／説明
量的	量的	<t01-1-1> <h01-1-1>
量的	質的	<t01-1-2> <h01-1-2>
質的	量的	<t01-2-1> <h01-2-1>
質的	質的	<t01-2-2> <h01-2-2>

List A.1. (Continued)

```
=====[ta02-2]=====

外的基準がない〈ha02-2〉
==============                                    用語の解説 = 〈DOS dict〉

  ・複数個の項目をより少ない項目で代表させて説明する時に使われる解析手法は、
    幾つか用意されています。
  ・「説明する項目」の測定タイプの組み合わせによって
    そのデータに適した解析手法が異なります。
  ・あなたが持っているデータに合致するタイプを以下の表から選んで下さい。

          説明する項目   |  手法／説明
        ------------------------------
              量的       |  〈t03-1-1〉
                         |  〈h03-1-1〉
        ------------------------------
              質的       |  〈t02-2-2〉
                         |  〈h02-2-2〉


=====[treg1-1]=====

回帰分析を行う前のチェック項目〈hreg1-1〉
============================                      用語の解説 = 〈DOS dict〉

  回帰分析を始める前に、以下の事柄をチェックした方が良いと思います。

    ・負のウェイトがないかチェックする〈DOS nweight〉
                                      〈hnweight〉
    ・欠損値がないかチェックする〈DOS miss〉
                                 〈hmiss〉
    ・説明される項目（変量ｙ）についてチェックする〈treg1-1-.3〉
                                                  〈hreg1-1-.3〉
    ・説明する項目（変量Ｘ）についてチェックする〈treg1-1-.4〉
                                                 〈hreg1-1-.4〉
    ・多重共線性についてチェックする〈DOS multico〉
                                   〈hmultico〉
```

List A.1. (Continued)

```
=====[hreg1-1]=====

回帰分析を行う前のチェック項目
==============================                    用語の解説 = <DOS dict>

  ・回帰分析を始める前に、回帰分析を行えるようなデータかどうかを
    チェックしておいた方が良いと思います。

  ・チェックする項目としては以下のものが考えられます。
      ◎負のウェイトがないかチェックする。
      ◎欠損値がないかチェックする。
      ◎説明される項目（変量ｙ）についてチェックする。
      ◎説明する項目（変量Ｘ）についてチェックする。
      ◎多重共線性についてチェックする。


=====[hreg1-2]=====

回帰分析を行う
==============                    用語の解説 = <DOS dict>

  [説明]

      ・説明される項目も説明する項目も共に量的に測定されている
        データに対しては、回帰分析を行うことによって分析できる。

      ・例えば、「ボール投げの飛距離」を、基礎的な体力を示す「握力」
        「身長」「体重」でどの程度説明できるか解析するような場合である。

      ・この例の場合、説明される項目は、「ボール投げの飛距離」であり、
                      説明する項目は、「握力」「身長」「体重」である。


=====[treg1-3]=====

回帰分析を行った後のチェック項目<hreg1-3>
==================================                    用語の解説 = <DOS dict>

  回帰分析を行った後に、以下の事柄をチェックした方が良いと思います。

    ・残差の正規性<DOS qqplot>
                  <hqqplot>

    ・各説明変数の偏回帰プロット<DOS pregpl>
                                 <hpregpl>

    ・系列相関（Ｄｕｒｂｉｎ－Ｗａｔｓｏｎ統計量）<DOS durbin>
      残差の関連性、残差の傾向（トレンド）          <hdurbin>

    ・影響の大きなケースの検出（ＤＦＦＩＴＳ、Ｃｏｏｋ）<DOS influe>
      Ｏｕｔｌｉｅｒの発見                              <hinflue>
```

List A.1. (Continued)

```
=====[hreg1-3]=====

回帰分析を行った後のチェック項目
================================                    用語の解説 = <DOS dict>

  ・回帰分析を行った後に、回帰分析が仮定している条件を満たしているかを
    チェックした方が良いと思います。

  ・チェックする項目としては以下のものが考えられます。
    ◎残差の正規性
    ◎各説明変数の偏回帰プロット
    ◎系列相関（Ｄｕｒｂｉｎ－Ｗａｔｓｏｎ統計量）
      残差の関連性、残差の傾向（トレンド）
    ◎影響の大きなケースの検出（ＤＦＦＩＴＳ、Ｃｏｏｋ）
      Ｏｕｔｌｉｅｒの発見


=====[hnweight]=====

負のウェイトがないかチェックする
================================

  [説明]

      ・ケース毎の重み付けは、正数の必要がある。

      ・入力ミス等で、負数が入力されていると間違った結果を導くので
        負のウエイトがないかチェックする。

      ・負のウェイトが有った場合は、ウエイトを正数に修正するか、
        そのサンプルを取り除く作業が必要である。


=====[hmiss]=====

欠損値がないかチェックする
==========================

  [説明]

      ・データ採取時に測定できなかった測定値に付いては
        以下に示すような処置を行う必要がある。

      ・欠損値のあるサンプルを取り除く。

      ・欠損値をある値に置き換える。
```

List A.1. (Continued)

```
=====[treg1-1-.3]=====

変量ｙのチェック<hreg1-1-.3>
===============                                    用語の解説 = <DOS dict>

  目的変数について調べる

    ・データの散らばり方を調べる<DOS spac>
                                  <hspac>

    ・データの偏り具合を調べる<DOS skew>
                                <hskew>
        Ｘについての間違いか？

    ・１つの値に集中していないか
      特定の値にかたまっていないか調べる<DOS atypical>
                                        <hatypical>


=====[treg1-1-.4]=====

変量Ｘのチェック<hreg1-1-.4>
===============                                    用語の解説 = <DOS dict>

  説明変数について調べる
    ・データの偏り具合を調べる<DOS skew>
                              <hskew>

    ・１つの値に集中していないか調べる
                              <DOS atypical>
                              <hatypical>
```