USING DATA MINING IN EDUCATIONAL RESEARCH:

A COMPARISON OF BAYESIAN NETWORK WITH MULTIPLE REGRESSION

IN PREDICTION

by

Yonghong Xu

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2003

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA ®
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Yonghong Xu

entitled USING DATA MINING IN EDUCATIONAL RESEARCH: A COMPARISON OF BAYESIAN NETWORK WITH MULTIPLE REGRESSION IN PREDICTION

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Darrell Sabers Darrell L. Sabers	11/19/2003 Date
Patricia B. Jones Patricia B. Jones	11/19/2003 Date
Jerome V. D'Agostino Jerome V. D'Agostino	11/19/2003 Date
	Date
	Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Darrell Sabers
Dissertation Director Darrell L. Sabers — 11/19/2003 Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: 

ACKNOWLEDGEMENTS

Upon the completion of this dissertation, the author would like to thank Drs. Darrell L. Sabers and Patricia B. Jones, who have encouraged and challenged me throughout my academic program and never accepted less than my best efforts. This dissertation could not have been written without their invaluable input and their willingness to share their knowledge and experiences; their friendly supports have been invaluable for me during my study, which happened to be a relatively difficult period of my life. They served not only as my academic advisors, but also as my professional and personal role models.

Drs. Sabers and Jones and other committee members, Dr. Jerome V. D'Agostino and Dr. Olivia L. Sheng, have patiently guided me through the dissertation process with their unselfish demonstration of individual expertise. My appreciation to them all.

DEDICATIONS

I should like to dedicate this dissertation to my beloved parents Yuchang Xu and Xiuying Yan. They may never understand a single word in this dissertation, but their love and support are the sources of all my accomplishments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Advances in technology have altered data collection and popularized large databases in areas including education. To turn the collected data into knowledge, effective analysis tools are required. Traditional statistical approaches have shown some limitations when analyzing large-scale data, especially sets with a large number of variables. This dissertation introduces to educational researchers a new data analysis approach called data mining, an analytic process at the intersection of statistics, databases, machine learning/artificial intelligence (AI), and computer science, that is designed to explore large amounts of data to search for consistent patterns and/or systematic relationships between variables.

To examine the usefulness of data mining in educational research, one specific data mining technique--the Bayesian Belief Network (BBN) based in Bayesian probability--is used to construct an analysis model in contrast to the traditional statistical approaches to answer a pseudo research question about faculty salary prediction in postsecondary institutions. Four prediction models--a multiple regression model with theoretical variable selection, a regression model with statistical variable extraction, a data mining BBN model with wrapper feature selection, and a combination model that used variables selected by the BBN in a multiple regression procedure--are expounded to analyze a data set called the National Survey of Postsecondary Faculty 1999 (NSOPF:99) provided by the National Center of Educational Services (NCES).

The algorithms, input variables, final models, outputs, and interpretations of the four prediction models are presented and discussed. The results indicate that, with a

nonmetric approach, the BBN can effectively handle a large number of variables through a process of stochastic subset selection; uncover dependence relationships among variables; detect hidden patterns in the data set; minimize the sample size as a factor influencing the amount of computations in data modeling; reduce data dimensionality by automatically identifying the most pertinent variable from a group of different but highly correlated measures in the analysis; and select the critical variables related to a core construct in prediction problems. The BBN and other data mining techniques have drawbacks; nonetheless, they are useful tools with unique advantages for analyzing large-scale data in educational research.

CHAPTER 1 INTRODUCTION

The science of statistics is often used to transform a set of raw data into useful information. Statistics became a critical component of the research methodology in the social and behavioral sciences because researchers wished to use a scientific approach of objective observation, hypothesis testing, and quantitative modeling of data structures (Allinson, 1999). In fact, the use of statistical techniques played a critical role in the formation of traditional quantitative research methods during the early part of last century with the classical procedures including linear regression and analysis of variance (ANOVA) to address effectively the needs of data analysis in agricultural and laboratory-based experiments.

In a traditional approach, a study begins with one or more research questions or testable hypotheses. The *a priori* knowledge or assumption specifies the underlying constructs being studied and determines the measurable variables. Then data are collected through a survey or a designed experiment. Because the cost is proportional to the total amount of data collected, most of the time the size of the data set is no more than absolutely needed to make valid statistical inferences (Wegman, 1988). As a result, the amount of data is parsimoniously small. In addition, variables are often assumed to be independently and identically distributed (IID) and a specific functional form of the model (often linear as in linear regression or ANOVA) is assumed. The analysis of data based on the functional form of the model leads to an answer to the hypothesis test.

If *traditional statistical methods* can be defined to cover all the parametric statistical techniques and procedures and their nonparametric counter-functions if

available (under this definition, general linear models including ANOVA, regression, and alike are typical examples of traditional methods; whereas most computation-intensive procedures, such as the bootstrapping and Markov chains, are not part of the traditional statistical methods), they apparently have been very successful in treating data collected through carefully designed experiments.

However, with dramatic progress in computing science, methods of data collection are undergoing changes. In the last decade, electronic data acquisition and database technology have allowed data collections that are substantially different from the traditional methods (Wegman, 1988; Hand, Mannila, & Smyth, 2001). For instance, credit card companies keep databases that track every transaction made by every customer. Banks have data ranging from customers' demographical information to every detail of their financial situation. There are even companies specializing in collecting and selling data of every aspect of human life. Overall, the low-cost electronic instruments result in data sets large in sample size and high in dimensionality stimulated by the belief that the information collected might be useful for some purposes in the future (Huber, 1999; Wegman, 1995). With frequent encounters of large-scale data in every aspect of management- and research-related areas, statisticians and data analysts have gradually realized the limitations of traditional statistics when trying to extract useful information from data of large volume.

Limitations of Traditional Statistics in Handling Large-Scale Data

Different from the small, low-dimensional homogeneous data sets collected in traditional research activities, large-scale data collections not only result in overwhelming

numbers of observations and variables (Hand et al., 2001; Wegman, 1995), but also produce the types of data ranging from regular numeric and text values to graphical and multimedia information (e.g., finger prints and pictures are saved electronically for police departments to track down criminals). Also, without systematical monitoring of the sampling process, most of the large data sets are collected from convenient or opportunistic samples; the potential selection bias menaces the validity of statistical inferences and generalizations (Hand, 1999; Hand et al., 2001). Moreover, in contrast to the clean and well-structured information in most experimental studies, noisy data and loose structures are common in large data sets and databases.

Statistical techniques are vulnerable to several problems when used to analyze large-scale data sets that are characteristically different from the small and clean data sets collected in designed studies. First, traditional inferential methods are not sufficient for dealing with the noisy data and irregularities inherent in the non-experimental data sets. The statistical methods are usually not robust against violation of the strong model assumptions including normality, linearity, and independence (Hand et al., 2001).

Second, with some predetermined model structures, the traditional approaches are not intended to uncover previously unknown patterns, which happen to be the potential values in most loosely-focused large data sets. Third, hypothesis testing is oversensitive to minor differences when the sample size is large (Rocke, 1998; Glymour, Madigan, Pregibon, & Smyth, 1997). Fourth, the definite "Yes" vs. "No" answer to a research question fails to quantify the magnitude of the observed differences when information about effect size is omitted.

Fifth, although traditional data analysis methods can handle a large number of cases through sampling and/or some sufficient statistics (e.g., mean and variance), they face another primary challenge, the so-called "curse of dimensionality", when a data set has a large number of variables (Hand, 1999; Hand et al., 2001). In this situation, the computational difficulty increases exponentially as the number of variables increases. High dimensionality also makes it nearly impossible to determine the model structure in advance as traditional statistical approaches require.

Finally, traditional inferential analysis is primarily concerned with variability of one or several dependent variables and questions related to this variability; but with a more complex data set, it is unavoidable to ask more questions to gain insight into the data structure and the functional relationships among the variables. Therefore, attention is shifted from statistical inference to structural inference (Wegman, 1988). Fully aware of the drawbacks of traditional statistics in analyzing large-scale data, Wegman (2000) made the argument that the changes in the nature and modes of data collection demanded new tools and techniques in data analysis.

From Computational Statistics to Data Mining

Statisticians have been paying close attention to the changes in data collection and the different needs in data analysis; the orientation of theoretical research in statistics under this climate is being discussed. Many statisticians, including Friedman (1997), Hand (1999), Parzen (1997), and Wegman (2000), support the idea that statisticians should be involved in topics of massive data set analysis, and statistics as an applied science should continue to build from the needs of other fields of research and practice

instead of focusing on the basic discipline. Taking the stand as a *computational* statistician, Wegman (2000) suggests that exploratory and structural inferences should be emphasized within a more function-oriented nonparametric framework and that algorithmically- and computationally-efficient analysis methods are necessary to understand large data sets of complex data structure, mixed data types, and a huge number of observations and measures.

In Wegman's argument, *computational statistics* is different from traditional statistics --"mathematical statistics" under his definition--because the latter are procedures that can be manually completed, whereas computational statistics refer to techniques that are only feasible with intensive computer work (e.g., bootstrapping, advanced visualization, and applications of Monte Carlo methods). According to him, traditional mathematical statistics is essentially a completed discipline: "the general principles are well-developed and contemporary statistical theory of traditional parametric and nonparametric methods is largely a solved problem" (Wegman, 2000, p. 3). Therefore, computational statistics, under a broad enterprise of "data analysis and inference", should be the future of statistics with the presence of digital data collection and rapid data growth. However, to many statisticians (e.g., Hand, 1998; Friedman, 1997), *computational statistics* as a research area gained far less attention than another data analysis approach called *data mining*, which Wegman (2000) introduced as "exploratory data analysis with little or no human interaction using computationally feasible techniques, i.e., the attempt to find interesting structure unknown *a priori*" (p. 6).

Data mining is not a new term to statisticians; it has been associated with negative connotations of blind exploration of data without *a priori* hypotheses to verify (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Fayyad, 1997a; Hand, 1999). However, it has taken on some new meaning as an automated and intelligent data analysis approach in the business world in recent years due to its ability to extract useful information in a timely manner from an ever-growing data set of a large volume. Due to its applied importance, data mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty.

It is difficult to find a unanimous definition of data mining in the literature (Friedman, 1997). Theoretically, data mining is a rapidly evolving area of research that is at the intersection of several disciplines, including statistics, database management, machine learning/ artificial intelligence (AI), and computer science (Mannila, 1996; Fayyad, 1997a; Friedman, 1997; Parzen, 1997; Hand, 1999). Technically, data mining is an analytic process designed to explore large amounts of data to search for consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new sets of data. The process thus consists of three basic stages: exploration, model building or pattern definition, and model validation/verification (StatSoft Electronic Textbook, n.d.).

The Relationship between Data Mining and Statistics

There is a strong connection between traditional statistics and data mining. Above all, data mining shares with statistics the same goal of discovering structures in data

(Hand, 1999). In its mission of finding useful but previously unknown structures, data mining also uses quite a few statistical techniques, including classification, clustering, artificial neural networks (ANN) and genetic algorithms, nonparametric regression, trend/ sequential analysis, density estimation, and many types of visual display methods (Friedman, 1997; Hand, 1998, 1999; Mannilla, 1996; Parzen, 1997; Wegman, 2000). Nevertheless, data mining is "not simply a reworking of some tried and true statistics techniques" (Wegman, 2000, p. 6). Data mining techniques and software have advantages in working with massive data sets, the size of which are typically overwhelming to regular statistical packages (Hand, 1999; Mannila, 1996). With a huge amount of data to process, algorithms are in the spotlight in data mining because of its reliance on machine learning and other related disciplines to accelerate computational efficiency (Hand, 1999; Mannila, 1996), whereas in traditional statistics, a good explicit model with quantified confidence or probability statements is the top priority to understand the data structure. Moreover, the exploratory nature of data mining is in opposition to traditional statistical methods commonly addressing confirmatory analysis and model fitting (Hand, 1999).

Acknowledging its close relationship with statistics, statisticians hold mixed feelings toward data mining as a new discipline. On the one hand, mainly due to their common mission of data analysis and the fundamental role of exploratory statistics and some multivariate techniques in data mining, some statisticians (e.g., Friedman, 1997; Hand, 1999; Parzen, 1997) argue that data mining should be a subset of statistics, or at least it would be so if the statistical community did not miss the chance to take full advantages of the modern computing science. They would like to see that statistics

continue to build in the area of massive data set analysis, which is essentially a statistical problem (Parzen, 1997; Petersen, 1997; Wegman, 1995, 2000).

On the other hand, some statisticians (e.g., Elder & Pregibon, 1996) are strongly opposed to the automated operation of data mining techniques. They argue that human expertise is crucial in model identification because not only can automated model search procedures often be fooled by anomalous association patterns, but also the algorithmic optimality in data mining is incapable of including qualitative distinctions between competing models of similar size for cost-effectiveness comparisons. Thompson (1997) insists that computing advancement is not a reason to drive statistics away from the traditional model fitting, and data-driven analysis methods cannot replace the objective reality of statistical model building. Statisticians should use powerful computers for more efficient and more general modeling techniques than was possible before such technology was available.

The core of this argument concerns the future of statistics as a science of data analysis in the age of high-speed computing and information explosion. The concerns are realistic given the fact that advanced computing power makes it fairly easy and rapid to collect or accumulate large volumes of data, but the traditional data analysis approaches become inefficient and expensive facing such data sets (Mannila, 1996; Fayyad, 1997b; Elder & Pregibon, 1996; Hand, 1999).

Analyzing Large-Scale Data in Educational Research

Although theoretical arguments regarding statistics as a science may not be a very strong concern to researchers in fields that only use data analysis procedures for

application purposes, it is surely a strong interest to all if some new and effective tools can be available to probe through a large amount of data and draw previously unknown information. Educational researchers are in need of such analytic tools because they have witnessed in their discipline an increasing number of large-scale data sets ranging from student information in school districts to national surveys of defined populations. For instance, many national survey databases--including data of, but not limited to, faculty, students, and institutions from K-12 to postsecondary education--are maintained by the National Center of Educational Statistics (NCES) for educational and research interests.

Many studies have been done with the large data sets available from national databases and other resources. Among them, Zheng (1996) used the School and Staffing Survey (SASS) to study how principals' instructional management behaviors were conditioned by contextual factors; Huang (1995) summarized the use of the Common Core of Data (CCD) and the National Education Longitudinal Study (NELS:88) in studying rural education in the United States. Studies have also been done with resources from other databases including the U.S. Census, the American Association of Community Colleges (AACC, 1998), and so on.

In general, however, the statistical methods used in most of the educational studies with large data sets are limited to traditional analysis procedures. Among some of the popular approaches, one trend is to present itemized summary information graphically or in tables. For instance, the AACC offered a quick comparison of community college enrollments and median household incomes from state to state by presenting the summary data from the national databases in some charts (AACC, 1998). A second

approach is to present descriptive statistics and simple comparisons of means as findings. One example is a study by Russell (1991) based on the data set of the National Survey of Postsecondary Faculty 1988 (NSOPF:88). She presented her results in a descriptive manner and ran simple analyses, such as comparing percentages of male vs. female and minority vs. nonminority faculty at each rank, comparing average years of service and numbers of publications in different gender and ethnic groups, and similar comparisons on other important factors.

Third, a subset of observations and/or a subset of the variables are taken from a large database to answer some specific research questions or to meet some particular research interests. For example, Zheng (1996) selected a group of variables from the national database based on a predetermined model and examined the significance of several particular factors in his study of principals' instructional management behaviors.

In summary, although theoretical explorations have been underway for advanced techniques of large-scale data analysis (e.g., Daszykowski, Walczak, & Massart, 2002), in practice educational researchers are prone to use simple statistical methods or graphical presentations to outline the relationships among some important variables. If there are preconceived research questions, researchers tend to select the group of participants of interest and to choose a small number of variables from a large pool of measures based on some well-known theories, previous studies, or sometimes educated guesses. That is, the database is tailored to fit a study in the traditional research manner.

Some disadvantages of using the traditional approaches to analyze large-scale data are exposed. First, descriptive statistics are seldom sufficient for uncovering the

important information in a data set. Second, with traditional inferential methods to analyze of a subset of a massive data collection, the major problem lies in the quality of data (e.g., sampling biases, missing data, and outliers) collected from convenient or opportunistic samples. The large database is treated only as a convenient way to save the trouble of collecting data, and there is no guarantee that all the important variables are measured and quantified as desired by the researchers. Third, researchers are likely to leave out variables that are not of interest or that are not evidently related to the questions they have in mind. It may not seem wrong to select a few important variables when the traditional research approach is taken as long as a complete model can be built without bias; however, when the goal of data analysis is to look for useful information from a collection of data, it is troublesome to ignore the potential values hidden in the large data set in which so many new variables exist. Therefore, the argument is that researchers should not be restricted to the traditional approaches; they should actively look for new methods to attack new problems; especially when working with large-scale data with a large number of variables, researchers should look for the right tools that can help to reveal the hidden patterns and to gain insight of the variable relationships that were previously unknown.

Research Questions

Currently, not many choices are available for analyzing data sets with large numbers of variables. Statistically, multivariate analysis techniques, including exploratory factor analysis (EFA), principal component analysis (PCA), and cluster analysis (CA), can be used to simplify the data structure when many variables exist and

some of them are highly related to one another (Fayyad, 1997b). After the dimensionality of a large data set is reduced, some traditional data analysis methods become applicable.

Recently, in business and scientific research, data mining has become a new approach to large-scale data analysis. As previously mentioned, data mining offers a group of new techniques specifically designed to perform efficient analysis on massive data sets by selecting critical variables and building accurate models. With data mining being used in business and scientific information management for over a decade, and as a research discipline continuing to grow with the inputs from statistics, machine learning, and computer science, it is time to see whether this new approach can benefit educational researchers with regard to large-scale data analysis.

So far, not many educational studies have used multivariate procedures in analyzing large data sets, and no publications are found with data mining as the tool for data analysis. Data mining is a still a novelty to most academic researchers who are familiar with traditional statistical methods. With the advancement in technology and data collection, the goal of this study is to introduce this new data analysis approach and to determine its usefulness in educational research, an area that regularly applies traditional statistical data analysis methods. This dissertation demonstrates the analysis of a large education-related data set with several different approaches, including data mining, traditional statistical methods, and a combination of these two; by laying different analyses of the same data set side by side, a solid case is provided to assess data mining techniques in comparison to comparable statistical methods. A clear-cut answer may be difficult regarding the specialties and advantages of individual approaches;

however, looking at a problem from different viewpoints itself is the essence of the study and hopefully it can provide critical information for researchers to make their own judgment about how well these different methods work to provide insight into the structure of and to extract valuable information from large-scale data sets.

Because data mining shares a few common concerns with traditional statistics including estimation of uncertainty and construction of models in defined problem scope (Glymour, Madigan, Pregibon, & Smyth, 1996), in order to narrow down the research problem, prediction functions are chosen to see whether data mining can offer any unique outlook when processing large data sets. Different models are set to search for factors that were most efficient in predicting faculty salary. On the statistical side, multiple linear regression is employed because it has been used as a dynamic procedure of prediction for a long time; for data mining, the prediction is performed with the *Bayesian Belief Network* (BBN), an algorithm based on Bayes's Theorem.

For the purpose of demonstration, a relatively large data set on faculty compensation is analyzed to answer the pseudo research question of salary prediction. The data set is considered appropriate because it is an education-related survey data set, not too large for traditional analysis approaches, and not too small for data mining techniques.

By using the most appropriate techniques in statistics and data mining to predict faculty salary, this study is conducted to examine the data thoroughly and to build the best models of prediction through different analytical approaches. With the presentation and discussions of the algorithms, the input variables, the final models, the outputs, and

the interpretations of the results of different analysis approaches, the objective is to assist readers to think independently and search for answers to the following questions. First, what are the similarities, differences, strengths, and weaknesses of the data mining and traditional statistics when performing predictions with a *large* data set? The definition of "large" is two-fold: a large number of observations (records) and a large number of variables (measures or attributes). Second, when used for exploratory purpose, which approach uncovers more useful information about the data structure: traditional statistics, data mining, or a combination of the two at different stages of analysis?

CHAPTER 2 REVIEW OF LITERATURE

For a good understanding of data mining and what it does differently from statistics, the review of literature is started by looking at the evolution of data analysis and the driving force behind it. Simply, this chapter is organized in the following order: the first section is a brief review of statistics and data mining from developmental point of view, including a comparison of the two disciplines at a conceptual level and a brief discussion of their relationship. In the second section, the analysis techniques used in this study are introduced. Final discussion is about some model evaluation criteria used in statistics and in data mining.

Conceptual Review of Statistics and Data Mining

Both statistics and data mining are concerned with analysis of data. Statistics as a discipline is more than one hundred years old and rooted in probability theory; whereas data mining is a relatively new methodology influenced by several disciplines including statistics, machine learning, and database management. As briefly mentioned in the previous chapter, the two disciplines have common goals and share some techniques, but they also have substantial differences as well. In this section, the theoretical development and conceptual structures of the two disciplines are outlined, followed by a discussion of the relationship between the two.

Statistics

Statistics is defined in the American Heritage Dictionary (2000) as "the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling."

Statistics has a long antiquity but a relatively short history as an applied science. As early as 2000 B.C., statistics were used in China to collect and compile data for use in public policy making (Rao, 1997). It was not until the 18th century that, with the growth of probability theory and the need to make decisions or predictions in face of uncertainty, the idea of drawing inferences from sampled data drove statistics to acquire a new meaning as methods of interpreting data or extracting information from data for making decisions (Rao, 1997; Wolfram, 2002). Since then and during the 19th century, the progression of statistical theory offered and elaborated concepts and techniques including the method of least squares, law of error, Law of Large Numbers, correlation, regression, χ^2 distribution, significance tests, and the ANOVA (Hald, 1998).

Also in the early 1900s, Fisher (1966) stressed the control of measurement bias and errors in data through well-structured experiments. His work initiated experimental design and related topics on research methodology, including sampling methods, survey design, and the study of reliability and validity, as new subdivisions of statistics. The methodological development gradually intensified confirmatory data analysis in the application of statistics. In the confirmatory approach, research questions are predefined, data are gathered through designed procedures, and collected evidence is then evaluated to provide an answer to the research questions in form of statements of significance and confidence provided by traditional statistical inferences (Hoaglin, Mosteller, & Tukey, 2000). It is called "confirmatory" to reflect the facts that the sample data are used to confirm the characteristics of the assumed population distribution and hypothesis testing is used to confirm *a priori* research hypotheses.

The essence of traditional statistical inference is that a sample consisting of randomly selected cases is considered to be representative of the population distribution. If the distribution of the sample follows some known probabilistic functions, predictions and inferences can be made about the population based on the observed behaviors of the relatively small group of selected subjects (Quenouille, 1958; Minium & Clarke, 1982). Therefore, a critical step in data analysis is to model the frequency and/or the probability of the randomly sampled data in order to make quantitative inferences about the population from which the data were drawn (Savage, 1972).

Furthermore, in inferential statistical analysis, imposed models on the sample data are both structural and stochastic (Shapiro & Gross, 1981). Structural models include, for example, regression, ANOVA, and other functional forms of models, which are predetermined to capture the relationships among variables. Meanwhile, some common stochastic models are assumed, including the errors (the difference between predicted and observed values) of the structural model following some known probabilistic distributions (Shapiro & Gross, 1981). With the structural model taken statistically as a general hypothesis and the sample data mathematically described with certain probability distributions, final conclusions can be made to answer the research questions. Because a structural model is presumed, estimation of the parametric properties of the model from the sample data with deductive calculations was at the core of classical statistics by the early 1900s.

Obviously, for making valid statistical modeling and sample-based inferences, the right probability distribution function (PDF) must be picked. However, when the sample

data collected in a study do not obey the model assumptions, which unfortunately happens quite often with small sample sizes, imprecise measurements, and ill-distributed populations (StatSoft Electronic Textbook, n.d.), a major threat to most classical parametric models is that the estimated model parameters may be misleading and the validity of the statistical findings can be in danger. To deal with this type of problem, the statisticians started to look for robust methods.

Some simple forms of robust and resistant procedures developed in first half of 1900s include nonparametric methods that are comparable to parametric methods (Rao, 1997), but stay distribution-free. Also, later in the 1960s, to decrease the instability of the estimators of location and regression coefficients caused by outliers or some small amounts of contamination in data, the influence function and diagnostics were proposed to further free statisticians of narrow models with unrealistic assumptions (Elder & Pregibon, 1996). In the 1970s, a group of statisticians, including Dempster, Laird, and Rubin (1977), presented many ways of solving estimation problems with incomplete data. Meanwhile, more studies were done to enhance the statistical techniques of dealing with categorical data. For example, Bishop, Fienberg, and Holland (1975) presented loglinear models that are analogous to linear models for continuous data (Elder & Pregibon, 1996).

A paradigm shift in statistics actually happened around 1970 when John Tukey pioneered exploratory data analysis (EDA) to promote the notion that statistical insights and modeling should be driven by data (Hoaglin et al., 2000; Rao, 1997). Different from the Fisherian confirmatory approach, EDA offers an alternative and wider view of

statistics in its methodological context, and stresses data- and objective-driven analysis, where visual interaction with the user plays a key role. As Hoaglin et al. (2000) stated:

> Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst. It often provides the first contact with the data, preceding any firm choice of models for either structural or stochastic components, and it also serves to uncover unexpected departure from familiar models. An important element of the exploratory approach is flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive steps of analysis uncover (p. 1).

The flexible probing of data provides analysts with an extensive repertoire of methods for in-depth study of a set of data. Advocates of EDA (e.g., Hoaglin et al, 2000) believe that, because of the imperfection of data in contrast to strict probabilistic model assumptions, it is necessary to look at details of the data before performing further analysis such as calculating summary statistics and carrying out data modeling and tests of hypotheses. Instead of being only "a set of tools" in seeking answers from available data, the primary concern of EDA is to define "a set of problems" in analyzing data sets (Friedman, 1997). Instead of settling with a single "right" answer, there are many answers in nearly all situations of data analyses (Elder & Pregibon, 1996).

EDA encompasses a lot of graphical and computational techniques to achieve four general missions: residual analysis, data visualization, data transformation or re-expression, and resistance procedures (Velleman & Hoaglin, 1981). With residual analysis EDA introduced a key change in statistical modeling: the decomposition of data

into structure and noise. Instead of partitioning somewhat an unnatural measure of variability--the squared units of the variance--into explained and unexplained or within-group and between-group variance, data are decomposed as:

data = model + error

or data = fit + residual

The "fit" or "model" describes the expected values of the data while the "error" or "residual" estimates the values that deviate from that expected value (Hoaglin et al., 2000). By iteratively examining the residuals, the analysts can assess the model adequacy and identify and move additional structure into the fit. This concept of residual is critical because diagnosis and treatment is only possible on the observed scale (Elder & Pregibon, 1996).

Another major credit of EDA is the renaissance of graphical methods (Elder & Pregibon, 1996). In data visualization, "a picture is worth a thousand words." With graphical tools including box plots, histograms, scatter plots, brushing, data smoothing, and much more, it is easier to spot outliers, discriminate clusters, check distributional and other assumptions, examine relationships, compare mean differences, and so on. As Tukey (1977) argued, good graphical methods allow unexpected values to present themselves so that the model can be modified to account for them.

The shift of statistics from confirmatory analysis to EDA happened at the same time as the early development of computer science (Elder & Pregibon, 1996). The rapid advancement in computing power and resources had strong impact on both data collection and data analysis. With the size of collected data sets increasing, the notions of

EDA were carried on to more liberal approaches in data analysis and to more advanced and computation-intensive procedures in the second half of last century. Generally, several major trends in the progression of data analysis were clear: first, multivariate analysis techniques were attracting more attention; second, statistics as a scientific means to overcome uncertainty in the process of decision making was emboldened and statistical decision theory was expanded from single-stage problems to sequential decision modeling; and also the notion of algorithmic function estimation through statistical learning from empirical data became an official component of the statistical theory.

Multivariate Analysis

Methodological changes in statistics are almost always driven by application needs. Easier data collection led to the development of multivariate analysis techniques capable of coping with increasing number of variables. With the different emphases, techniques under the name of multivariate analysis can be grouped into three major categories.

First, some traditional univariate dependence methods, including regression and ANOVA, were expanded to include multiple dependent variables. Dependence methods are defined as the procedures that try to explain the variance of some of the variables by the others, and so there is assumed dependency relationships among variables. Although dependence methods (e.g., ANOVA and multiple regression) are intrinsically "multivariate", only one dependent variable is generally predicted or explained by the independent variables and covariates. To include more dependent variables, multivariate

analysis of variance (MANOVA) and multivariate regression techniques were developed; although the mathematical operations are much more complex, multivariate dependency techniques are the same in logic and nature as their univariate counterpart. For example, MANOVA can be used to analyze several groups for their location in a space based on many variables. As in ANOVA, additional questions are examined about the relative contribution of each variable to the MANOVA.

This group of multivariate techniques are parametric; in other words, they have model assumptions including linearity, normality, and independence. Soon after, a theoretical breakthrough extended the usual normal linear model to "a much wider class of models that included probability models other than the normal distribution, and structural models that were nonlinear" (Elder & Pregibon, 1996, p. 4): the *Generalized Linear Models*. By decomposing the variation in a dependent variable into systematic and random components, Generalized Linear Models provide a unifying theory for regression-like models for categorical data and continuous data with non-normal distribution.

The second major category of multivariate methods refers to a group of techniques that look for patterns or relationships among a group of variables simultaneously and summarize data by reducing the number of variables necessary to describe the underlying structure. They are called independence methods because the study of variable relationships is conducted with no dependent variables. By looking at the interrelationships among variables, these multivariate methods are employed to develop taxonomies or systems of classification or to investigate useful ways to

conceptualize or group items (Johnson & Wichern, 1988). All of the techniques require that variables be in some form of interrelationships that can be quantified as similarity (e.g., correlation) or dissimilarities (e.g., distance). Examples of such techniques are EFA, PCA, projection pursuit regression, and CA.

The importance of dimensionality reduction techniques increases with the number of variables that need to be analyzed simultaneously in a data set; meanwhile, the popularity of these techniques is also reinforced by easy-to-use statistical software packages that can easily implement them on high-speed computers with large storage capacity (Krzanowski & Marriott, 1994; Dillon & Goldstein, 1984).

With few exceptions, most of the multivariate techniques for data reduction have assumptions of linearity or homogeneity. Exploratory projection pursuit (EPP) is one of the exceptions. The goal of EPP is to represent the input data items in a lower-dimensional space in such a way that certain properties of the structure of the input data set are preserved as faithfully as possible (Friedman , 1987; Friedman & Tukey, 1974). During the analysis, the data are projected linearly to reveal as much as possible the non-normally distributed structure of the data set. Based on how much the projected data deviate from the normally distributed data in the main body of its distribution, a numerical "interestingness" index can be assigned to each possible projection, and a projection angle is chosen to maximize the index. Projection pursuit tries to express some nonlinearities, but if the data set is high dimensional and highly nonlinear it may be difficult to visualize it with linear projections onto a low-dimensional display even if the "projection angle" is chosen carefully (Friedman , 1987; Friedman & Tukey, 1974).

A third category of multivariate analysis focuses on the causal relationships among variables. A well-known technique is called structural equation modeling (SEM), defined by Ullman (1996) as a method that examines a set of relationships between one or more independent variables and one or more dependent variables. SEM can deal with both observed and latent variables. An observed variable is a variable that can be observed directly and is measurable. A latent variable is a variable that cannot be observed directly and must be inferred from measured variables. In SEM, latent variables, also known as factors, constructs, or unobserved variables, are implied by the covariances among two or more measured variables.

SEM is a combination of multiple regression and factor analysis. Path analysis is a subset of SEM used to examine causal relationships between two or more variables. It is based upon a linear equation system and was first developed by Sewall Wright in the 1930s for use in phylogenetic studies. Path analysis was adopted by the social sciences in the 1960s and has been used to study relationships of measured variables (Maruyama, 1997). For causality examination, SEM and path analysis cannot statistically test directionality in relationships. The directions of arrows in a SEM analysis represent the researcher's hypotheses of causality within a system. The SEM approach remains useful in understanding multivariate relational data because of its abilities to distinguish between indirect and direct relationships between variables and to analyze relationships between latent variables without random errors (Maruyama, 1997).

The above-mentioned multivariate techniques rely more or less on some forms of probability distribution as the model base. However, statistics started to outgrow the

parametric models after the 1960s. Driven by needs in field applications, statisticians explored new possibilities and developed new theories and methods that are more flexible in terms of data handling and model building. Two major schools of theories--decision analysis and statistical learning--are introduced next. They are different from the multivariate procedures already introduced mainly because the selection of the functional form of a model is part of the analytical consideration.

Decision Analysis

With the wide application of statistics in broad areas, the interest in statistics as a means to overcome uncertainty in the process of decision making grew significantly strong (Rao, 1997). To meet the increasing complexity of data structures, decision making in statistics gradually led to the expansion from simple data analysis to sequential decision models and the formation of *decision analysis* as a subdivision of statistics (Chou, 1972; Shafer, Pearl, & Kaufmann, 1990).

Decision-making is part of everyday activity, but the systematic study of complex decision analysis did not appear until tools for coping with volumes of data were available. With incremental gains in the knowledge of decision making process and human thought, decision theory tries to support the process by analyzing disparate data using sophisticated statistical techniques (Raiffa, 1968). As a research subject, the basic ideas of decision theory started from the utility theory and the work by frequentist schools of probability as early as the 1930s and substantiated by the game theory in the 1940s. Then, when modern computers made immense computational tasks possible in the 1960s, the tree structure was formally introduced to the decision makers' lexicon by

Howard Raiffa (1968). Decision theory also rekindled the interest in a statistical theorem due to Bayes and its extensions including Markov chains. Since the late 1960s, the term *decision analysis* has been generally used to refer to the methodology that uses decision trees, Bayesian networks, and a number of new techniques, including multiattribute utility assessment and influence diagrams, to support decision making (Shafer et al., 1990). The following is a brief introduction of decision tree and Bayesian networks.

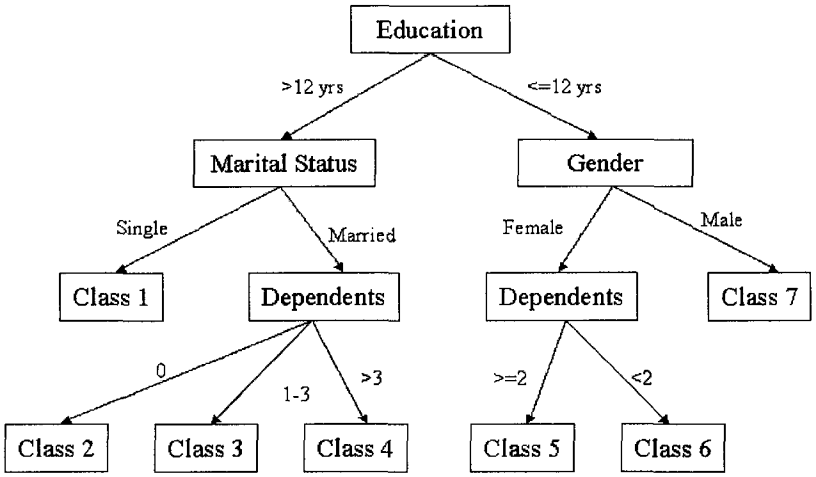


Figure 2.1. An example of the decision tree structure.

Decision Trees. With an appealing transparent structure, the decision tree is one of the most widely used methods for inductive inference (Mitchell, 1997). From the beginning, decision trees (also called classification trees) have been a kind of robust method of classification that approximate discrete-valued functions. Figure 2.1 is an example of a decision tree. A decision tree is a simple knowledge representation structure consisting of internal nodes (e.g., marital status, gender, dependents, and education in

Figure 2.1), leaf nodes (e.g., classes 1-7 in Figure 2.1), and branches. Each internal node is labeled with an attribute name and stands for a decision or test point. Each branch corresponds to a possible outcome of the test or a possible value of the attribute in the parent internal node. Each leaf node represents a predicted class, the output measure of classification. The number of levels of internal nodes is called the depth of a tree, which is functionally determined by a "purity" measure calculated from the data (Mitchell, 1997). The ideal purity is that only objects of the same class end up in the same leaf node.

A decision tree completely classifies data to a finite number of classes: Objects are classified by following a path down the tree from the root to the leaf. Each object goes through the internal nodes and takes the branches that correspond to the values or test results of its attributes. Technically, most algorithms of tree building are variations of a recursive top-down search through the entire space of possible decision trees. Statistical tests are used to evaluate each attribute on how well it classifies the instances in the data, and measures of homogeneity of the instances in every leaf node are also necessary (Mitchell, 1997).

BBN. Bayesian probability, often used to update the plausibility of a given statement in light of new evidence, started from a set of rules discovered by Thomas Bayes in the 1760s and updated by Laplace in the 1810s (Gillies, 2001). The basic version of *Bayes's Theorem* starts with a well-known product rule of probability for independent events (Cumming, 2003):

$$P(AB) = P(A) \times P(B), \tag{2.1}$$

where P(AB) means the probability of A and B happening together. The rule is a special case of the following product rule for dependent events, where P(A | B) means the probability of A given that B has already occurred:

$$P(AB) = P(A) \times P(B \mid A), \tag{2.2}$$

and $$P(AB) = P(B) \times P(A \mid B). \tag{2.3}$$

It is clear that

$$P(A) \times P(B \mid A) = P(B) \times P(A \mid B), \tag{2.4}$$

so, a simple version of Bayes's Theorem can be given as

$$P(A \mid B) = [\, P(A) \times P(B \mid A) \,] \;/\; P(B), \tag{2.5}$$

which provides the probability of event A happening given that event B has happened, calculated in terms of other known probabilities.

Similarly, different versions of the Bayes's Theorem can be developed. For instance, suppose that P(AB) needs to be calculated given that a third event, I, has happened. Written as P(AB | I), with the product rule Equations 2.2 and 2.3, P(AB| I) can be turned into

$$P(AB \mid I) = P(B \mid I) \times P(A \mid BI), \tag{2.6}$$

or $$P(AB \mid I) = P(A \mid I) \times P(B \mid AI). \tag{2.7}$$

Therefore,

$$P(B \mid I) \times P(A \mid BI) = P(A \mid I) \times P(B \mid AI), \tag{2.8}$$

that is

$$P(A \mid BI) = [\, P(A \mid I) \times P(B \mid AI) \,] \,/\, P(B \mid I), \tag{2.9}$$

which is another version of Bayes's Theorem that calculates the probability of event A happening given that event B and event I have happened. If A stands for hypothesis H, B for evidence E, and I for context C, the theorem can be written as

$$P(H \mid EC) = [\, P(H \mid C) \times P(E \mid HC) \,] / P(E \mid C), \tag{2.10}$$

where P(H | EC) is the probability of Hypothesis H given Evidence E in Context C.

The product rule eventually helps to chain more probabilities together. For instance, to find the probability of H given that E_1, E_2, and C have happened:

$$P(H \mid E_1E_2C) = [\, P(H \mid C) \times P(E_1E_2 \mid HC) \,] / P(E_1E_2 \mid C). \tag{2.11}$$

To find the probability of H given that E_1, E_2, E_3 and C have happened:

$$P(H \mid E_1E_2E_3C) = [\, P(H \mid C) \times P(E_1E_2E_3 \mid HC) \,] / P(E_1E_2E_3 \mid C). \tag{2.12}$$

Note that

$$\begin{aligned} P(E_1E_2E_3 \mid C) &= P(E_1 \mid E_2E_3C) \times P(E_2E_3 \mid C) \\ &= P(E_1 \mid E_2E_3C) \times P(E_2 \mid E_3C)\, P(E_3 \mid C), \end{aligned} \tag{2.13}$$

which can be used to calculate two of the values in the above equation.

The chained Bayes's Theorem can be made simpler with conditional independence, an assumption indicating that each variable was independent of its nondescendants in the network given the state of its immediate parents. For example, given that C is true, E_1 being true will not affect the probability of E_2 being true, then a simpler version of the chained Bayes's Theorem is possible:

$$\begin{aligned} &P(H \mid E_1E_2C) \\ &= [\, P(H \mid C) \times P(E_1 \mid HC) \,] \times P(E_2 \mid HC) \;/\; [\, P(E_1 \mid C) \times P(E_2 \mid C) \,]. \end{aligned} \tag{2.14}$$

This version makes it very easy to introduce new evidence into the situation. However, conditional independence is only true in some special situations.

A well-known terminology associated with Bayesian probability is the *Prior Probability* P(H | C), the probability of the hypothesis in context C regardless of the evidence. The notion of prior probability can be viewed positively or negatively. Some view it as positive because it allows the inclusion of input from domain experts for predicting future probability. Others take it as negative because subjectivity of human judgment in expert input is something often frowned upon. Relatively, P(H | EC), P(H | E_1E_2C), and such are called *Posterior Probability*, the modified probability based on available evidence.

A typical implementation of the Bayesian probability is found in tree-like Bayesian networks. In statistical modeling, Bayesian networks belong to a subset of the class of loglinear models for nominal data called graphical models (Elder & Pregibon, 1996). However, there is a conceptual difference between models based on Bayesian probability and the traditional statistical models: The model parameter H in the Bayesian net itself is considered a random variable, which is estimated and modified based on new input evidence.

Bayesian probability has been used in decision theory for over half a century, but the amount of computation in the network building process is overwhelming with even small data sets, so the field application of BBN was not realized until powerful computers became available in recent years.

Decision theory has helped to nudge the community of statistics from focusing on model estimation to model selection. Gradually, more methods were developed in which the modeling process is a search over structure space as well as parameter space. Assisted by advanced computing power, it is becoming common to consider many thousands of candidate structures in a data modeling process. In practice, some statisticians even explicitly blend the outputs from several models from different families to achieve estimates with reduced variance and better accuracy (Elder & Pregibon, 1996).

Statistical Learning Theory

Statistical learning theory was originally introduced for pattern recognition. Over the past 30 years, a general model of statistical learning has been developed for various statistical learning tasks including density estimation and regression function due to the work of Vladimir Vapnik. The general problem setting of statistical learning is to estimate predictive models from data. In traditional statistics the assumption is that the functional form of the correct prediction model is known and the goal is to estimate the model parameters; whereas in statistical learning theory the presumption is that the correct form is unknown and the goal is to identify the best possible model from a set of competing and admissible models (hypotheses; Pednault, 1999). According to Vapnik (1999), a general learning model can be described using three components:

1. A group of random variables or a generator of random vectors **X**, drawn independently from a fixed but unknown distribution P(x).
2. A supervisor that returns an output value y for every input vector **X**, according to a fixed but unknown conditional distribution function P($y \mid x$).

3. A learning machine capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters. The elements $\alpha \in \Lambda$ are not necessarily vectors, that is, α can be any abstract parameters. Therefore, $f(x, \alpha)$ in fact can be any set of functions.

The problem of learning is that of choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one that predicts the supervisor's response $\boldsymbol{Y}$ in the best possible way; that is, to select the best prediction model from the defined model space. The data set used for learning is called training data or "supervisors", which need to be large because the convergence and approximation in the learning fucntions are in essence extensions of the Law of Large Numbers to spaces of functions (Vapnik, 1999).

In order to choose the best available approximation function, a risk functional is defined as the expected value of the loss, which is the discrepancy between the response y in the "supervisor" training data (the observed value) and the response to a given input provided by the learning machine (the predicted value). Because the distribution is unknown, the risk functional is used to calculate expected loss for each admissible hypothesized function and the final model is the outcome of "learning" from empirical data with minimized risk value. This so-called Empirical Risk Minimization (ERM) is a general principle that has been used in least-squares methods in regression and Maximum Likelihood in density estimation. Vapnik (1995) also showed that the statistical learning could work with unrepresentative data because a large enough data set will keep the loss within a certain additive distance from the best hypothesis.

The core issue in statistical learning is to evaluate the performance of competing models; the theory provides a solid statistical basis for assessing adequacy of models that may have different mathematical forms and none of them need to be correct. Unfortunately, although introduced in the late 1960s, the flexibility and data-driven nature kept the statistical learning theory as a purely theoretical analysis tool and not well accepted by the mainstream of statisticians as an established approach until the 1990s (Vapnik, 1999). In the meantime, the same exploratory notions were highly welcomed in the fields of machine learning and pattern recognition, in which the learning theory is implemented in data preparation, evaluation of model fit and model consistency, handling of large numbers of variables, and other data-intensive tasks. In a recent book by Hastie, Tibshirani, and Friedman (2001), *The elements of statistical learning: Data mining, inference, and prediction*, statistical learning theory was reviewed as a bridge connecting statistics with data mining, and many methods within learning theory (e.g., kernel methods, neural networks, and support vector machines) have been realized in data mining algorithms.

Summary

The brief review of the history showed that statistics as an applied science have its fingerprints on almost all the aspects of data analysis. Different schools of statisticians advocate different fashions. The evolution of statistics is driven by two major factors: One is self-improvement, the other application needs. As Rao (1997) stated, "statistics is a peculiar subject without any subject matter of its own. It seems to exist and thrive by solving problems in other areas" (p. 158).

Statistics as a means of data analysis has dominated quantitative research for almost a century. However, in the last several decades, statistical research centered on some core theoretical issues and somewhat lost touch with the reality of field application (Parzen, 1997; Peterson, 1997). Also, reluctance to accept flexible approaches such as exploratory model selection limited statistics from keeping analysis power in line with data complexity. Even though decision theory and statistical learning theory have been around for decades, they did not attract much attention until recently. As a result, many opportunities are left for new ideas and new application tools designed to meet the changing demands of customized data analysis. Data mining is a major example of data analysis tools that have begun to fill the void in this area.

Data Mining

Data mining is a process of nontrivial extraction of implicit, previously unknown, and potentially useful information from a large volume of data. It uses machine learning, statistics, and visualization techniques to discover and present knowledge in a form that is easily comprehensible (Frawley, Piatetsky-Shapiro, & Matheu, 1991).

The advancement in computer technology and electronic data acquisition made data mining possible and necessary (Mannila, 1996; Fayyad, 1997b; Parzen, 1997; Hand, 1999; Friedman, 1997). According to its advocates (e.g., Aptė, 1997), data mining has prevailed as an analysis tool for dealing with large data sets because it can efficiently and intelligently probe through an immense amount of material to discover valuable information and make meaningful predictions that are especially important for decision-making under uncertainty.

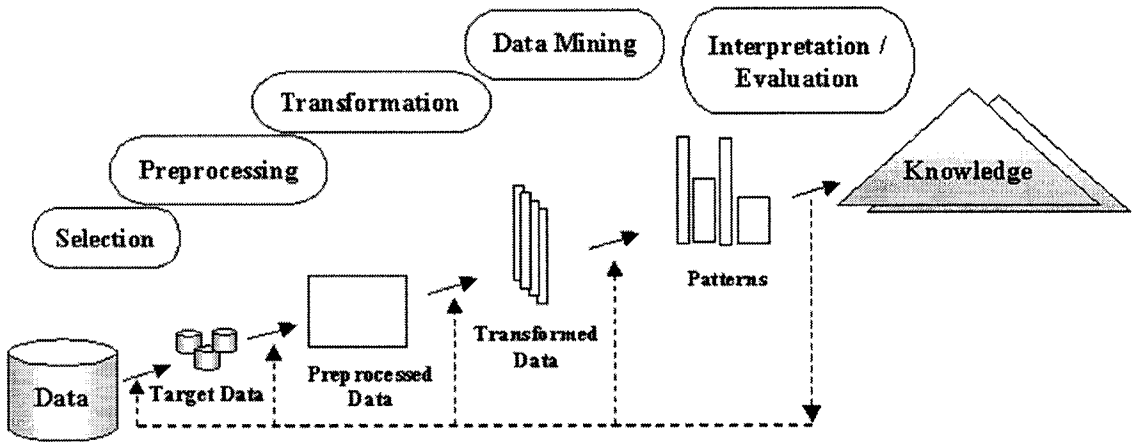


Figure 2.2. Data mining as a step in KDD (Fayyad et al., 1996).

Conceptually, data mining can be viewed as an embedded step in the process of *knowledge discovery in databases* (KDD) as illustrated in Figure 2.2 (Fayyad et al., 1996; Hand et al, 2001). The term KDD first appeared at a 1989 workshop to emphasize that knowledge "is the end product of a data-driven discovery" (Fayyad, 1997b, p.4). The KDD process starts with a large database or a data warehouse, from which the target data set is selected for analysis. As in statistical analysis, the selected data set has to be checked for missing data and other irregularities during preprocessing; then, any necessary transformation of variables is performed, including binning of continuous measures into category-like intervals. Finally, different algorithms of data mining are tried to develop an optimal single or combined model, which has to be evaluated and interpreted by experts before being used as extracted knowledge (Fayyad, 1996, 1997b).

Because data mining is meaningless without other steps in the KDD process, *data mining* is often used interchangeably with the term KDD.

Given the many data preprocessing methods and techniques offered by traditional statistics, out of the three basic stages in data mining (data preprocessing and exploration, model building or pattern definition, and model validation / verification), data miners center their major research interests and discussions on data exploration and model structure definition. Feature selection/extraction theoretically happens in the exploration stage; data mining algorithms are the essence of structural modeling and pattern discovery.

Feature Selection / Extraction

The size of a data set is determined by the number of variables and the number of observations. Variables are usually called attributes, features, or properties in data mining; observations are also called instances, cases, or records. In this paper, "variable" and "feature" are used interchangeably; so are "case", "instance", and "observation". For a large data set, both the number of variables and the number of observations can be enormously large. When the number of variables is large, data mining, as inductive learning from the observations, becomes increasingly difficult. First, unrelated or redundant feature dimensions in the data can slow down the learning algorithm dramatically. Second, the high dimensionality of feature space demands enormous computational resources and makes the analysis cost-inefficient. Third, the data mining learning algorithms may experience low prediction accuracy due to learning information from irrelevant variables. Finally, redundant features can lead to an overfit result with

poor generalizability. Hence, using an appropriate feature selection method to choose an optimal set of features is extremely important in data mining, especially with the consequence that the selected feature subset is the only source of information available for the subsequent learning algorithms of data modeling. Piramuthu (1999) summarized that:

> The goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a good chance that the information content in the set of features is low. On the other hand, if too many (irrelevant) features are selected, the effects due to noise present in (most real-world) data may overshadow the information present. Hence, this is a tradeoff which must be addressed by any feature selection methods. (p. 296)

Feature selection is defined as the process for obtaining a minimal set of representative variables that can retain the optimal salient characteristics of the data (Mitra, Murthy, & Pal, 2002). A wide variety of feature selection methods are in use including statistical, geometrical, information-theoretic measures, and dynamic programming. With some evaluation criteria, they all work to choose a smaller set of features by removing irrelevant and redundant features. The reduced number of features should be able to lead to more concise results and better comprehensibility (Liu & Motoda, 1998a).

Feature selection is a step interacting with the inductive learning algorithms in data modeling; together they make a difference in the results (models or patterns) as well as predictive accuracy, speed of learning, and interpretability. Therefore, it is essential to

take into consideration both the input data structure and the induction learning algorithms in the process of feature selection. As a matter of fact, a very common classification of feature selection methods is to differentiate the wrapper approach from the filter approach (described below) in terms of their dependency on the induction algorithms. The wrapper approach of feature selection is integrated with the data mining induction algorithm because it uses the algorithm as part of the function evaluating the feature subsets (shown in Figure 2.3). The filter approach (Figure 2.4) serves to eliminate the irrelevant and irrelevant features independently of the induction algorithms (Yuan, Tseng, Wu, & Zhang, 1999).

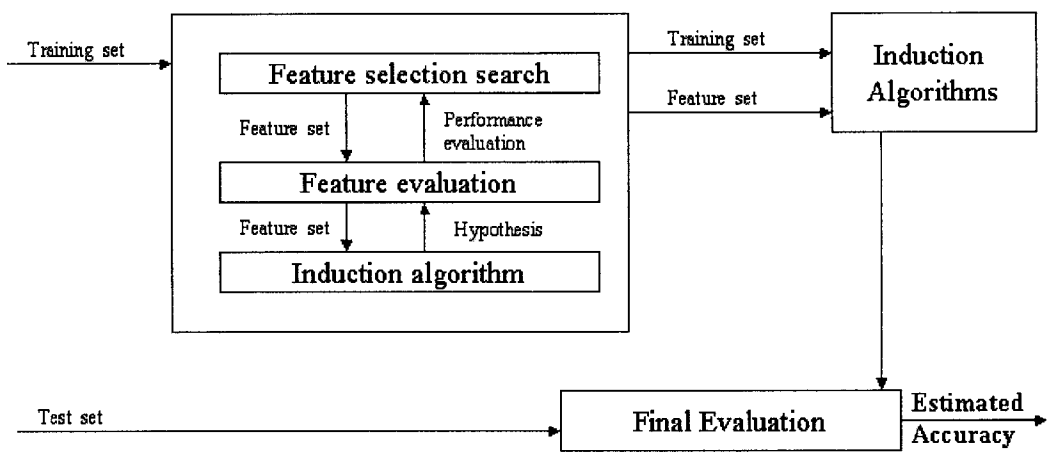


Figure 2.3. The wrapper approach to feature selection: The induction algorithm is used as a "black box" evaluation method in the subset selection (Liu, & Motoda, 1998b).

Wrapper feature selection. Wrapper feature selections are generally black-box operations that are "wrapped" into the induction algorithm. The induction algorithm is

repeatedly run on the data set using various feature subsets. Certain evaluation criteria are used to assess the performance of each subset, and the feature subset with the highest evaluation is chosen as the final set for the induction program to run with (Kohavi & John, 1998). Users can only know that the chosen subset has the highest estimated value on a performance criterion of the induction algorithm. A minor drawback is that the inductive and representational biases of a particular algorithm inevitably take a part in the selection of the final subset. Therefore, a particular subset of variables may not lead to the best predictive accuracy *vis a vis* other induction algorithms because of the dependency of feature selection on the induction algorithms.

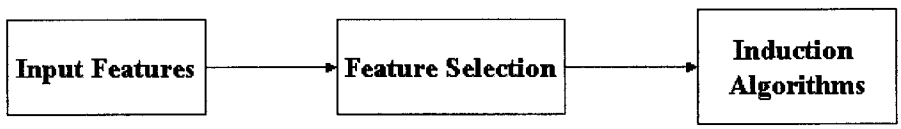


Figure 2.4. The filter approach to feature selection: The subset feature selection is independent of the induction algorithms (Liu, & Motoda, 1998b).

Filter feature selection. The filter approach of feature selection attempts to assess the merits of features from the data alone (Kohavi & John, 1998). The selection of features is done in the preparation stage of the data mining process, independent of the induction algorithms. Filter feature selection is open to a wide range of possible methods, and makes heavy use of statistical measures of variable relationships. A simple example is the multiple regression techniques used to select features that account for a significant

portion of predicted variance. Another example is a feature weighting system called *Relief* proposed by Kira and Rendell (1992). Relief selects features that are statistically relevant in terms of some distance between feature values. Implicitly, the method chooses a subset by evaluating features using a distance measure.

Although a filter feature selection is more transparent than the wrapper approach, the filter approach does not take into account the biases of the induction algorithms. In view of the fact that the optimal selection of features is not independent of the inductive and representational biases of the algorithms in most data mining cases, the total ignorance of such biases by filter feature selection may not lead to the optimal performance of the subsequent induction algorithm.

Experiments with artificial data (Kohavi & John, 1998; Piramuthu, 1998) have shown that feature subset selection can improve the data mining results and that a wrapper method should be preferable to a filter, despite the fact that the wrapper may cause over-fitting and require a large amount of computer resources.

Search strategy of feature selection algorithms. Liu and Motoda (1998a) introduced some representative feature selection algorithms in terms of their search strategies: exhaustive, heuristic, and stochastic. In an *exhaustive* search, the whole search space is completely covered: All the combinations among all features are evaluated by a consistency (or inconsistency) measure in order to find the optimal subset. An exhaustive search may not be computationally feasible for data with a very large number of variables in complicated algorithms. An *heuristic* search sacrifices the promise of an optimal subset for a quick solution by sequentially adding or removing features from a subset. A

simple example is to induce a classifier (e.g., a decision tree) and select the features used by the classifier. Different evaluation criteria are used for different heuristic search methods. One drawback of this approach is that the subset may only be near optimal as a tradeoff between efficiency and optimality. In contrast, to search for an optional subset without exhaustive search, a *stochastic* method tries to randomly generate subset features and make the selection based on certain utility functions. The genetic algorithms used for feature selection are examples of stochastic search methods. A potential problem with stochastic search is that it may result in a local optimum.

Feature selection methods are often used independently; the critical issue is which method should be chosen for a particular mining problem among many available ones. Usually some prior knowledge about data and application is necessary to make the decision with a good understanding of what each feature selection method can offer.

Besides feature selection, other approaches are also available to transform variables for the purposes of condensing information and reducing data dimensionality. Among them, feature extraction produces a smaller number of new features from the original set without losing important information. The extraction methods are originated in statistical methods for finding the intrinsic dimensionality of a data set: the minimum number of variables required to represent the data accurately (Liu & Motoda, 1998a). Some common feature extraction methods include PCA and self-organizing maps. With some efforts to decide on ideal parameters, these methods extract a smaller number of new features that are different from the original set in nature. Also, the representation of data is changed after the feature extraction so that techniques of visualization and

classification can be conveniently used (Bursteinas & Long, 2000; Liu & Motoda, 1998a).

For a thorough discussion of feature selection, extraction, and more, please refer to Liu and Motoda (1998a, 1998b).

Model Building and Pattern Definition

Assuming that an optimal subset of features is available through some feature selection or extraction methods, the next question is what data mining algorithms to use to extract previously unknown, valid, and actionable information from the data.

Unlike traditional statistics, which try to explain the overall data structure with a fitted model, data mining offers techniques to discover overall structures as well as local patterns existing in subsets of a big data set. Thus, for discussion on specific data mining techniques, a distinction is necessary between "model" and "pattern". According to Hand (1998), a model is "a global representation of a structure that summarizes the systematic component underlying the data or that describes how the data may have arisen" (p. 116). Patterns are only local structures compared to the comprehensiveness of models.

Both model building and pattern detection in data mining are considered empirical because they are processes of seeking either global or local relationships without basing them on any underlying theory. In terms of the specific tasks performed, Hand et al. (2001) grouped data mining techniques into five categories: exploratory data analysis, descriptive modeling, predictive modeling, discovering patterns and rules, and content retrieval.

Exploratory data analysis. As the name indicates, this function in data mining is a close relative of EDA, the branch of statistics initiated by Tukey in the 1960s. Using graphical display methods, data miners can explore the data for unexpected trends, patterns, and anomalies in an interactive manner. The methods adopted from EDA make visualization of small and low-dimensional data very effective. In case of a large number of observations, the indiscernible cloud of points can be reduced to a "lower resolution" with the cost of possibly missing important details (Hand et al., 2001). When the dimensionality is higher than three or four, statistical methods (e.g., projection) can help to produce informative low-dimension presentation. Examples of advanced graphical exploratory analytic techniques in data mining software include brushing, data smoothing, function fitting and plotting, overlaying and merging of multiple displays, splitting and merging subsets of data in graphs, aggregating data in graphs, and more.

Descriptive modeling. Part of the data mining mission is to offer digested descriptive information that can help the user to gain a better understanding of the overall data structure. Examples of descriptive modeling tasks include density estimation (probability distribution of the data), dependency relationships among variables, and cluster analysis and segmentation (Hand et al., 2001). The most successful applications of descriptive models are the clustering techniques in cluster analysis and segmentation.

Both cluster analysis and segmentation use clustering techniques originated from statistics. A clustering technique separates data items into subsets that are similar to each other without any output measure. According to Hand et al. (2001), in data mining, the cluster analysis is used to discover "natural" groups in data, whereas segmentation is a

way to partition the instances into a predefined number of homogeneous groups so that similar cases are in the same group.

With some type of distance or similarity measures, most clustering functions in data mining are a powerful and automated version of the clustering techniques from the multivariate methods in statistics. More than statistical approaches, however, the techniques in data mining software include score functions which help to automatically select the optimal model out of numerous partitions of data (Hand et al., 2001). K-means cluster analysis (KMC, a type of cluster analysis with iterative partitioning) and some form of ANNs (e.g., self-organized maps) are the most commonly seen methods of clustering in data mining software.

Predictive modeling. Predictive modeling is intended to estimate the values of some measure based on other variables in the same database or data set. If the measure being predicted--the output--is a continuous variable, the prediction is a regression problem; if the output is a categorical measure, it turns into a classification problem (Fayyad, 1997b; Hand et al., 2001). A wide variety of techniques is available for predictive modeling. Because linear multiple regression combined with non-linear transformation of independent variables (inputs) works so well for predicting continuous variables, data mining can only claim its advantage in classification problems for predicting the most likely state of a categorical variable (the class) using some machine learning algorithms along with statistical techniques including logistic regression and discriminant analysis. Examples of classification techniques are decision tree, ANNs, Bayesian networks, and other rule-based classifiers.

The decision tree algorithms used in data mining are an automated version of the tree structures used in statistical decision analysis. Some human input in statistical decision process is replaced by embedded evaluation criteria in model building and selection. The automation is made possible by the techniques introduced from machine learning, AI, and computer sciences. Another popular prediction algorithm not commonly used in statistics is ANNs.

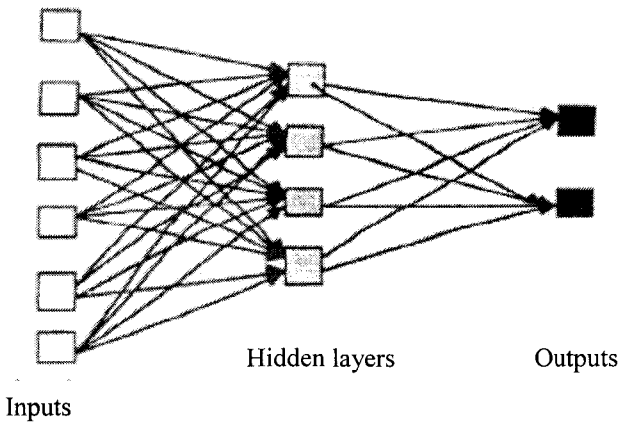


Figure 2.5. An example of simple-structured ANN.

The interest in ANNs started from the concept of an "artificial neuron" that mimics the process of a neuron in the human brain as early as in the 1940s in the field of AI, but did not come into favor until 1980s when studies showed that algorithms of perceptrons with nonlinear thresholding units connected in multiple layers can serve both clustering and predictive computations (Groth, 1999).

In an ANN, processing elements (PE) are used to mimic neurons in the human brain. One PE is limited in ability, but when many PEs (nodes in inputs, hidden, and

outputs layers, as shown in Figure 2.5) are connected, they together form an intelligent model. The strength of the connections between PEs is called weight, which can be modified with a mathematical method--the learning rule--during the iterative induction process of the network. During the process, PEs summarize the transformed data, and the connections between PEs receive different weights until the neural network satisfies some stopping criterion. In other words, a network tries various weighting formulas for modeling the input data in a way it sees fit (Mitchell, 1997).

ANNs provides a flexible approach to approximating functions for real-valued, discrete-valued, and vector-valued data; they are among the most effective methods currently available to interpret complex real world sensor data and other similar problems. The algorithms of ANNs are robust against errors and other noises in training data, and they are also applicable to problems such as interpreting visual scenes, speech recognition, and learning robot control strategies (Mitchell, 1997). One drawback is that, unlike statistical models with straightforward functional equations, ANNs take a black-box approach that is not friendly for human understanding of the internal structures (Mitchell, 1997).

Although the idea of ANNs was first introduced in statistics, it failed to achieve widespread use as a statistical technique because of its high flexibility and low interpretability. Currently, assisted by high computational power and machine learning algorithms, ANNs are used extensively in data mining for both predictive modeling and descriptive modeling.

Discovering patterns and rules. Pattern detection is a component of data mining research that has no "official" counterpart in statistics. Patterns are found by sifting the data to look for co-occurrences of particular values on particular variables (Hand, 1998). A well-known pattern detection algorithm in data mining is called the association rule analysis, which can be best explained using an example of the Market Basket (MB) analysis. The MB studies the aggregate associations among different items sold in catalogs or at a retail store. Taking the point of sale transaction data, the MB analysis provides information and recommendations in the form "if customers buy product A they also tend to buy product B, x% of the time" to exploit product associations and customer-purchase behaviors (Groth, 1999).

One problem with the basic association rule analysis is that it does not distinguish between random and non-random item associations with its regular support-confidence framework. To make this crucial distinction, a dependency framework is suggested for advanced association analysis that looks at associations where the percentage of co-occurrence is significantly different from random chance. A χ^2 test of statistical independence is used for this purpose in most association algorithms (Groth, 1999).

Pattern detection is also very useful in finding anomalies, a problem similar to outlier detection studied by statisticians. It is difficult to decide what constitutes truly unusual behaviors in the context of normal variability, but domain knowledge and expert interpretation encoded in data mining algorithms can help a great deal.

Content retrieval. Content retrieval is a kind of task most commonly used for text and image data sets. The web search engine *Google.com* is a typical application of

content retrieval. Another example is the automatic spelling check in word processing programs. For text search, the retrieval is to use a set of keywords to find relevant documents. For image search, a sample image or a description of image is necessary. Because text and images are not numerical data, the definition of similarity and the search strategy are two critical elements in content retrieval algorithms (Hand et. al., 2001). In statistical data analysis, most of the data being analyzed are numbers. Content retrieval is a new subject exploited by data miners.

Machine Learning in Data Mining

As well acknowledged, the core of data mining techniques includes many statistical procedures, such as regression analysis, classification methods, projection pursuit, numerical taxonomy, multidimensional analysis, stochastic models, time series analysis, nonlinear estimation techniques, and any methods for summarizing data (Hand, 1998; Hand et al., 2001). It may seem like data mining is nothing more than statistics in numerical data analysis after the brief review of some of the specific techniques and procedures. Nevertheless, use of statistics is only one side of data mining. The other side is machine learning, a subdivision of AI that has contributed substantially to the automation and computational efficiency of the data mining process.

Machine learning is a field of research first conceived around the 1960s with "the bold objective to develop computational methods that would implement various forms of learning, in particular mechanisms capable of inducing knowledge from examples or data" (Kubat, Bratko, & Michalski, 1998, p. 3). Profiting from an AI idea that

Program = algorithm + data + domain knowledge,

machine learning employs learning systems that acquire high-level concepts and /or problem-solving strategies through examples (input/training data) in a way analogous to human knowledge induction to attack problems that lack algorithmic solutions or have only ill-defined or informally stated solutions (Kubat et al., 1998). The notion of "concept" is as vital to machine learning as numbers to mathematics.

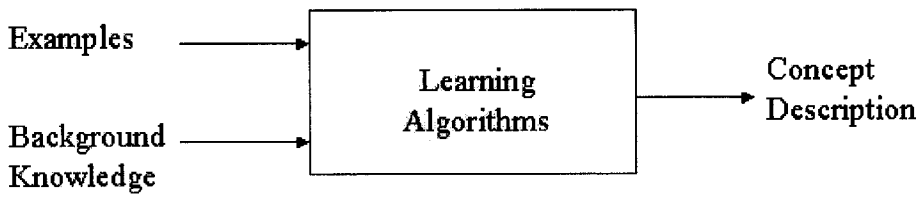


Figure 2.6. Machine learning task (Kubat et al., 1998).

The general framework for machine learning is depicted in Figure 2.6. The output of a learning system is a description of a given concept from a set of concept examples provided by the input data and from the background knowledge.

As with statistics, machine learning aims at building models to understand and interpret a set of observations or to be able to predict properties of new objects. However, with embedded background knowledge, machine learning uses inductive concepts learning to create descriptions of rules through ANNs, decision trees, or genetic algorithms that do not assume any parametric form of the appropriate model. The distribution-free techniques allow machine learning to use a toolbox approach to identify the appropriate model structure directly from the input data rather than doing model fitting with probabilistic assumptions.

Because of the inductive nature of the learning process, the correctness of the qualitative descriptions learned from the input data improves with the number of instances. Consequently, machine learning methods typically require much larger data sets than parametric statistical analysis does, and in order to avoid optimistic bias (i.e., overfitting), the models need to be validated using independent sets of observations.

Machine learning algorithms are computer programs that learn to adapt to environment by "understanding" the input data and achieve a desired outcome based on the learned knowledge. From a machine learning perspective, no matter how complicated the algorithms used, models always stay in the scope of two general missions: prediction and discovery. Prediction depends on supervised learning algorithms, in which the training data set acts as a "teacher" who "supervises" the learning and induction of the model structures. Prediction techniques are analogous to dependence analysis in statistics because both have output measures. Discovery depends on unsupervised learning algorithms, which refer to the collection of techniques in which the pattern identification and model building are done without using dependent variables and any training data.

In supervised learning, an induction algorithm is typically presented with a set of training data, where each of the records or instances is described by a set of features (the predictors or independent variables) and a class label (the predicted or dependent variable). The task of the induction algorithm is to induce from the training data a classifier rule or structure that is useful in grouping future cases. The classifier is essentially a mapping from the space of feature values to the set of class values. As an outcome of "learning" from data, the prediction is a process of classifying new cases with

the induced classifier. In some machine learning techniques, besides learning from the data, the algorithmic structure used for prediction can also be supplied by human experts in terms of constraints, such as independence, conditional independence, higher order conditions on correlations, and so on. Prediction models in data mining, including regression and classification algorithms, are in a general sense members of supervised-learning family because the models are built from a training data set that has known output values, and the induced model can be used to classify new cases.

Unsupervised learning algorithms dominate discovery in machine learning. Discovery, perceived by some people as looking for nuggets of information in a mountain of data, is used to model the relationships among a set of variables without making the distinction among inputs and outputs. The goal is to look for systematic relationships, for example, correlations among variables and/or similarities among subjects. Discovery is suspected by some as synonymous with "fishing" or "data snooping", which are terms used to describe the process of trawling through data to look for previously unknown patterns (Fayyad, 1997b; Hand, 1998). Association rule and segmentation are typical examples of data mining methods that make use of unsupervised learning algorithms from machine learning.

Machine learning uses computer programs and AI algorithms for automation and computation-intensive model selection, but it does not shun taking advantage of parametric statistical assumptions when they hold. As an inductive means prone to concept descriptions, machine learning prefers handling categorical data (e.g, nominal and ordinal variables) more than traditional statistics. When dealing with continuous

values, a majority of machine learning algorithms need to discretize the values in order to proceed.

Statistical techniques play important roles in data mining. However, with embedded background knowledge, machine learning takes the data modeling and interpretation in data mining one step further by summarizing variable relationships at a conceptual level and offering causal explanations of dependencies; by describing regularities in logic statements; by offering descriptive statements about clusters of entities and hypothesizing reasons for the entities being in the same clusters (Michalski & Kaufman, 1998). In short, equipped with a substantial amount of background knowledge and machine learning functions, data mining automates the data analysis process and takes over some tasks of result interpretation and knowledge discovery routinely performed by human analysts in traditional statistical analysis. The types of algorithms and properties data mining inherited from machine learning makes itself an approach of data analysis quite different from traditional statistics. The following section offers more discussion on the theoretical similarity and differences of statistics and data mining.

Comparison of Traditional Statistics and Data Mining

The emergence of data mining as a new research area is to satisfy the need for turning enormous collections of data into useful task-oriented knowledge. Although data analysis techniques available in traditional statistics have been widely used in extracting information from quantitative data and solving practical problems, they have inherent limitations when dealing with very large volumes of data (Hand et al., 2001; Kubat et al., 1998; Wegman, 1995). Because the size of data sets is believed to be a critical reason for

the existence and popularity of data mining, for further discussion it is necessary to have an unambiguous profile of data size.

Table 2.1

The Huber Taxonomy of Data Set Sizes (modified from the table in Wegman [1995])

Descriptor	*Data set size (bytes)*	*Storage mode*
Tiny	10^2	Piece of paper
Small	10^4	A few pieces of paper
Medium	10^6	A floppy disk (1 MB)
Large	10^8	Hard disk, compact disk (CD, 2.5 inches)
Huge	10^{10}	Multiple hard disks, e.g. RAID storage
Ridiculous	10^{12}	Robotic magnetic tape storage silos

Size Definition of Data Sets

Huber (1994) proposed a system of taxonomy of large data sets as shown in Table 2.1 (as cited in Wegman, 1995). The scale used for this size definition is a computer storage measure "byte", which is roughly the memory space taken by a single letter or a single digit. Since data are often arranged in a matrix form of r rows and c columns, the size of data set can be understood as $n = r \times c$. The taxonomy by Huber is helpful because it quantifies the meaning of *tiny, small, medium, large,* and *huge* data sets. Based

on this system, most statistical and visualization techniques are computationally feasible with tiny and small data sets (Wegman, 1995).

For data sets of different sizes, Wegman (1995) estimated their numbers of operations in some common statistical algorithms of various computational complexities, and evaluated the computational feasibility of those algorithms on a regular PC. Based on his estimates, Wegman argued that data sets of large and beyond required intellectual attention, and made the statement that

> "It is probably axiomatic that as the size of the data set increases, so does its complexity... Applying traditional statistical methods to what in Huber's taxonomy are medium, large or huge data sets is doomed to failure. Homogeneity is almost surely gone. Any parametric model will almost surely be rejected by any hypothesis testing procedure. Fashionable techniques such as bootstrapping are computationally too complex to be seriously considered for many of these data sets. Random subsampling and dimensional reduction techniques are very likely to hide the very substructure that may be pertinent to the correct analysis of the data." (p. 292).

Almost a decade after Wegman's estimates, the regular PCs are now able to deal with medium (and occasionally large) data sets with regular statistical software; however, the number of data sets of large, huge, and beyond has also dramatically increased with the rapid data collection supported by modern computing resources. Computational complexity and feasibility remain problematic in this trend, and major points of Wegman's argument still hold for large-scale data sets that have size of 10^8 or larger.

The solution proposed by Wegman (2000) for analyzing very large data sets was nonparametric analytic techniques and graphical exploratory analysis tools, ideas that have been unconventionally realized in data mining software in recent years.

Traditional Statistics and Data Mining

Both statistics and data mining serve to study data structures and generate useful knowledge, and data mining employs many statistical techniques in its algorithms. What are the differences of data mining from statistics that endow it with the power of dealing with very large data sets?

Part of the reason is related to the fact that a large data set can have a large number of cases and/or a large number of variables. If a data set has millions of cases but only a few variables, even if the algorithmic scale in terms of hardware resources and computational time may be overwhelming to statistical computations as Wegman (1995) argued, traditional statistics can resort to sufficient statistics and sampling to tackle the problem and to fulfill the goal of data modeling (Hand et al., 2001). However, it turns more problematic when the number of variables is large in a data set. The "curse of dimensionality" presents many challenges because the necessary computational resources increase exponentially with the number of variables (Hand et al., 2001). In a high dimensional space defined by a large number of variables, it is difficult to find accurate estimates of probability densities. To make the matter worse, most data reduction methods in statistics assume linear relationships or calculate geometric distances among variables. Linearity may not be realistic when many categorical variables exist; "nearest"

neighbors in high dimensions could be very far even if there is a way to actually calculate the distances.

To deal with high dimensional data, data mining takes advantages of both inductive machine learning algorithms and deductive statistical methods when applicable. Machine learning algorithms can be used for optimum subset search; together with the statistical learning theory, different functions are available to sift the variables for most useful ones. For instance, depending on the circumstances, exhaustive or stochastic search is used to find the "best" subset of variables for a particular data modeling purpose. Meanwhile, statistical techniques can be used to quantify variable relationships and, when appropriate, are also used to reduce the number of variables by combining highly correlated original variables into fewer new measures.

Overall, the deductive reasoning in traditional statistics is sufficient and efficient when analyzing tiny, small, and medium data sets; but when a data set is large with many variables, inductive reasoning in data mining becomes more flexible and adaptable without the need to make probabilistic assumptions. To improve the quality of inductive conclusions, a large number of observations is always required, which is another reason why data mining accommodates and works better with data sets of a large number of records.

The preference of traditional statistics to small and clean data sets is determined by their deductive nature in parametric data modeling based on restricted distribution assumptions. Based in probability theory, statistical methods are used to estimate the parameters of the population distribution from sample data, taking into account the

chance factors associated with sample selection (Minium & Clarke, 1982). Clearly, sampling is an important step to ensure that the data collected are representative of the population and so a part of the endeavor in statistics is devoted to experimental design and data collection. Carefully designed experimental or surveying procedures lead to sample data that are mostly clean and well structured. Because data collection is expensive, the sample size is usually no bigger than necessary to keep the significance tests sensitive to meaningful differences. In such cases, data analysis is intended primarily to answer some preconceived questions.

On the other hand, according to Hand (1998, 2001), data mining is a secondary and retrospective data analysis that comes into play when the data have already been collected. Other than the preference for large data sets, data mining techniques are different from traditional statistics in the types of data they are used with. Without any control over data collection, data mining techniques are mostly applied to analyze observational data from convenient or opportunistic samples, or that constitute the population. As a result, data mining faces problems including noisy data, irregular distributions, and loose structure. Unable to estimate probability distributions in high dimensionality for determining the model *a priori*, data mining emphases searching for suitable model structures; the process is exploratory rather than confirmatory.

The deductive and inductive natures of statistics and data mining, respectively, also determine the types of variables preferred. Mathematically processing sample data, traditional statistics have more elasticity working with continuous data although categorical variables can be handled. Data mining algorithms, due to the heavy use of

machine learning rule inductions, favor categorical variables for inducing conceptual statements from input data (Hand et al., 2001). When continuous variables are included in the data mining analysis, they need to be binned into discrete intervals. For instance, statistical regression uses the *actual* value of a continuous variable (e.g., age) in the analysis, but for data mining purpose the same measure has to be transformed into 1 (0-5 years old), 2 (6-10 years old), 3 (11-13 years old), and so on. The downside of binning continuous variables in data mining is the loss of information and measurement accuracy.

With a small number of variables, the functional model structure is predetermined in traditional statistics, meaning that the role of each variable is clear in the structural model (e.g., ANOVA and multiple regression). The analysis is conducted to make the best estimates of the model parameters for statistical inferences. The estimation of parameters is supported by some distributional assumptions (e.g., normality and homogeneity) associated to the structural model. In the case that some of the variables do not confirm the probabilistic assumptions, data transformation can be used.

In data mining, predetermination of model structure is not feasible because of the data complexity with a large number of variables, and because the huge record number makes it difficult to examine or assume probabilistic distributions (Hand et al., 2001). Thus, the analysis has to start with model selection, a process in which different structural models are evaluated to find the one that best matches the data by criteria called score functions. Estimating parameter values is done as a part of structural model building. In essence, the emphasis of data mining analysis is structural inference rather than the statistical inference.

In statistical analysis, data analysts need to monitor closely the deductive estimates of model parameters from the sample data: selecting analysis procedures, interpreting output, formulating mathematical models, and reporting final findings. However, the analysis process is much more automated in data mining, and even the use of statistical procedures is fully embedded in the machine learning algorithms powered by sophisticated computer programs. Once the input data are ready, the software programs may carry out almost every analytical steps without human intervention: determining the underlying structural or functional form of the data, evaluating the quality of a candidate model, optimizing the search over different model/pattern structures, handling the data access efficiently during the search and optimization, and even making statements and conclusions based on the findings and the encoded background information.

Furthermore, the outputs of data mining and statistics are different. In statistical analysis, a model is usually expressed in forms of mathematical functions; whereas in data mining, the output is seldom such functions. It can be a graph (e.g., a tree structure or a prediction network) companied by descriptive statements. For example, when making a prediction, a multiple regression equation may lead to a predicted value of 60 with a standard deviation of 3; whereas in data mining, the prediction is most likely made in a statement like "there is a 70% probability that a simple-functioned microwave made in the U.S. is priced between $58 to $65".

As introduced previously, many statistical techniques used in data mining (e.g., decision tree, projection pursuit) have been around since 1960s, but they never became

popular in statistical applications as they now are in data mining. Out of many possible reasons, a major one is that the priority of statistical data analysis is model interpretability with the objective of understanding variable relationships and finding answers to some predefined research questions. Statisticians are willing to sacrifice some performance or accuracy for a clear and coherent model structure that underlies "data-generating mechanism" (Elder & Pregibon, 1996, p.10), so they do not hesitate to retreat back to additive models with distributional limitations when nonlinear and sequential structures are too flexible to interpret.

Data miners seem to emphasize different qualities in the models they seek: optimal accuracy of a model is almost always the most desirable; black-box approaches and opaque model structures are all tolerable as long as the findings can bring valuable mechanism and knowledge into future applications, including predictions and decision making. Therefore, some very complicated statistical algorithms, such as ANNs and polynomial networks, are used widely in data mining because of their robustness against noisy data and flexibility in model formulation.

Data mining also has some unique research interests that are not shared by statistics. For instance, data mining not only works on numerical data of very large volume, but also extends the analysis to text and multimedia data. Another difference worth noting is the local pattern detection in data mining, a useful function not discussed much in traditional statistics. An example of this difference is how mining, anomalies may be studied as a type of potential local patterns of special interests (Aptė, 1997). Some of the anomaly detections have been successfully applied to business anomalies

Table 2.2

Differences between Traditional Statistics and Data Mining

	Traditional statistics	Data mining
Data Collection	Experimental or by survey	Retrospective analysis of observational data in databases
Characteristics of data	Few variables; clean data	A big number of variables; noisy data and convenient sample
Data size	Typically small to medium	Typically medium to ridiculous
Favorite data type	Continuous	Categorical
Model	Stochastic modeling with parameter estimation	Model selection and parameter estimation; pattern and anomaly detection
Output	Usually mathematical functions	Graphs, conceptual rules in descriptive statements
Priority	Model interpretability	Application; model accuracy
Techniques	Deductive statistics based on frequent probabilistic assumptions	Multivariate statistics embedded in inductive machine learning algorithms, less model assumptions
Process	Human intervention	Automation with background knowledge
Goal	To answer predefined research questions	To search for useful, previous unknown pattern & structures
Objectives	Statistical inferences, variable relations & predictions	Structural inferences, predictions and decision making

are treated in data mining and statistics. In statistics, the anomalies are often called outliers, and checked for clerical accuracy or removed from the analysis. In data management, for example, to catch credit card fraud in customer protection. Pattern detection can have potential applications in areas including education. For instance, Oracle Corporation helped some universities finding factors related to student dropout based on the student databases with their data mining software.

A summary list of the above-discussed differences between traditional statistics and data mining is available in Table 2.2.

The Reasons for Data Mining in Educational Research

The major differences between statistics and data mining show that although the two areas overlap, they cannot replace each other. Data mining has obtained popularity because it can efficiently extract information from a very large amount of data to guide decision-making. Statistics provides many powerful tools in dealing with relatively small and clean data sets for explaining variable relationships and making inferences. The motivation for discussing the applicability of data mining in educational research is that the number of large databases and large data sets are dramatically increasing with the advances of modern computing technology; whereas a thorough review of literature indicated that techniques that have been used in analyzing data of large data sets, especially those with a large number of variables were very limited: some studies offered only descriptive information or simple comparisons of means, others conducted the analysis in a confirmatory style by selecting a part of data from a large data set. It is unfortunate that some hidden values stay uncovered due to the lack of appropriate tools.

With practical problems presented by large volumes of data to traditional statistical methods, efficient handling of massive data sets is emphasized in data analysis procedures; even statisticians (e.g., Thompson, 1997) who distrust the model selection approach sense the need to develop statistical methods to incorporate high computer speed. So far, data mining has been the most successful example of the marriage between statistics and computer science, connected by machine learning and other related fields. Supported by strong computing power, the exploratory nature of data mining has the potential to discover values from large volumes of data and to stay robust against loose but complex structure.

In the area of educational research that is accustomed to using traditional statistics, data mining as a new approach is worth studying in the following aspects. First, for functions that can be completed by both statistics and data mining (e.g., prediction), what do they do differently to fulfill similar data analysis goals? Second, what can data mining offer to educational research, especially for analyzing large data sets, that are not readily available in traditional statistical approaches?

Analysis Procedures

This section covers the specific statistical and data mining techniques used in the study to provide a better understanding of the analysis procedures. In addition to the statistical variable reduction methods, multiple regression and the BBN with wrapper feature selection are introduced as prediction models. However, the discussion is by no means complete and thorough. If interested, readers can seek more information about specific topics in cited references.

Variable Reduction

When dealing with data sets of a large number of variables, both statistics and data mining seek to simplify data structure with variable reduction or feature selection/ extraction techniques, respectively, before any data modeling and structural inference are performed. Statistics offer a number of choices to simplify the variable space; several of them--PCA, EFA, KMC, and multidimensional scaling (MDS)--are introduced for being part of procedures used in this study. Feature selection/extraction in data mining has been introduced early in this chapter. Only the wrapper feature selection is used as an embedded component of the BBN prediction model; the technical details are discussed as a part of the learning algorithms later in this chapter.

If variables in a large data set contain redundant or irrelevant information, they present a more serious problem to statistical analysis than to data mining. To overcome this difficulty, quite a few techniques have been developed for variable reduction and dimensional simplification. Among them, PCA, EFA, KMC, and MDS are briefly introduced here as part of the techniques used in this study.

Quantification of Variable Relationships

Quantification of variable relationships is the first step in order to decrease the complexity of the dimensional structure of a variable space. Relationships between variables can be measured by either their similarity or dissimilarity. In statistical terms, variable similarity can be assessed with proximity measures (in which large values mean objects are similar) and dissimilarity with distance measures (in which large values mean objects are dissimilar). The most popular proximity measure is the Pearson product-

moment correlation, which is appropriate for linearly related variables on interval or ratio scales. However, in practice, the application is often extended to ordinal scale (Jones & Sabers, 1992). For variables on nominal scales, associations can be measured with χ^2 statistics or ϕ (the correlation coefficient of nominal data).

There are a few distance measures suiting different requirements. The most straightforward way of computing distances between objects in a multi-dimensional space is the Euclidean distance, which simply is the geometric distance of data points in a *p*-dimensional space. The Euclidean distance is computed as:

$$E\text{-}distance(x,y) = \sqrt{\sum_i (X_i - Y_i)^2}, \qquad (2.15)$$

where $i = 1, 2, \ldots, P$.

Other distance measures include squared Euclidean distance, city-block (Manhattan) distance, Chebychev distance, and so on. Generally, variables need to be on comparable scales to be eligible for most of the distance calculations. When they are not, data are often standardized before fed into the distance formula for a meaningful result.

Techniques using different measures of variable relationships may be sensitive to different properties of variable groupings such as density, variance, dimension, shape, and separation (Aldenderfer & Blashfield, 1984). For example, methods using correlation coefficients as the similarity measure emphasize shape, whereas KMC tries to minimize the variance within each cluster. For choosing a proper variable grouping method, it is always recommended to consider the nature of the classification, the variables included, and the similarity measures used to quantify the resemblance between variables.

PCA, EFA, KMC, and MDS are popular multivariate techniques for simplifying complex data structure; they can be used together to examine the data from different angles given that they use various similarity or dissimilarity measures, have different underlying assumptions, and take different approaches in grouping variables. When several different techniques are used to look for a common parsimonious explanation of a data structure, the potential biases associated with each individual approach can be alleviated or even avoided, and the uncovered dimensional structure can be more stable and robust. On one hand, the operational differences of these techniques can be reflected by the disagreements of their output structures; on the other, the resemblance of the parsimonious outcomes produced by different techniques can be a strong evidence of a stable and consistent intrinsic data structure independent of analytical methods.

The discussion below compares these four variable reduction/classification techniques.

PCA

According to Dillon & Goldstein (1986), PCA transforms the original set of variables into a smaller set of principal components (PC)--uncorrelated linear combinations of the original variables--that account for most of the variance of the original set of variables. The goal is to compress the data so that as much variation in the data can be explained by as few PCs as possible.

The extracted PCs are linear combinations of the original variables in the form

$$PC_{(i)} = w_{(i)1}X_1 + w_{(i)2}X_2 + ... + w_{(i)p}X_p \ , \qquad (2.16)$$

where p is the number of original variables, the weights $w_{(i)1}$, $w_{(i)2}$, …, $w_{(i)p}$ are chosen to maximize the ratio of the variance of $PC_{(i)}$ to the remaining total variation, with the constraints that $\sum_j^p w_{(i)j}^2 = 1$ for each linear combination and that all $PC_{(i)}$ are orthogonal to each other. Consequently, the first extracted component, $PC_{(1)}$, accounts for the largest portion of the total variation in the data. The variation of the original data accounted for by each of the extracted PCs decreases sequentially.

The PC extraction starts from transforming the raw data into the correlation or covariance matrix of the p variables. Since PCA is scale sensitive, the correlation matrix is desirable when the scaling is arbitrary. One inevitable question is how many PCs to retain. If fewer orthogonal PCs are kept to represent the main structural feature of a multivariate data set, the data can be better understood and easily graphed. However, important information may be lost if too few PCs are retained.

PCA can be used to choose an optimal subset of original measures. When some of the variables feature weakly in the first few PCs, they might be uninformative or only carry redundant information. They can be discarded from future analysis, but it is a subjective decision (Krzanowski & Marriott, 1994).

EFA with Factor Rotation

Without any *a priori* hypothesis other than linearity, EFA "attempts to simplify complex and diverse relationships that exist among a set of observed variables by uncovering common dimensions or factors the link together the seemingly unrelated variables and consequently provides insight into the underlying structure of the data"

(Dillon & Goldstein, 1984, p. 53). EFA is a common data reduction technique that usually works with the correlation matrix of observed variables to extract a number of latent or common components called factors. The observed variables can be approximately expressed as linear combinations of the factors. If X_i stands for the observed measures, and $F_1, F_2, \ldots, F_p$ stand for the p extracted factors, their relations can be written as

$$X_i = a_{1i} F_1 + a_{2i} F_2 + a_{3i} F_3 + \ldots + a_{pi} F_p + U_i, \qquad (2.17)$$

where U_i is the unique portion of a variable not related to the extracted common factors. The weights $a_{1i}, a_{2i}, a_{3i} \ldots a_{1i}$ are factor loadings. They differ from variable to variable because the relationships of observed variables with the common factors vary. A high factor loading means a strong relationship between an observed variable and a common factor.

The number of extracted factors can be as large as the number of observed variables in EFA. However, to search for a parsimonious structure underlying the data, only the most important factors are kept for interpretation. Again, the problem is how many factors to keep because there are no criteria for "important factors." As in PCA, one commonly used method is to examine the eigenvalues produced from the correlation matrix. An eigenvalue can be interpreted as the share of total variance explained by a given factor. In some cases, only factors with eigenvalues greater than a specified value are accepted; in other cases, eigenvalues are plotted against the corresponding factors in a "scree" plot. If an outstanding change exists in the slope of the plot, it suggests that factors to the left of the point be retained.

Different algorithms for factor extraction are possible. For instance, factors can be extracted with the principal component procedure. EFA with PC extraction defines factors such that each factor accounts for the maximum possible amount of the variance contained in the set of variables being factored. This procedure is very similar to PCA (Dillon & Goldstein, 1986). The major conceptual differences between EFA with PC extraction and PCA can be described as follows. PCA is used to reduce a set of observed variables into a smaller number of artificial variables called principal components (PC) that are dependent on variables. EFA is used to identify the number and the nature of the underlying factors that are responsible for the covariation in the observed variables, so the variables are dependent on the extracted factors. Also, error is an explicit part of EFA modeling (U_i in Equation 2.17), but it is not the case in PCA. In addition, unlike PCA, which can begin with either a correlation or covariance matrix, EFA usually works with a correlation matrix. Computationally, PCA works with a regular correlation or covariance matrix, but some factor extraction methods in EFA use a correlation matrix in which the diagonal elements (1s) are replaced by the respective variable's communality estimate. The communality of a variable is defined as the portion of its own variance that is attributable to the common factors (Jones & Sabers, 1992).

Another difference between EFA and PCA is the uniqueness of solution. PCA always produces a unique result, but EFA provides the solution in a different manner. Imagine in a multidimensional space, the selected common factors can be considered as axes; the factor loadings are the coordinates determining the location of individual cases in the space. The factor analysis solution allows the axes representing the selected factors

to rotate to different positions, and every position is a unique EFA solution given that the factor loadings change with the position of the axes. Rotation is a positive feature of EFA because it facilitates a more meaningful and usually simpler interpretation of the extracted factors.

Rotation can be either orthogonal or oblique. In orthogonal rotation, the perpendicular orientation between factors is preserved after the rotation. With oblique rotation, the factor axes can be rotated independently and so the perpendicularity may not exist after rotation. Multiple rotation methods are available for both orthogonal and oblique rotations. The definition of a "simple structure" is the key to choosing among the rotation methods; prior knowledge about the structure can also help. For exploratory purposes, different rotations can be tried to look for the one with a simpler interpretation. For more discussion of PCA and EFA, please refer to Dillon and Goldstein (1986).

Given the fact that EFA uses linear correlation as a measure of variable relationships (similarity), in order to avoid potential biases in variable grouping, its result should be cross-validated with alternative techniques (e.g., cluster analysis) that quantify variable relationships in a different approach.

KMC

Cluster analysis refers to a class of exploratory data analysis techniques for solving classification problems. The objective is to sort cases or variables into groups (clusters) so that the degree of association is strong between objects within the same cluster and weak between objects belong to different clusters.

Cluster analysis techniques can be used collaboratively with EFA in studying variable relationships because they do not assume linear relationships among objects or variables. Instead, cluster analysis methods work with distances or dissimilarity measures to group objects based on their closeness in space.

Everitt (1980) vaguely described clusters as continuous regions of a space containing a relatively high density of points that are separated from other such regions by space containing a relatively low density of points. The definition left enough room for the fact that cluster analysis actually encompasses a number of different classification algorithms. In general, hierarchical and nonhierarchical techniques are two major categories of cluster analysis methods.

Hierarchical techniques always classify variables into non-overlapping clusters, but they have a major drawback: the partitioning process works through the data only once and so there is no chance for modification of a poor initial partition in subsequent steps. Nonhierarchical cluster techniques also consist of a variety of methods, and some of them can result in overlapping partitions. KMC is discussed next as an example of nonhierarchical partitioning methods.

KMC, one of the iterative partitioning methods, is often used in both statistics and data mining for optimally partitioning data into a predefined number of final clusters. Operationally, it can be thought as "reversed" ANOVA (StatSoft Electronic Textbook, n.d.). The program starts with *k* predefined random clusters, and then moves objects among those clusters with the goal to minimize variability within clusters and to maximize variability between clusters. This is analogous to "ANOVA in reverse" in the

sense that, with the hypothesis that the means of the groups are different from each other, the significance test in ANOVA computes the *F* value to evaluate the between-group variability against the within-group variability. In KMC, the program tries to move objects (e.g., cases or variables) in and out of groups (clusters) to get the most significant ANOVA result. The iterative process can be outlined as follows (Aldenderfer & Blashfield, 1984):

1. An initial partition is chosen to set the data into a specified number of clusters. The centroid of each cluster is computed.
2. Every data point is assigned to the cluster that has the nearest centroid based on a distance measure.
3. A new centroid is calculated for every cluster. Clusters are updated when all the data have been worked through.
4. Step 2 and 3 are repeated until no data points change clusters.

KMC can be a better choice over hierarchical cluster analysis methods for the major reason that iterative partitioning methods can compensate for a poor initial partition by working through the data multiple times. Data points can be reassigned to different clusters through the iteration process in order to minimize the variance within clusters.

Ideally, the iterative partitions should be able to discover the optimal solution by forming all possible partitions of a data set. Unfortunately, this is computationally difficult except for trivial problems. As a less perfect alternative, a wide range of heuristic procedures have been developed to help sampling a small subset of all possible

partitions of a data set with the hope that at least an approximate optimal partition can be found from the sample. Because a sample can only cover a small portion of all possible partitions, a potential problem with KMC is that a suboptimal or locally optimal partition may be chosen as a solution by the procedure.

Caution is given that KMC measures the Euclidean distances between data points, and so the variables usually need to be standardized if they are on different scales. Also, in most statistical software, KMC is only available for partitioning cases, records, or subjects into clusters; in order to use KMC for variable classification, the input data matrix has to be transposed.

Determined by the similarity measures used in the procedures, EFA and KMC are not appropriate for grouping nominal variables. In order to consider nominal variables in the dimensionality analysis, another technique, MDS, can be used to study the relationships among all variables.

MDS

Using measures of similarities or dissimilarities between objects (e.g., cases or variables), MDS works to detect meaningful underlying dimensions that allow a reasonable explanation of the observed data structure. A variation of projection pursuit, the MDS attempts to arrange "objects" in a space of a specified number of dimensions, defined by axes, so as to reproduce their observed distances. Within a defined space, objects are moved around and some criterion is specified to check how well the distances between objects can be reproduced by a new configuration. Once an acceptable

dimensional definition is reached, it can "explain" the distances between objects in terms of the defined dimensions.

Similar to EFA, the actual orientations of the axes in unweighted MDS analysis are arbitrary (it is a different matter in weighted MDS), and can be rotated in any directions for an easy interpretation of the dimensions. The positions of objects are recorded as their coordinates on the axes. In order to simplify the dimensional structure, it is not required that the coordinates of objects recovered by the MDS produce distances the same as the proximity of the input data as long as the order of distances is preserved. Because it supports graphical representations of the proximities between objects spatially on a map such that their relative positions in the space reflect the degree of their perceived proximities (similarity), MDS often yields more readily interpretable solutions with fewer dimensions than the outcome of an EFA.

In MDS, the more dimensions used to reproduce the distance matrix, the better the reproduced matrix fits the observed matrix. Actually, a perfect reproduction of the observed distance matrix is possible if as many dimensions as variables are used. However, given the intention to reduce the observed complexity of data while keeping the approximate relative distances between objects, a proper number of underlying dimensions has to be determined with some objective measure. Whatever the measure is, it should be able to indicate the adequacy of the spatial representation. *Stress* is one such measure. A smaller Stress value usually means a better representation of the proximity data. Unfortunately, the Stress value not only decreases when the number of dimensions increases, it is also related to the number of objects being scaled, the type of similarity

measure used, and the reliability of the data (Jones & Sabers, 1992). Therefore, Stress is a convenient but not an ideal criterion for choosing the appropriate number of dimensions in MDS. In order to make the final decision, other information, including parsimony of structure, adequate representation of data, and *a priori* knowledge, should be considered along with the Stress value.

Once the number of dimensions is determined, the location of objects in the reproduced space can be checked to understand the structure of the data. The arbitrary axes may not be meaningful; thus the interpretation often relies on the grouping of objects and the way in which they are ordered in the space. Due to the subjective manner of structure interpretation, plenty of room is left for disagreements when interpreting the MDS result.

Graphical display of the data is an important feature of MDS. When the number of dimensions is small, a graph of data points can be much clearer than a mathematical model for revealing the data structure. However, a drawback associated with the emphasis on graphical descriptions of data is that the maximum number of dimensions is usually limited to six or less. When the expected number of dimensions is more than six, graphical outputs of the MDS analysis are difficult and offer little help for understanding data structure. Instead, only the numerical output (e.g., stimulus coordinates) may be useful to study the relative positions of variables.

MDS may provide further information of data structure not revealed by the EFA and KMC analyses because their different approaches in data modeling. EFA and MDS are fundamentally different methods. The former requires that the variable relationships

be linear and uses the Pearson correlation as the similarity measure; the latter imposes no such restrictions and is open to a variety of measures of variable associations. Meanwhile, EFA assumes an additive model in which the variance of a variable is explained by a weighted sum of the factors. MDS, like KMC, does not require an additive model and so has less restrictive assumptions. Also, MDS and KMC share some measures of variable similarity, have less restrictive model assumptions than EFA, and search iteratively for the underlying data structures. However, MDS is different from KMC in that it is not required to classify all objects into mutually exclusive groups.

In addition to the graphical display of relative positions of objects, MDS has another advantage over EFA and KMC: when used on a nonmetric level, MDS assumes that the level of measurement is at nominal or at best ordinal scale, so categorical variables can be included in the analysis.

Prediction Procedures

Variable reduction or feature selection is one of the preparation steps for building a valid and efficient model. In this section, techniques used for building prediction models are discussed. Statistically, multiple regression has been used as a dynamic approach for prediction. In data mining, many prediction techniques are available. Limited by space, only the BBN model, the one used in the current study for prediction, is covered in this section.

Multiple Regression Analysis

The primary concern of regression analysis is the estimation and/or prediction of the expected values of a dependent variable conditioned on the known values of some

independent or predictor variables (Dillon & Goldstein, 1986). The goal of regression is to find a line or a plane in a multidimensional space that provides the best fit to the sample data points. A regression model is in the form:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} + e_i, \quad (2.18)$$

where b_0 is a constant intercept value, b_1, b_2,..., b_p are the model parameters called regression coefficients, and e_i is the residual term (prediction error) associated with the ith observation. The model gives the expected value of Y conditional upon the input values of X_1, X_2,..., X_p, plus the error component.

A popular computational approach in regression analysis is to use the ordinary least squares (OLS) procedure to estimate the values of the unknown parameters so that a linear model is created having the minimum average prediction error (measured as the summed squared differences between the predicted and observed values). In order to make unbiased estimates of model parameters and valid statistical inferences, the implementation of the OLS approach requires a number of assumptions about the residual terms, including normality, homogeneity, and independence, in order to make valid estimates of parameters and statistical inferences about the population.

As in all areas of mathematical modeling, a regression model as parsimonious as possible is almost always most desirable with a good measure of fit. The parsimony in regression is achieved by using only as few parameters b_i as possible in the model. To do so, two important criteria used in building regression models are the measure of model goodness-of-fit and the tests of parameter significance.

A number of measures of model fit are available indexing how well the estimated regression function explains the data. The most popular ones are R^2 and adjusted R^2; both are restricted to the interval [0,1] with greater values indicating a better fit. R^2 is called the coefficient of multiple determination, which represents the proportion of total variance in the predicted variable that is accounted for by the regression model. Although it can never go beyond 1, R^2 always increases with the number of predictor variables included in the model. Unfortunately, a model with more predictor variables may not be better in terms of stability and generalization. High values of R^2 can also be possible in situations where the assumptions of the regression model are violated. So R^2 is not a very reliable measure of model adequacy.

Adjusted R^2, a modification of R^2 by taking into account the number of predictor variables in the model and the sample size, is recommended when comparing different models. Because it takes into consideration the effect of losing degrees of freedom when more variables are added to the model, adjusted R^2 may decrease when the increase in the R^2 is offset by the loss of degrees of freedom. Relatively, adjusted R^2 is a better measure of model fit than R^2 and most statistical packages provide both in the regression output.

Significance tests of model parameters are used to decide the inclusion or exclusion of one or more candidate independent variables in a regression model. For instance, a simple F test can be used to determine statistically whether one or more predictor variables can be removed from the model due to their nonsignificant contribution in explaining the variance of the dependent variable through a comparison of models with and without these particular variables. The same statistic can also be used to

guide the variable selection process and to test important hypotheses concerning relationships among model parameters.

Mechanically, each different combination of candidate predictor variables makes a regression model. The objective of a variable selection process is to look for the most parsimonious regression model with a satisfactory fit. Most statistical software provides sequential variable selection methods including forward, backward, and stepwise selections.

The backward selection starts with a model in which all predictor variables are included. The procedure works to exclude variables one by one from the model if the significance test on the drop in R^2 associated with a particular variable (i.e., partial R^2 calculated as the squared partial correlation with the dependent variable) produces a p-value greater than the predetermined α value. Rather than building down, the forward selection starts with the single predictor variable that has the highest correlation with the predicted variable and adds others according to their partial R^2 until none of the excluded variables has a p-value smaller than the predetermined α in the significance test of unique variable contribution. For both the backward and forward selection methods, the decision of removal or inclusion of a variable is irreversible. It is different for the stepwise selection method. Stepwise selection is similar to the forward selection with regards to the way variables are entered, but it reevaluates the partial R^2 of each included variable as if it were the last entered when more variables are added into the model. As a result, a previously entered variable may be removed because its correlation with other variables in the model reduces its p-value to lower than the α on the test of its partial R^2 with the

dependent variable. The iterative process stops when a model is reached where no excluded variables can be added and no included variables can be removed based on the specified criteria (e.g., predetermined α).

With straightforward criteria for variable evaluation, the forward, backward, and stepwise selection methods always provide a single "optimal" (according to their own criteria) regression model. However, their final pick may not be the best on different standards of model adequacy given that they may not yield a model with the maximum R^2, and also that the R^2 is not a perfect measure of model fit. Other advanced techniques are available, including C_p value, Max R, and Min R. Please refer to Neter, Kutner, Nachtsheim, and Wasserman (1996) for detailed discussions.

All these variable selection methods and model evaluation criteria share the goal of searching for an adequate model although they emphasize different aspects of the model fit. It is hardly possible to have a model satisfying all the criteria, but different measures of model fit should always be considered to make the optimal decision. For more discussion on multiple regression including the diagnostic measures, please refer to Cohen and Cohen (1975), Neter et al. (1996), and Dillon and Goldstein (1986).

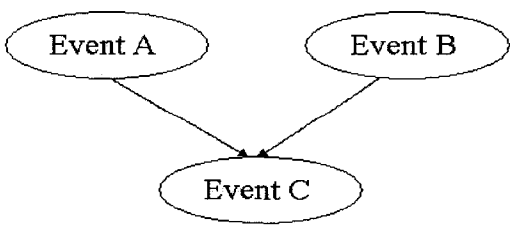


Figure 2.7. A BBN model of dependency.

BBN

A BBN is a network-structured model based on the Bayesian probability. Also known as Bayesian network or causal probabilistic network, BBN takes a graphical model to represent relationships between variables, even if the relationships involve uncertainty, unpredictability, or imprecision.

BBN as a Prediction Model. Based on the Bayesian probability, the following is simple demonstration of how BBN works as a prediction technique (Cumming, 2003). First, to set up a BBN, various variables or events have to be defined, along with the dependencies among them and the conditional probabilities involved in those dependencies. A BBN can use the given information to calculate the probabilities of various possible paths being the actual path leading to an event or a particular value of a variable. For instance, event C can be affected by events A and B (as shown in Figure 2.7) and the following probabilities are known:

Table 2.3

Conditional Probabilities in the Bayesian Network Example

	Event A: true P(A) = 0.1		Event A: false P(~A) = 0.9	
	Event B			
Event C	True P(B) = 0.4	False P(~B) = 0.6	True P(B) = 0.4	False P(~B) = 0.6
True	P(C | AB) = .8	P(C | A~B) = .6	P(C | ~AB) = .5	P(C | ~A~B) = .5
False	P(~C |AB) =.2	P(~C | A~B)= .4	P(~C | ~AB) = .5	P(~C |~A~B) = .5

When dependencies converge, some conditional probabilities need be filled in from subjective information or though calculation from data, the rest can be available with the rule that the probabilities for each state should sum to 1, as shown in Table 2.3.

With the above given or derived probabilities, the "initialized" probability of C can be calculated by summing the various combinations in which C is true and breaking those probabilities down into known probabilities:

$$
\begin{aligned}
P(C) &= P(CAB) + P(C\sim AB) + P(CA\sim B) + P(C\sim A\sim B) \\
&= P(C \mid AB) \times P(AB) + P(C \mid \sim AB) \times P(\sim AB) \\
&\quad + P(C \mid A\sim B) \times P(A\sim B) + P(C \mid \sim A\sim B) \times P(\sim A\sim B) \\
&= P(C \mid AB) \times P(A) \times P(B) + P(C \mid \sim AB) \times P(\sim A) \times P(B) \\
&\quad + P(C \mid A\sim B) \times P(A) \times P(\sim B) + P(C \mid \sim A\sim B) \times P(\sim A) \times P(\sim B) \\
&= 0.8 \times 0.1 \times 0.4 + 0.5 \times 0.9 \times 0.4 + 0.6 \times 0.1 \times 0.6 + 0.5 \times 0.9 \times 0.6 \\
&= 0.518
\end{aligned}
$$

So as a result of the conditional probabilities, C has a 0.518 chance of being true in the absence of any other evidence except for A and B.

If it is known that C is true, the "revised" probabilities of A or B being true (and therefore the chances that they caused C to be true) can be calculated with Bayes's Theorem and the initialized probability:

$$
\begin{aligned}
P(B \mid C) &= [\, P(C \mid B) \times P(B) \,] / P(C) \\
&= [\, P(C \mid AB) \times P(A) + P(C \mid \sim AB) \times P(\sim A) \,] \times P(B) / P(C) \\
&= (0.8 \times 0.1 + 0.5 \times 0.9) \times 0.4 / 0.518 \\
&= 0.409
\end{aligned}
$$

$$P(A \mid C) = [\, P(C \mid A) \times P(A) \,] / P(C)$$

$$= [\, P(C \mid AB) \times P(B) + P(C \mid A\sim B) \times P(\sim B) \,] \times P(A) \,/ P(C)$$

$$= (0.8 \times 0.4 + 0.6 \times 0.6) \times 0.1 \,/ 0.518$$

$$= 0.131$$

So the result shows that given C is true, B is more likely to be its cause than A.

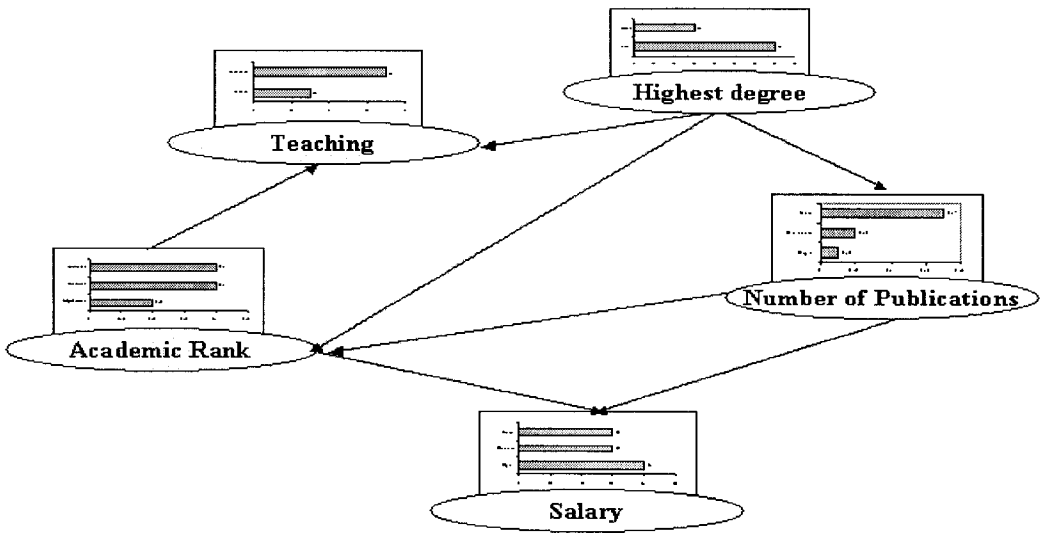


Figure 2.8. An example of the BBN model. This graph illustrates the three major classes of elements of a Bayesian network; all variables, edges, and CP tables are for demonstration only and do not reflect the data and results of the current study in any way.

BBN in Data Mining. A BBN is expressed as a special type of diagram together with an associated set of probability tables, as in the example shown in Figure 2.8. The three major classes of elements are a set of variables presented as nodes, a set of directed edges (arcs) between variables showing the causal/relevance relationships between variables, and a conditional probability table $P(A| B_1, B_2,\ldots, B_n)$ attached to each variable

A with parents B_1, B_2, …, B_n. The conditional probability describes the strength of the belief given that the prior probability is true (Winkler, 1972).

Assisted by computer programs, a BBN may be developed automatically from data files, created by experts, or by a combination of the two. Building a BBN is a collaborative work of probability, machine learning, and decision analysis. Once the variables are entered and topology is defined to build the network, an extensive iteration is underway to construct a full joint probability distribution over the entire product state space (i.e., the combinations of distinct values of all variables) of the model variables. The computation task is enormous because the elicitation at a later stage in the sequence results in back-tracking and changing the information that has been elicited at an earlier point (Yu & Johnson, 2002). With the iterative feedback and calculations, BBN uses probabilistic inferences to update and revise belief values.

BBN as a type of probability model has a unique feature that cannot be accomplished by any inferential models in classical non-Bayesian statistics: it permits the introduction of prior knowledge into the probability calculations and the model building (Hecherman, 1997). Prior knowledge--such as the beliefs about the dependencies among some variables articulated by domain experts--can be introduced into the BBN as subjective probabilities, and joined by evidence from objective data to propagate consistently through the network to impact the probabilities of uncertain sequential outcomes. The assumption of prior beliefs is a key element of Bayesian inference, which can be a weakness as well as strength: a BBN can only be as useful as the prior

knowledge is reliable. Either an excessively optimistic or pessimistic expectation of the quality of these prior beliefs can distort the entire network and invalidate the results.

BBN is a powerful tool also because it not only enables reasoning under uncertainty, but combines a sound mathematical basis with the advantages of an intuitive visual representation. The graph of the probability network is a format people can easily understand, and it allows a clear visualization of the relationships involved. In addition, knowledge captured in the modular form of the network can be transported easily from one situation to another.

The rigorous mathematical meaning of BBN is an advantage for its realization in computer programs, but software tools that can interpret BBN and perform the complex computation were not available until very recently due to the practical difficulty of performing the propagation even with the availability of high-speed computers. Even a small BBN demands a quite onerous computation that was intractable not long ago due to the fact that, to calculate the probability of any branch of the network, all branches must be calculated in exploring a previously unknown network (Niedermayer, 1998). While the resulting ability to describe the network can be performed in linear time, this process of network discovery was a task which might either be too costly to perform, or computationally impossible given the number and combinations of variables.

Currently, the BBN is becoming popular in data mining as a type of predictive model. The learning of a BBN is an intensive model selection process. However, due to the immense amount of computational resources required, an exhaustive search for the optimal model remains impossible in practice when the number of variables and their

product state space are relatively large, especially for a BBN to be learned from a large objective data set without any prior knowledge.

As a compromise, data mining researchers have developed some utility functions to help random feature subset selection in the network discovery process and to guide the search for the optimal subset with an evaluation function tracking the classification error rate of every attempted model. That is, a stochastic feature subset selection is wrapped into the BBN algorithm. The wrapper feature selection function conducts a search for the optimal subset of features using the BBN itself as a part of the evaluation function, the same algorithm that will be used to induce the final prediction model. The stochastic wrapper feature selection adds computational overhead to the BBN model building, but it is still cost-effective because an exhaustive examination of the entire model space can be avoided.

Finally, a remark is necessary concerning the probabilistic distributions induced for data modeling in the BBN discovery. Although explicit discussions of variable distributions are rare in BBN literature, selecting the proper distribution functions to describe the data has a notable effect on the quality of the resulting network (Niedermayer, 1998).

Evaluation Criteria of Prediction Models

Model evaluation is an integral part of data analysis in both statistics and data mining. The goal is to assess whether a model fits the data in an effective and efficient manner for a given problem (Vehtari & Lampinen, 2002). Model comparison is important when there is more than one candidate model summarizing the same set of data. The

multiple candidate models can be the outputs of a same technique, or they may result from different model building procedures. Either way, when different models are available for the same data set, the one with the best fit should be chosen through some comparison procedures based on appropriate evaluation criteria. Because traditional statistics and data mining share some practical model building techniques but have different emphases on model structures and their future applications, researchers in the two disciplines take related but different approaches in model evaluation and comparison. Following is a brief discussion of the criteria of model evaluation and the procedures for model comparison in statistics and data mining.

Model Evaluation and Comparison in Statistics

In traditional statistics, goodness-of-fit statistics have been commonly used to ascertain whether a given model fits the data (Sobel, 1995). The general procedure starts with defining a test statistic, which is a function measuring the discrepancy between the hypothesis (the model) and the data; and then, assuming the hypothesis is true, the probability is calculated of obtaining data having a still larger value on this test statistic than the value observed. This probability is the significance of the test or the confidence level. Small probabilities indicate a poor fit. However, especially high probabilities (close to one) often correspond to a fit that may be too good to happen very often, and indicate a mistake in the way the test was applied, such as treating data as independent when they are correlated or having a small data set relative to the number of estimated parameters.

Some popular goodness-of-fit tests are the χ^2 test for goodness-of-fit, Kolmogorov-Smirnov test, Anderson-Darling test, Shapiro-Wilk test, Von-Mises test,

and more. For instance, when variables are discrete, the χ^2 statistic can be used to test whether the population distribution from which the sample data were drawn is the same as the hypothesized distribution. Kolmogorov-Smirnov and Anderson-Darling are used to test whether the data fit a specified distribution. Actually, the Anderson-Darling test can be applied to any distribution, but tables of critical values for some distributions are sometime difficult to find except for the normal and lognormal distributions.

With the notion that data can be partitioned into fit and error components, explicit residual statistics become another choice in checking model fit (diagnostic analysis) and in making model selection. The residuals are supposed to be small if the model is a good fit to the data. For example, in multiple regression analyses, the residual analysis is very important for model evaluation and modification. If the model fits the data well, the residual terms should be relatively small and distribute randomly around zero. Substantial residual terms of a model and/or any systematic pattern in residual distribution suggest lack-of-fit of the model (Neter et al., 1996).

Model comparison is a topic related to but more complex than model evaluation. Adequacy of capturing data characteristics is the emphasis of model evaluation, but much more than model adequacy needs to be considered in model comparison. For instance, residual statistics in multiple regression analysis are not enough for choosing a final model from a class of alternatives (Sobel, 1995). In addition to goodness-of-fit tests, some other measures, including model R^2, complexity, and parameter stability should be taken into consideration for model comparison (Freund & Wilson, 1995). For example, to compare two competing models, the difference of their R^2s is tested for significance. In

addition, complexity is defined as the effective number of parameters in a model; it is considered as important as the fit because too many parameters hinder the model interpretation and sometimes cause problems including instability or model overfit. Thus, a number of model selection procedures try to balance the fit and the complexity (simplicity or interpretability) when making the final selection.

In some cases of model comparison, the candidate models are nested because they are products of the same model-building procedure. The models have comparable structures and parameters, such as the number of variables included, the residuals, goodness-of-fit measures, and so on. Model comparison in such cases is much easier than the comparison of models built by different procedures. In the latter case, lack of comparable parameters can make the indirect comparison quite difficult.

Model Evaluation and Comparison in Data Mining

As contrasted with traditional statistics, model evaluation and comparison are closely connected with each other in data mining because the selection of the final model is a process of evaluating and comparing the candidate models available in the defined model space. In general, even with the shift of research interest from data modeling to field application, classification accuracy as well as model adequacy remain at the top of the criteria list for model evaluation and comparison. The major concern is to show that a fitted model explains the data and makes predictions better than chance (Datta, 1999).

In data mining, model evaluation and comparison mainly rely on statistical procedures. For instance, to measure model accuracy in machine learning classification, Mitchell (1997) suggested using the *t* test and confidence interval as evaluation criteria.

With the model--a function *f* called the "hypothesis"--he defined "error" as the probability a model misclassifying an instance drawn at a random from the hypothesized distribution. When the estimated error rate $error_S$ is obtained from the sample data set, the $error_S$ is considered an unbiased estimate of the true error rate. Based on the binomial distribution, the mean and standard deviation of the sampling distribution of the estimated true error rate can be calculated. When there are two candidate models, the error rates of the two models can be compared through a simple *t* test.

In recent years, some computation-intensive methods, including Markov chain simulation and the Bayesian formula in general, have been tried as advanced techniques to test the fit of complex models because they allow flexible treatments of the probability models (Gelman & Meng, 1995; Ibrahim, Chen, & Sinha, 2001). The Bayesian approach to model evaluation and comparison covers a broad collection of topics including Bayes factors, model diagnostics, and goodness of fit measures.

One method of Bayesian model assessment is based on the calculation of the posterior model probabilities. For example, studies conducted by Rubin (1981, 1984) and Gelman, Meng, and Stern (1995) explored using the posterior predictive to assess model fitness via realized discrepancies in situations where the simple χ^2 bound is too crude. The general idea is that, when an observed data set *y* is modeled by hypothesis *H* in a multidimensional space with θ unknown parameters, a replicated data set y^{rep} can be produced by the same model *H* with the same set of θ parameters, a test statistic *T* is used to map a test statistic from data space to the real numbers. The comparison of the observed data and replicated data leads to a classical *p* value on *T* as

$$p_c(y,\theta) = P_A[T(y^{rep}) \geq T(y) \mid H,\theta)], \tag{2.19}$$

where a p value very close to 0 indicates the lack of fit in the direction of the test statistic T. This procedure is an example of Bayesian probability used for goodness-of-fit test. When there are multiple candidate models, the one with a better fit is preferred.

The posterior predictive approach has gained attention because it has the virtue of constructing a well-defined reference distribution with a corresponding tail-area probability that is easy to simulate for any test statistics. However, the most valuable point of this and some other Bayesian model comparison techniques is that they try to improve the model comparison from depending only on point estimates (e.g., a t-test of the classification accuracies) to developing a distribution of the expected utility estimate and drawing a conclusion based on probabilistic test results (Vehtari & Lampinen, 2002). The distribution of an expected utility estimate is analogous to the sampling distribution of estimated parameters in statistics. A decision based on a probabilistic sampling distribution of the utility estimates is expected to be more reliable than based on a simple comparison of point values of utility estimates.

Research Questions

The review of literature introduced a data analysis technique called data mining and discussed its characteristics in comparison to traditional statistical methods. In spite of its implementation of statistical procedures, data mining is substantially different from traditional statistics in its approach, methods, result presentations, and application objectives. Three reasons made this study necessary: the increasing number of large-scale data sets in educational fields, the inadequacy of traditional statistics when dealing with

such large data sets, especially those with a large number of variables, and the popular application of data mining as a powerful tool for extracting knowledge from massive data sets in business and scientific research. Using several different models, including multiple regression, the data mining BBN model, and a combination of the two, to work on a post-secondary faculty data set for comparison purpose, this study is designed to demonstrate the usefulness of data mining in educational research. Researchers can make their own assessment of data mining as an analytical tool for studying large volumes of data, and search for answers to following questions:

1. What are the similarities, differences, strengths, and weaknesses of data mining and traditional statistics when doing prediction with a large data set?
2. Which approach is more adapted to uncover useful information and to understand data structure given a data set with a large number of variables: traditional statistics, data mining, or a combination of the two at different stages of analysis?
3. How are the special features of data mining that may not be readily available from statistical software packages beneficial to educational researchers in processing data of very large volume?

Direct comparison of the multiple regression and the data mining BBN models is difficult, but hopefully a demonstration of the different methods can help to explore the potential of using data mining in education research.

CHAPTER 3 METHOD

Data Source

To answer the research questions, a data set is needed as a laboratory setting to build the statistical and data mining prediction models and to compare their characteristics. The data set chosen is the National Survey of Postsecondary Faculty 1999 (NSOPF:99). The reasons for using this particular data set are, first, it is a large education-related data set with more than 18,000 records and more than 400 variables. The size is large according to Huber's taxonomy introduced in previous chapter. Data of this size can be analyzed through both statistical and data mining approaches. Second, the data set is a collection of information on a specific group of people: faculty members in the postsecondary institutions that met certain standards. The structure of the data is relatively loose as most large data sets, and variables contain redundant and irrelevant information. However, it provides enough information to answer a pseudo research question: what are the factors highly related to the faculty salary level? That is, the faculty salary determination is a predictive question that can be answered by both statistical methods and data mining models with this data set. Finally, many studies are available on faculty compensation in postsecondary institutions, which make possible a theoretical model as a comparison to other prediction models in this study.

Background

NSOPF:99 was a survey conducted by the NCES in 1999. The initial sample included 960 degree granting postsecondary institutions and 27,044 full- and part-time faculty employed at these institutions. The sampled institutions represented all public and

private not-for-profit Title IV-participating, degree-granting institutions in the 50 states and the District of Columbia. The faculty population included all those who were designated as faculty, whether or not their responsibilities included instruction, and other non-faculty personnel with instructional responsibilities. Under this definition, researchers, administrators, and other institutional staff who held faculty positions but did not teach were included in the target population.

Both the sample of institutions and the sample of faculty were stratified and systematic samples. The institution sample was stratified by the Carnegie classifications that were aggregated into fewer categories. The faculty sample was stratified by gender and race/ethnicity. At the end, approximately 18,000 faculty and instructional staff questionnaires were completed for a weighted response rate of 83 percent. The response rate for the institution survey was 93 percent.

For the purpose of this study, only faculty data were used. The faculty data file of NSOPF:99 included 18,043 records and 439 original and derived measures. It contained faculty information about their backgrounds, workloads, responsibilities, salaries, benefits, attitudes, future plans, and so on. For more information about this survey and its data, please refer to the methodology report available on the NSOPF:99 CD-ROM.

Data Preparation

Data preprocessing steps were taken to prepare the NSOPF:99 faculty data for the analysis. To focus on the compensation structure of regular fulltime faculty with teaching responsibility, respondents were eliminated from the analysis if they were part-time or had a researcher or administrative title including post-doc, visiting positions, and other

academic titles such as managers, supervisors, coaches, chaplains, counselors, mentors, advisors, librarians, curators, research associates/assistants, secretaries, miscellaneous clericals, adjunct faculty, teachers, and other unspecified. Faculty assigned by religious order were excluded as well. In addition, some respondents' records had to be removed from the data set to eliminate invalid measures of salary. With a reasonable assumption that there were 35 weeks in an academic year and the paid hours per week was 40 or less, respondents with a reported salary lower than $4.75 per hour on average were deleted. As a result, the total number of records kept for the analysis is 9,963. Two thirds of the records were randomly selected as training data and used to build the prediction models; the remaining one third were saved as testing data for the purpose of cross-validation.

After the respondent cases were cleaned, actions were taken to process the variables. First, 112 original variables were manually screened out of the entire set of more than 400 variables. The screening was made in a way so that only the most salient measures of professional characteristics were kept to quantify factors considered relevant in determining salary level according to the general guidelines of salary schema in postsecondary institutions and to the compensation literature in higher education.

Then, in order to avoid redundant or over-specific information, in several cases, some highly related variables were combined to form a single comprehensive measure. One example is the measures of fund and support availability for professional development. Six questions were asked about internal tuition remission funds, internal professional association membership fee, internal professional travel funds, internal training funds, internal release time from teaching, and internal sabbatical leave. Because

answers to the questions were on a same ordinal scale, they were added together as a single indicator of the internal support for professional improvement (Q61SREC). However, multiple measures were kept on teaching, publication, and some other constructs because they quantified different aspects of the general constructs; the redundant information among these variables provided a chance of testing the differentiation power of the different variable selection and reduction techniques. The total number of variables reduced to 91 after the preprocessing. Table A1 in Appendix A provides more information on the 91 variables.

The NSOPF:99 data were imputed by the NCES with regression-based, hot-deck, or logical methods. Thus, the originally unanswered questions were imputed or identified as legitimate skipped or missing data. For the purpose of the current study, additional systematic recoding of missing values was performed. First, when it was reasonable to take "legitimate skip" as "not applicable", the missing answers were recoded so that the participants who skipped the question were treated as a separated group in future analysis. For example, a few participants who had no instructional responsibilities during the 1998-99 academic year did not provide an answer when asked what type of courses (credit or non-credit) they were teaching. The original coding of "legitimate skip" –6 was recoded into 0 so that they could be included in future analyses as a uniquely identifiable group. Totally, the "legitimate skips" of 16 variables were recoded into a number that was valid in the subsequent analyses. Missing values of other variables were coded as SYSTEM MISSING.

Other data preparation procedures included reverse coding of some ordinal variables, changing calendar years to the numbers of years until 1999, and so on.

Data Analysis

After the data preprocessing, descriptive statistics of all the 91 remaining variables were checked for range, distribution, variable intercorrelations, and possible irregularity including missing data, outliers, and other noises with EDA graphical and descriptive methods. Descriptive information of the major variables is provided in Tables A2 and A3 in Appendix A. The preparation and examination steps were critical for both statistical and data mining analysis.

Data Modeling

Once data preparation was completed, four different models were built for faculty salary prediction. For each of the models, the process started with examination and simplification of the variable space: Variable reduction or extraction was necessary in order to remove redundant and irrelevant information. For traditional statistical methods that only work well with a small number of variables, this step helps to reduce the number of variables to a manageable level so that the multiple regression analysis is feasible, effective, and interpretable. Feature (variable) selection/extraction is necessary for the data mining prediction model because, as explained in the previous chapter, it helps to reduce the required computer resources, improve efficiency of model searching, and enhance model accuracy and validity.

The four models used different methods for variable reduction or extraction; the technical details are covered next in the description of individual models.

Model I: Theoretical Model

The first model was a multiple regression model constructed with 13 independent variables selected from the NSOPF:99; the selection of the 13 variables was based on established theories about faculty compensation. This theory-driven modeling is a well-accepted approach in published studies of faculty salary and compensation analyses. In this study, this regression model served as a base model for comparison purposes only, similar to a control group in a designed experiment.

Many studies have been conducted on faculty compensation in postsecondary institutions. Although it is still arguable whether some variables should be included and how some variables should be quantitatively measured, many of the researchers (e.g., Bella, Ritchey, & Parmer, 2001; McLaughlin & McLaughlin, 2003; Moore, 1993; Simpson & Sperber, 1988) agree on including variables such as academic rank, tenure status, seniority, and regional economic factors, in studies of diverse faculty samples. A thorough literature review suggests the following variables be chosen from the 90 candidate predictor variables from the NSOPF:99.

1) Total classroom credit hours taught. This was the total number of classroom credit hours taught by a respondent during 1998-99 academic year; noncredit teaching responsibility was not included. Quite a few measures were available on teaching activities, including undergrad/graduate classroom credit hours, total classes taught, and total credit classes taught, but total classroom credit hours taught is the most objective quantification of the teaching load.

2) Principal field of teaching/research. There were a total of 10 general categories: agriculture and home economics, business, education, engineering, fine arts, health sciences, humanities, natural sciences, social sciences, and all other programs. This variable was included because studies have shown that faculty salary in different academic disciplines differ substantially.
3) Department chair. Whether the respondent was a department chair. Because school administrators were excluded, department head was a variable measuring administrative responsibility.
4) Tenure status. Whether a respondent was tenured, tenure eligible, or not on the tenure track in an institution having a tenure system.
5) Academic rank. Seven general categories were defined: full professor, associate professor, assistant professor, instructor, lecturer, other ranks, and no rank. Both the tenure status and the academic rank measure privileged status in academic settings; they are correlated, but both were included because it was not clear which was the most appropriate.
6) Total number of publications. This variable was a respondent's total number of publications during one's career, including articles published in juried and nonjuried media, published reviews of books or chapters in edited volumes, and textbooks and reports (a sum of the first five publication variables Q29A1-Q29A5). The unweighted sum of the total number of publications was used because ranking publications in different media has always been difficult and lacks a unanimous standard.

7) Years since first job in higher education. This was a measure of number of years a respondent had worked in higher education institution untill 1999. This variable is a measure of experience, a factor that can influence the salary positively or negatively (e.g., salary compression).
8) Highest degree type. The highest degree held by a respondent had eight categories: first-professional degree, doctoral degree, master of fine arts or social work, other master's degree, bachelor's, associate degree or equivalent, certificate or diploma for completion of undergraduate program, and others.
9) Total funds from all sources. This variable indicated the total funds (in U.S. dollars) received by a respondent from all sources during the 1998-99 academic year.
10) Respondent's gender.
11) Respondent's ethnicity. The gender and ethnicity information was included for examining salary equity.
12) Institution type. Three general categories of the institutions were doctoral, four-year non-doctoral, and not-for-profit 2-year. Another measure of institution type, the Carnegie classification, is more "official", but its 15 categories take away 14 degrees of freedom. Thus, this classification on ordinal scale was used instead.
13) Institution BEA (Bureau of Economic Analysis) region code. This was a variable created by the NCES to classify the institutions according to their geographical region using the BEA region codes. The ten categories were

New England, Mid East, Great Lakes, Plains, Southeast, Southwest, Rocky Mountain, Far West, U. S. service school, and not in IPEDS. Regional factors should be included in the regression because the NSOPF:99 collected data from a national sample, and literature supports the argument that cost of living is a factor in salary determination.

With the compensation measure (basic salary of the 1998-99 academic year) as the predicted variable, a total of 14 original variables were selected to build the first multiple regression model. In addition, based on Johnson's argument (1999) that female faculty were paid less than their male colleagues partly due to female faculty being unfairly delayed in moving up the ladder of the academic ranks, the interaction term between gender and academic rank (RANKGEN) and between gender and tenure status (TENUGEN) were also included as a part of the initial regression model. A third interaction term, the one between the total number of publications and teaching load by classroom credit hours, was included to evaluate the argument that research was more rewarded than teaching in postsecondary institutions, and teaching faculty were paid less because the heavy teaching responsibilities left them with less time for research activities.

After the three multilevel nominal measures (principal field of teaching/research, ethnicity, and BEA region code) were recoded into binary variables, the initial multiple regression started with 35 input variables for the prediction of faculty salary. Following the human capital theory in Economics and previous studies, salary was log-transformed to alleviate its non-linear relationships with the independent variables. The data in this study also confirmed this transformation: the R^2 was greater than .42 for the model with

log-transformed salary, and was only about .36 for the model with untransformed salary as dependent measure given the same group of initial independent variables.

To select the optimal regression model, several variable selection techniques were used including forced entry (*enter*) and stepwise selection. Forced entry of variables provides information of the importance of individual variables, but the order in which the variables entered into the regression model make differences in the outcome. In contrast, stepwise selection orders variables in accordance with their partial R^2s. If any of the variables were significant in one variable selection method, but nonsignificant in the other, individual tests on those variables were conducted to make the final decision whether to include them in the final regression model. In addition, the proposed model was cross-checked with all-possible-subsets regression techniques including Max R and C_p evaluations to make sure the model was a good fit in terms of the model R^2, adjusted R^2, and the C_p value.

Model II: Statistical Model

The second model was also a multiple regression model; but different from Model I, which used theory-driven method for variable selection, Model II relied on statistical variable reduction techniques to extract important measures from the 90 NSOPF:99 variables (salary measure excluded). To differentiate it from others, this model was named as "statistical model". Other than the different variable reduction methods, it did not differ from the theoretical model in terms of the prediction model building.

The variable reduction for Model II was completed in two steps. In the first step, the dimensional structure of the variable space was examined with EFA, KMC, and

MDS; based on the results of the three techniques, variables were classified into a number of major dimensions. In the second step, different variable extraction methods were tried; the one resulting in a higher model R^2 was used that extracted one measure with the highest R^2 with the predicted variable from each identified variable group.

During the first step, multiple techniques were used to scrutinize the underlying dimensional structure of the variables and to reduce the potential bias associated with each of the individual approaches. EFA measures variable relationships by linear correlation; only 82 variables on dichotomous, ordinal, interval, or ratio scale were included. To obtain as much information as possible about the underlying structure of the 82 variables, different factor extraction methods were tried and followed by both orthogonal and oblique rotations of the extracted factors. The variable grouping was decided based on the matrices of factor loadings: Variables that had a minimum loading of .35 on the same factor were considered belonging to the same group.

In the KMC analysis, the Euclidian distance was used to measure similarities of the same group of 82 variables. Usually, the number of output clusters needs to be specified to start a KMC analysis. When the exact number of variable clusters is unknown, the results of other procedures (e.g., EFA) can provide helpful information to determine a range of the possible number of clusters. Then the KMC can be run several times, each time with a different number of clusters specified within the range. The multiple runs of the KMC can also help to reduce the chance of getting a locally optimal solution. However, the KMC procedure in SAS does not require a cluster number; it produces the number of clusters when the convergence criterion is met. Because

variables were separated into mutually exclusive clusters, the interpretation of cluster identity was based on variables that had short distance from the cluster seed (the centroid). The results of the KMC analysis were compared with the findings of the EFA for similarities and differences.

The MDS procedure was performed on all 90 variables; to accommodate the inclusion of categorical measures, χ^2 was used to measure variable association. Stress values, results of the EFA and KMC procedures, and some *a priori* knowledge were considered to determine the number of dimensions in the MDS. Because the number of dimensions was potentially relatively large, making it unlikely to take advantage of the graphical display, the numeric coordinates of variables in the suggested dimension were evaluated for variable clusters. The MDS provided very little useful information for understanding the variable space due to the lack of clear dimensionality.

A final dimensional structure of the variable space was determined based on the consensus of the EFA, KMC, and MDS outputs; each of the variable dimensions was labeled with a meaningful interpretation. Because of the different clustering methods used, variables in the same group might not share linear relationships. Thus, extracting artificial variables through linear techniques (e.g., PCA) was not an optimal choice. Therefore, two other approaches were tried to extract one variable from each cluster: One approach was to regress the log-transformed salary on the variables in the same cluster, and select the one with the greatest partial R^2 change. The second approach was to standardize the variables in the same group and then use their unweighted sum as one

artificial predictor variable. The first approach was finally taken because it resulted in a higher model R^2.

Variables that did not show any strong relationships with any of the major groups, along with categorical variables that could not be classified, were carried untouched into the initial multiple regression model as individual predictors and tested for their significance. Possible interactions among the predictor variables were examined and included in the model if significant. As in the theoretical model, the dependent variable, the basic salary of the academic year, was log-transformed and categorical variables were recoded into binary variables. Variable selection techniques used were forced entry and stepwise selection; if any of the variables was significant in one variable selection method, but nonsignificant in the other, an individual test on the variable was conducted for making the final decision whether to include the variable in the final regression model. As in Model I, the proposed model was cross-checked with all-possible-subsets regression techniques including Max R and C_p evaluations to make sure the model was a good fit in terms of the model R^2, adjusted R^2, and the C_p value.

In summary, the dimensional structure underlying the large number of variables provided a schema of grouping related measures and therefore made it possible to simplify the modeling process by means of variable extraction.

Model III: Data Mining BBN Model

The third model for prediction was the BBN-based data mining classification model. To build the BBN model, the 90 original variables were input into a piece of software called the *Belief Network Powersoft* ; variables on interval and ratio scales were

binned into category-like intervals because the network learning algorithms require discrete values for a clear definition of a finite product state space of the input variables. The salary measure remained its original values (no log-transformation) as the predicted class variable; log-transformation was not necessary because BBN is a robust nonmetric procedure not influenced by any monotonic variable transformation. The salary was binned into 24 intervals, with the first interval from the minimum to $29,600, the second from $29,601 - $32,615, and so on. Because the rule of binning is to keep the same number of cases in each bin, the width of bins may vary. Other binning methods are available, but with approximating to an even distribution in which chances are equal for respondents to fall in each bin, the current schema is chosen to ease the computation by avoiding complicated variable distributions while keeping the model prediction as accurate as possible. During the learning process, an embedded wrapper feature selection was performed internally on the predictor variables

The BBN model learning was an automated process after reading in the input data. According to Chen and Greiner (1999), two major tasks in the process are learning the graphical structure (variable relationships) and learning the parameters (CP tables). Learning the structure is the most computationally intensive task. The BBN software used in this study takes the network structure as a group of CP relationships connecting the variables, and proceeds with the learning by identifying the CPs that are stronger than a specified threshold value with some statistical functions (e.g., χ^2 statistic and mutual information test).

Table 3.1

Summary of the Four Prediction Models

Models	Variable selection	Variable type	Prediction model
Theoretical model	Theory-driven selection	Original variables from the NSOPF:99	Multiple regression
Statistical model	EDA extraction	Original variables from the NSOPF:99	Multiple regression
BBN model with wrapper selection	Wrapper feature selection	Original variables from the NSOPF:99	Data mining BBN model
Combination model	Variable selected by the BBN model	Original variables from the NSOPF:99	Multiple regression

The output of the BBN model was a network in which the nodes (variables) were connected by arcs (CP relationships between variables) and a table of CP entries (probability) for each arcs. Only the subset of variables that was evaluated having the best prediction accuracy stayed in the network. The prediction accuracy was measured by the percentage of correct classifications of all observations in the data set.

Model IV: Combination Model

The fourth model was a combination of data mining and statistical techniques: the variables selected by the data mining BBN model were put into a multiple regression procedure for an optimal prediction model. Although a wrapper feature selection is dependent on the inductive learning algorithms in data mining, it was still interesting to see whether the CP relationships among variables that were captured by the Bayesian

network could help the variable selection from a large number of candidate measures in prediction problems. Again, categorical variables were recoded and the basic salary as the dependent variable was log-transformed. Multiple variable selection techniques were used including forced entry and stepwise selection. If necessary, the candidate model was cross-checked with all-possible-subsets regression techniques including Max R and C_p evaluations to make sure the final model was a good fit in terms of the model R^2, adjusted R^2, and the C_p value.

Table 3.1 summarizes the four prediction models of this study.

Model Comparison

The goal of this study is to be suggestive rather than conclusive, so the details of the four models were displayed to facilitate discussion about their similarities, differences, and the factors that made any of them work better than others according to some specific criteria. The major concern was to provide solid examples of prediction models in different data analysis approaches given a large number of variables and a large number of cases, so that readers can have enough information to make their own evaluation of the strengths, weaknesses, and possible compatibilities of statistical and data mining techniques.

First, the theoretical model, the statistical model, and the combination model were comparable because they were all multiple regression models, but taking different sets of candidate predictor variables. The differences between the statistical and theoretical model indicated how well the statistical multivariate techniques worked for extracting fewer variables but keeping as much information as possible, and also revealed the

effectiveness of data-driven variable reduction methods in contrast with the theoretical variable selection. The differences of the combination model vs. the theoretical and statistical models showed how the Bayesian probability-based data mining variable selection worked differently from other variable selection or reduction methods when making prediction through multiple regression models. With the same model building technique, several model evaluation criteria were used including R^2, adjusted R^2, standard error of estimates, model generalizability, and so on. For example, R^2 and adjusted R^2 showed how well the model can explain the variance of the predicted variable. For models that had close R^2 values, the numbers of independent variables in the final models became another consideration. A model with fewer variables is easier to interpret and has less chance for multicollinearity. Along with standard error of estimates, graphical and mathematical examinations of the residual terms and the tests of lack of fit were also available. The performance of the three models on the testing data set was used as an important indicator of model stability, generalizability, and prediction accuracy.

Second, the BBN data mining model with wrapper feature selection was examined in contrast with the three regression models. Through the combination model, how the wrapper feature selection in data mining worked differently in comparison to the statistical variable selection/reduction methods was studied. The inputs, algorithms, outputs, final model presentations, model interpretability, and other aspects of the models were discussed. To compare the prediction accuracy of the BBN model with that of the regression models, the observed and predicted salary values in the regression models were binned the same way as in the BBN classification. Model stability was checked with

the testing data set: the overall model accuracy (e.g., classification accuracy, R^2, or standard error of estimates) should stay consistent for both the training and testing data set if the models were stable. Part of the qualitative comparisons was regarding the model simplicity, interpretability, and robustness against noisy data including outliers and missing data. A simpler model is always preferable if it does not a make substantial difference in the utility measure because of the cost-effectiveness and model interpretability.

Finally, the exploratory nature of data mining techniques was discussed in terms of their usefulness and uniqueness in understanding structures of large and huge data sets, what they do differently from most statistical exploratory techniques, and how they can be used in educational research.

Software

Two statistical software packages, SAS and SPSS, were used for statistical procedures including EFA, KMC, MDS, and linear multiple regression.

A piece of software was also needed to run the BBN data mining algorithm on the NOPSF:99 salary data. Because most of the commercial software with Bayesian learning algorithms is very expensive, due to financial limitation, a shareware available on the World Wide Web called BN PowerConstructor (Chen, 2001) was downloaded and used in this study. The BN PowerConstructor was the winner of the yearly competition of Knowledge Discovery and Data mining (KDD) – KDDCup 2001 Data Mining Competition Task One, being recognized for its best prediction accuracy among 114 submissions from all over the world. This system is an extension of the BBN learning

system to BBN based classifier learning. In terms of the feature selection task, the software is designed to perform a wrapper subset selection automatically in the process of building the BBN from the data.

CHAPTER 4 RESULTS AND DISCUSSION

The same set of data was analyzed with four different methods: a multiple regression model with theoretical variable selection, a multiple regression model with statistical variable extraction, a data mining BBN model with wrapper feature selection, and a combination model that used variables selected by the BBN in a multiple regression procedure.

To be systematic with the study design, this chapter first presents the results of individual models separately and then compares their major findings at the end. Due to the multiple analysis procedures used, the presentation of results in this chapter is limited to the major findings only. Unless specified otherwise, the significance level was $\alpha = 0.01$ throughout the study.

Model I: Theoretical Model

The first multiple regression model used 14 variables that were selected from the original data set based upon published studies and theories of faculty compensation as described in the previous chapter. Basic academic year income (SALARY) departed far from the normal distribution, but its log transformation (LOGSAL) was much better (see Figure A1 in Appendix A). LOGSAL was used as the dependent (predicted) variable in this regression model because of the support from literature, and in the current study, the scatter plots showed that some of the independent variables had a better linear relationship with LOGSAL than with SALARY.

Three multilevel nominal variables among the 13 predictor variables (i.e., ethnicity, BEA region codes, and principal academic disciplines) were recoded into

binary variables before input into the regression model. A total of 35 predictor variables (10 dichotomous, ordinal, and interval variables, 3 interaction terms as introduced in the previous chapter, and 21 binary variables recoded from the three categorical measures) were included in the initial regression model and account for 42.2% of the total variance (adjusted $R^2 = 0.4193$).

One obvious problem with the initial model was the eight binary variables recoded from the BEA region codes (X0_37). Using the U.S. service schools as the baseline, all eight categories had very close regression coefficients (between –0.315 to –0.444), indicating few differences would exit if the U.S. service schools were not included. The frequency information in Table A2 in Appendix A showed that only five out of the 6652 faculty were from the U.S. service schools, suggesting that they be considered outliers that did not fit the pattern of the rest of the data. In addition, the eight binary variables (BEA1- BEA8) had extremely high variance inflation factor (VIF; ranging from 59.83 to 259.57), a sign for variable multicollinearity. A VIF of 1.0 indicates that a variable is uncorrelated with other predictor variables. The higher the value of VIF, the more the variable overlaps with others. To reduce the model instability caused by the eight BEA binary variables, two solutions were possible. One was to change the BEA codes to U.S. service schools vs. others; the other was to treat the five respondents from the U.S. service schools as outliers and remove them from the analysis. The latter was chosen because it offered a chance to examine whether the regions could make a difference in the salary prediction when the U.S. service schools were excluded.

Thus, the BEA region codes were recoded into seven binary variables with *Far West* as the baseline, and the initial model was revised accordingly.

When the predictor variables were entered into the revised initial model altogether by forced entry method, tests conducted on individual variables showed that all three interaction terms were nonsignificant with $F(1, 6608)=2.38$ at $p = .123$, $F(1, 6608)=0.94$ at $p = .333$, and $F(1, 6608)=0.39$ at $p = .530$ for effects between academic rank and gender, between tenure status and gender, and between total classroom credit hours taught and the total number of publications, respectively. Ethnicity and total credits taught were another two nonsignificant variables with $F(3, 6608)=1.91$ at $p =.126$ and $F(1, 6608) =0.91$ at $p=.339$, respectively. The seven binary variables from the BEA region codes were significant when evaluated collectively as a group with $F(7, 6608) = 16.27$ at $p < 0.0001$.

A decision had to be made whether to include tenure status (Q10REC) because it was nonsignificant in the revised model with the forced entry ($F = 0.76$ at $p = .383$), but the model using stepwise selection captured it with a significant partial R^2 ($F = 29.76$ at $p< .0001$) when all the other variables remained significant or nonsignificant as in the revised model (Table B1 in Appendix B). The difference was mainly due to the strong correlation between the tenure status and the academic rank ($r = 0.498$) and the inclusion of the interaction between tenure status and gender ($r = 0.738$). With the consideration that this theoretical model was built as a comparison to other models with regarding to prediction accuracy, the final model was determined to include all the original variables except for the nonsignificant ethnicity, total classroom credit hours taught, and the three

interactions. A test on tenure status showed that it was significant at $F(1, 6615) = 29.25$ at $p < .0001$, indicating it was appropriate to include this variable in the final model.

To evaluate this model, two all-possible subset procedures, Max R and C_p, were also conducted. The greatest R^2 for all the 26-*df* models was .4219 ($R^2 = .4213$ for the selected final model), belonging to a model in which a couple of the nonsignificant BEA binary variables were replaced with two interaction terms. The best C_p for the 26-variable models was 25.2, presenting a model that replaced two BEA binary variables with two ethnicity variables. Given that binaries created from the same categorical variable should be evaluated as a group, the final model appeared to be a good choice in terms of the model R^2 and the C_p (= 30.5).

The summary information of the final regression Model I is shown in Tables 4.1 and 4.2 (model *df* = 26 because categorical variables were recoded into binaries).

Model II: Statistical Model

With a total of 90 potential predictor variables, two steps were taken to build the multiple regression Model II: variable reduction and model selection.

Variable Reduction

In order to reduce the number of input variables for the regression model, multivariate data reduction techniques were used to understand the variable structure. EFA and KMC were applied to the 82 dichotomous, ordinal, interval, and ratio variables, and nonmetric MDS was tried to understand the underlying dimensions of all the 90 predictor variables.

Table 4.1

Parameter Estimates of Model I

Variable name	Variable definition	Parameter estimate	Standard error	t value	$p > \|t\|$
Intercept	Intercept	10.1296	0.0359	282.10	<.0001
Q13	Chair of a department	0.0388	0.0118	3.30	0.001
Q59A	Total funds from all sources	2.16×10^{-07}	2.12×10^{-08}	10.17	<.0001
Q81	Gender	-0.0991	0.0089	-11.17	<.0001
Q24A1REC	Years since first job in higher education	0.0073	0.0004	16.42	<.0001
TOTPUB	Total number of publications	4.80×10^{-04}	4.08×10^{-05}	11.76	<.0001
Q16A1REC	Highest degree type	0.0824	0.0052	15.82	<.0001
Q10REC	Tenure status	0.0265	0.0049	5.41	<.0001
X01_8REC	Academic rank	0.0428	0.0036	12.02	<.0001
X08_0D	Institution type: doctoral, 4yrs, or 2 yrs	0.0886	0.0064	13.85	<.0001
BEA region code (Baseline: Far West)					
BEA1	New England	-0.0564	0.0209	-2.70	0.0069
BEA2	Mid East	0.0157	0.0157	1.00	0.3173
BEA3	Great Lakes	-0.0648	0.0153	-4.24	<.0001
BEA4	Plains	-0.1139	0.0177	-6.43	<.0001
BEA5	Southeast	-0.0889	0.0143	-6.23	<.0001
BEA6	Southwest	-0.0565	0.0178	-3.17	0.0015
BEA7	Rocky Mountain	-0.1151	0.0223	-5.17	<.0001

Variable name	Variable definition	Parameter estimate	Standard error	t value	$p > \|t\|$
Principal discipline of teaching/research (baseline: unspecified)					
DSCPL1	Agriculture & home economics	-0.1735	0.0314	-5.52	<.0001
DSCPL2	Business	-0.0485	0.0220	-2.21	0.0273
DSCPL3	Education	-0.2192	0.0210	-10.46	<.0001
DSCPL4	Engineering	-0.0903	0.0240	-3.76	0.0002
DSCPL5	Fine arts	-0.2739	0.0238	-11.53	<.0001
DSCPL6	Health sciences	0.0388	0.0177	2.19	0.0286
DSCPL7	Humanities	-0.2456	0.0180	-13.64	<.0001
DSCPL8	Natural sciences	-0.1705	0.0169	-10.10	<.0001
DSCPL9	Social sciences	-0.1880	0.0190	-9.88	<.0001
DSCPL10	All other programs	-0.1357	0.0186	-7.29	<.0001

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Table 4.2

The ANOVA Table of Model I

Source	*df*	Sum of squares	Mean square	*F*	$p > F$
Model	26	519.0913	19.96505	185.19	<.0001
Error	6615	713.1660	0.10781		
Corrected total	6641	1232.2573			

Note. Model R^2 = .4213 and adjusted R^2 = .4190; Root MSE (standard error of estimate) = 0.3283.

EFA

The factor analysis started with the correlation matrix of the 82 variables. Different factor extraction techniques were tried, but only EFA with principal component extraction was able to converge within a specified 250 iterations. Twenty-six principal factors with an eigenvalue greater than one were extracted, accounting for 64.24% of the total variance. If the scree plot of the eigenvalues was considered, the "elbow" around eigenvalue = 2 would qualify only six factors and about 31% of the total variance (Figure 4.1). Taking into consideration that the EFA was conducted to understand data structure, all factors with eigenvalues greater than one were kept for interpretation.

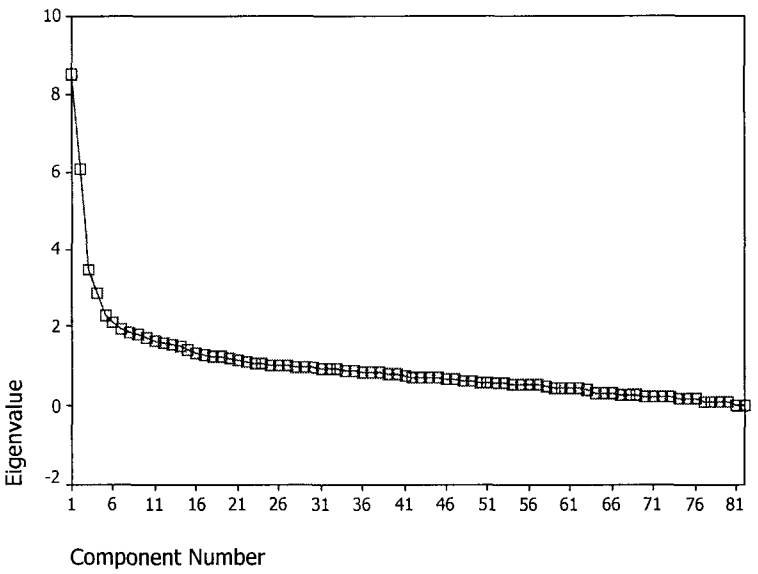


Figure 4.1. Scree plot of the factor eigenvalues in the EFA.

A simplification from 82 variables to 26 factors was a significant improvement; but the factors without any rotation were difficult to interpret due a complicated component loading structure. The component matrix showed that many variables had similar loading patterns, and some variables loaded almost equally on several factors. In addition, a general factor existed with high loadings from many variables. Therefore, two different rotation methods--VARIMAX for orthogonal and OBLIMIN for oblique rotation--were tried and successfully simplified the component loading matrix. The interpretability of the EFA results was greatly improved after the rotations. In both cases, the interpretation of factors was based upon variables that had a minimum loading of .35 (see Tables C1 and C2 in Appendix C for lists of variables and their rotated loadings on the factors). Given that both the orthogonal and the oblique rotations were carried out on the same component loading matrix, their rotated factors shared similar structures as expected.

As shown in Tables C1 and C2, most of the variables loaded on a same factor appear to share some common meaning. For example, Factor 1 in the VARIMAX rotation was very close to Factor 2 in OBLIMIN rotation; they had seven variables loaded highly (loading > .80) including years of teaching in higher education institutions, years on current job, years since first job in higher education, years since highest degree, and so on. It was clear that the factor was about experience. In short, the analysis results suggested that the interpretable factors shared by the two different rotation results were administrative responsibilities, education level, institution parameter (two factors), other employment, publications (multiple factors including books, creative works in juried

media and non-juried media, presentations, and reviews), research activity, experience, and teaching activities (multiple factors related to individual instruction, teaching load, graduate teaching, undergraduate teaching, etc.), and job satisfaction. One major difference between the results of orthogonal and oblique rotations was that more variables appeared in multiple factors in the oblique rotation because the factors were not required to be independent of each other.

With a relatively large number of factors extracted, variables related to some general constructs were refined to several factors with different emphases. Two typical cases were the factors related to teaching and publications. One of the reasons for this refinement could be the special data distributions that existed among some variables. For instance, distribution could be the main reason that two variables, undergraduate committees served and chaired, were isolated from other clusters, because 5,213 out of 6,652 respondents had 0 for number of undergraduate committees served, and they were part of the 5,805 who had 0 for number of undergraduate committees chaired. The extremely skewed distributions made variable correlations highly venerable to chance agreement and prevented the two variables being grouped with other related variables.

Meanwhile, the results also indicated that the EFA could be improved if some of the variables were excluded from the analysis given their independence from other variables. First, four variables in VARIMAX rotation and four in OBLIMIN rotation did not have a loading greater than .35 on any of the factors, including research appointment, hours/week unpaid activities at the institution, and whether the current position is primary employment. Second, with both rotation methods, several of the factors were not

interpretable because variables associated with them seemed unrelated with each other. For example, Factor 24 in VARIMAX and Factor 21 in OBLIMIN had only respondent's gender and number of dependents with absolute loadings higher than 0.35; it might suggest some male and female faculty had different trends in reporting number of dependents, but the factor could not be interpreted as a reasonable construct. And finally, several interpretable factors had only few variables loaded on them. For instance, one factor about institution characteristics showed that the ratio of student FTE to faculty FTE shared variance with private vs. public school types.

Therefore, the EFA was rerun after 12 variables that appeared unrelated to others were excluded. This time, a total of 23 factors had an eigenvalue greater than one and accounted for 68.4% of the variance of the remaining 70 variables. All 23 factors were interpretable and they confirmed the major factors extracted from the 82 variables. Given the fact that the correlations between the factors in the OBLIMIN rotation were generally small (the highest was .26 between Factors 5 and 8; the two factors shared two variables) and the results were constant between the VARIMAX and OBLIMIN rotations, only the factors from the orthogonal rotation were presented in Table C3 in Appendix C.

KMC

To examine the variable relationships from a different angle, a KMC was run on the same group of 82 variables using Euclidian distance measuring variable similarity. In contrast to EFA, KMC does not assume linear relationships, but the calculation of Euclidian distance requires the data to be standardized before analysis if variables being

clustered are on different scales. For the same reason, categorical variables had to be excluded as well.

With clusters having only one variable removed mandatorily in the iterations, the FASTCLUS procedure in SAS grouped the 82 variables into 15 clusters when the convergence criterion was met. The variables in every cluster and their distance to the centroid are listed in Table C4 in Appendix C. The sizes of the clusters were quite uneven and several of the clusters did not have a clear meaning with a mix of seemingly unrelated variables. Nevertheless, when variables distant from the centroid were excluded from consideration, the meaning of major clusters became clear: administrative responsibility, publications (multiple clusters), other employment, research and grants, experience, institution parameter, work environment index, and teaching activities (multiple clusters including teaching load, contact with students, etc.). The KMC output confirmed the most salient factors suggested by the EFA.

In some cases, variables loaded on different factors in EFA were grouped together in KMC. For instance, most of the measures of teaching load, including number of credits hours, and total classes taught, were grouped into different factors in EFA, but they gathered in one cluster in the KMC analysis, suggesting that the interrelationships among these variables were nonlinear.

Some factors produced by the EFA clustered into a single group in KMC. For example, in KMC output, work support availability shared the same cluster with experience measures (e.g., years on current job, years at current rank; Cluster 4 in Table C4 of Appendix C); institution parameter was combined with variables related to

graduate teaching in Cluster 7; the differences between the KMC and EFA variable groupings were the result of different similarity measures and clustering techniques.

Twelve of the 15 clusters were clearly interpretable; several of the other clusters seemed like a mix of more than one theme. Generally, the major variable groupings shared by the KMC and EFA were very helpful in understanding the variable relationships and the overall data structure. The considerable overlaps of the EFA and KMA results suggested a robust solution for simplifying the variables structure.

MDS

Although MDS may not be very helpful in studying the data structure of such a large number of variables, it was tried with variables on all measurement scales included to see whether it could offer anything helping the understanding of the categorical variables that were not part of the 82 variables covered by the EFA and KMC.

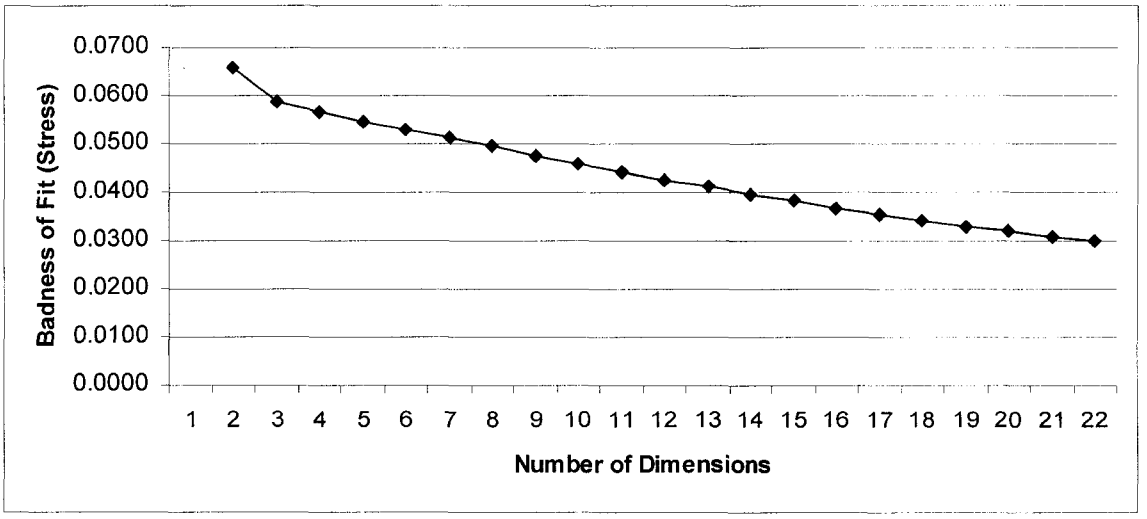


Figure 4.2. Scree plot of the Stress values in the MDS analysis: Number of dimensions ranged from 2 to 22.

Because of the inclusion of categorical variables, variable relationship was measured as dissimilarity with χ^2 independence, which is a statistic measuring the level of association between variables. The nonparametric MDS procedure in SAS was performed on the proximity matrix of the 90 variables with specified numbers of dimensions ranging from 2 to 22 in order to identify an optimally simplified structure that preserved well the relative positions of variables. Figure 4.2 is the scree plot of the Stress values; except for the not-very-obvious "elbow" at the three-dimension point, the curve showed a constant decrease of Stress values with the increasing number of the dimensions. The lack of clear dimensionality shown in the plot indicated that MDS is not appropriate when a large number of dimensions is expected in a complex variable space.

Final Variable Grouping

In order to keep as much as the original variance in variable extraction, the final grouping of variables followed several guidelines: First, every variable had a unique group identity; in other words, variable groups were disjoint and mutually exclusive. Second, clusters identified by both EFA and KMC were kept the way they were. Third, when several clusters in EFA were combined in KMC, they stayed in one cluster if a clear interpretation was available; otherwise, their separate cluster identities were kept as produced by EFA. If a variable loaded higher than 0.35 on more than one factor in EFA, the one with he highest loading was kept as the group identity. Finally, variables that did not have a clear cluster identity were carried untouched over to the regression model building along with the categorical variables.

The result of the variable space simplification was that 70 of the 82 variables were clustered into 17 groups; the groups and their variables are listed in Table C5 in Appendix C. Ten of the groups were distinct clusters that did not seem to overlap with each other: academic rank, administrative responsibility, beginning work status, education level, institution parameter, other employment, research, teaching, experience, and work environment index. The other seven groups were teaching (undergraduate committee), teaching (graduate), teaching (individual instruction), publications (books), publications (reviews), and institution parameter (miscellaneous).

Following the final grouping of variables, the next step was to extract one variable from each of the clusters. Because of the different clustering methods used, variables in the same group might not share linear relationships. Taking into consideration that the goal of the analysis was prediction, a method of extracting variables that account for more salary variance was desirable. Therefore, two different approaches were tried. One was to regress the log-transformed salary on variables in the same cluster, and select the variable having the greatest partial R^2 change. A second approach was to standardize the variables in the same group and then use their unweighted sum as one artificial predictor variable. A comparison of the results indicated that the first approach using 17 selected variables yielded an R^2 about .45, whereas the group of 17 artificial variables produced by the second approach yielded an R^2 of about 0.38. As a result, the first method was chosen; the 17 extracted variables are listed in the first section of Table C6 in Appendix C.

Model Building

The regression started with 37 independent variables: 17 extracted measures, one from each cluster, and 20 original variables (8 categorical variables and 12 variables that could not be clustered).

An initial run of the regression model with all the candidate predictor variables was tried to examine the contribution from each individual variable to the total variance of the dependent measure (LOGSAL). However, the model was not of full rank because one of the binary variables, STRATA11, recoded from Carnegie classification of institutions, was linearly dependent on another college classification variable, X08_0D, in the model. In other words, the least-square solution for the model parameters was not unique and some statistics might be biased and misleading. Therefore, X08_0D was removed to eliminate variable redundancy.

Another problem with the initial model was regarding the inclusion of BEA region codes as found in Model I. All eight binary variables were only different from the baseline of the U.S. service schools, but had little differences among themselves; and they brought strong multicollinearity to the model. However, excluding the five respondents in the baseline would make the sample different from the data used in the data mining procedure. Given that Model I still found BEA regions as significant factors using Far West as the baseline after the five respondents from U.S. service schools were removed, the nine categories of BEA were recoded for Model II with the Far West as the baseline while keeping the U.S. services school as a valid category.

A rerun of the regression model without X08_0D was done to identify variables having a significant relationship with the dependent variable. As shown in Table C7 in Appendix C, only 17 variables were significant at 0.01 level with forced entry variable selection, counting every categorical variable as one. The same group of predictor variables was also evaluated with stepwise selection, which ended with 19 variables as significant including the 17 significant variables identified by the forced entry method and two more: rank at hire for first job in higher education and the ratio of student FTE to faculty FTE. However, both of the two extra significant variables had a partial R^2 less than 0.001; such small values were found significant because of the large sample size. Due to the ignorable effect size, the two variables were excluded from the final regression model. Although significant in both variable selection methods, union status was also excluded because it had a small partial R^2 (.0009) at the expense of three degrees of freedom. The recoded BEA variables were significant when evaluated as a group with much lower VIF values.

The selected final model of sixteen predictor variables had 47 degrees of freedom (categorical measures recoded into binary variables). The model was also evaluated with the Max R and C_p procedures. The largest possible R^2 and the best possible C_p for models with 47 degrees of freedom was .5087 and 33.6, respectively, in comparison to .5036 and 82.7 of the selected model. Both procedures suggested models in which the nonsignificant binary variables created from the Carnegie classification and the BEA region code were replaced by other predictor variables. Given the constraint that binary variables from the same categorical variable have to be entered or excluded as a group,

the final model was a result of best efforts with some bias but a decent model R^2. The parameter estimates and model summary information are in Tables 4.3 and 4.4.

Table 4.3

Parameter Estimates of Model II

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
Intercept	Intercept	10.0399	0.0485	207.10	<.0001
Q29A1	Career creative works, juried media	0.0019	0.0002	11.87	<.0001
X15_16	Years since highest degree	0.0077	0.0004	17.82	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0011	0.0002	-6.04	<.0001
Q31A5	Time actually spent at administration (%)	0.0017	0.0003	5.95	<.0001
Q16A1REC	Highest degree type	0.0841	0.0050	16.68	<.0001
Q29A3	Career reviews of books, creative works	0.0018	0.0004	4.22	<.0001
Q76G	Consulting/freelance income	0.0000037	0.0000	5.75	<.0001
X01_66	Other aspects of job	0.0519	0.0058	8.89	<.0001
X01_8REC	Academic rank	0.0510	0.0031	16.27	<.0001
Q31A4	Time actually spent on professional growth (%)	-0.0023	0.0006	-3.86	0.0001
Q31A6	Time actually spent on service activity (%)	0.0013	0.0003	3.80	0.0001
Q81	Gender	-0.0667	0.0084	-7.97	<.0001
BEA region codes (Baseline: Far West)					
BEA1	New England	-0.0608	0.0058	8.89	0.0021
BEA2	Mid East	0.0082	0.0031	16.27	0.5788
BEA3	Great Lakes	-0.0545	0.0006	-3.86	0.0001

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
BEA4	Plains	-0.0868	0.0003	3.80	<.0001
BEA5	Southeast	-0.0921	0.0084	-7.97	<.0001
BEA6	Southwest	-0.0972	0.0198	-3.07	<.0001
BEA7	Rocky Mountain	-0.1056	0.0148	0.56	<.0001
BEA8	U.S. Service schools	0.1480	0.0142	-3.82	0.2879
Principal field of teaching/research (Baseline: legitimate skip)					
DSCPL1	Agriculture & home economics	-0.0279	0.0306	-0.91	0.3624
DSCPL2	Business	0.1103	0.0228	4.84	<.0001
DSCPL3	Education	-0.0643	0.0216	-2.98	0.0029
DSCPL4	Engineering	0.0695	0.0246	2.82	0.0048
DSCPL5	Fine arts	-0.0449	0.0241	-1.86	0.0627
DSCPL6	Health sciences	0.0933	0.0182	5.12	<.0001
DSCPL7	Humanities	-0.0641	0.0195	-3.29	0.001
DSCPL8	Natural sciences	-0.0276	0.0190	-1.45	0.148
DSCPL9	Social sciences	-0.0249	0.0202	-1.23	0.2173
DSCPL10	All other programs	0.0130	0.0194	0.67	0.502
Primary activity (Baseline: others)					
PRIMACT1	Primary activity: teaching	-0.0541	0.0169	-3.21	0.0013
PRIMACT2	Primary activity: research	-0.0133	0.0199	-0.67	0.5039
PRIMACT3	Primary activity: administration	0.0469	0.0203	2.31	0.0211
Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	0.0053	0.0236	0.22	0.8221
STRATA2	Private comprehensive	-0.0377	0.0263	-1.43	0.1525

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
STRATA3	Public liberal arts	-0.0041	0.0341	-0.12	0.9039
STRATA4	Private liberal arts	-0.0917	0.0260	-3.52	0.0004
STRATA5	Public medical	0.2630	0.0326	8.07	<.0001
STRATA6	Private Medical	0.2588	0.0444	5.82	<.0001
STRATA7	Private religious	-0.1557	0.0523	-2.98	0.0029
STRATA8	Public 2-year	0.0386	0.0247	1.56	0.1185
STRATA9	Private 2-year	-0.0061	0.0574	-0.11	0.9155
STRATA10	Public other	-0.0207	0.0563	-0.37	0.7127
STRATA11	Private other	-0.0879	0.0428	-2.06	0.0399
STRATA12	Public research	0.0792	0.0228	3.47	0.0005
STRATA13	Private research	0.1428	0.0259	5.51	<.0001
STRATA14	Public other Ph.D.	0.0005	0.0254	0.02	0.984

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Table 4.4

The ANOVA Table of Model II

Source	*df*	Sum of squares	Mean square	*F* value	$p > F$
Model	47	621.4482	13.2223	142.46	<.0001
Error	6599	612.4897	0.0928		
Corrected total	6646	1233.9379			

Note. Model R^2 = .5036, adjusted R^2 = .5001, and root MSE = 0.305

Model III: BBN Data Mining Model

The third prediction model was a Bayesian-network based data mining model. To start, all the 90 predictor variables and the basic salary of the academic year (SALARY) as predicted measure were imported into a database table. A program called *Data Preprocessor* as a component of the software *Belief Network Powersoft* read in the data and identified the variables on continuous scales for binning. Each continuous variable was cut into a specified number of bins in a way that same number of subjects was kept within each bin. The number of bins for each continuous variable was a subjective decision with two factors to consider: First, binning decreases the measurement accuracy; the smaller the number of bins, the greater the bin width, and the more the loss of information. Second, the construction of a BBN model requires an enormous amount of computation; small bin width leads to many bins, and so dramatically increases the variable product state space and therefore the required computational time and resources. The binning information of the continuous variables is listed in Table D1 in Appendix D.

The learning of a good BBN prediction model was an automated process once the binning of continuous variables was completed and the predicted variable was specified. To compare the findings of the automatic learning from data with the results of statistical analysis, the data mining started without any pre-specified knowledge that might include the order of variables in some dependence relationships, forbidden relations, or known causal relations. The software also makes it possible for users to decide a threshold value that determines how strong a mutual relationship between two variables is considered as meaningful; relationships below the threshold are omitted from subsequent network

structure learning (Chen & Greiner, 1999). If no threshold value is set, the program can make automatic selection in the learning process, but this overhead makes learning slower and more expensive.

When the threshold value is low, relatively weak dependencies are kept in the learning process and the final model consists of a relatively large number of predictor variables and too many arcs (dependency relationships) in a complex structure (Chen, Greiner, Kelly, Bell, & Liu, 2001). If too many variables and their conditional dependencies are kept, the model can be overfit, meaning that it only works well with the training data set but would have poor prediction accuracies on new data sets. Adjusting to a higher threshold value can alleviate overfit by building a model with a smaller number of predictor variables connected by relatively strong relationships. However, when a threshold is too high, it causes missing arcs and often results in very simple structure, a phenomenon called underfitting.

To search for an optimal model, a number of BBN learning processes were conducted, each with a different threshold value specified. When threshold was set to eight times the system default value, the learned network had 8 variables and 21 edges connecting them. The network had a classification accuracy of 46.84% on the training data set but only 10.57% for testing data set, a sign of overfit.

As shown in Table 4.5, the classification accuracy of the training data set decreased when the threshold value increased, but the prediction accuracy with the testing data set reached its peak at a threshold value 12.5 times the default. Noticeably, the two BBN models with threshold values 10 and 12.5 times the default value selected

the same group of six variables, indicating model stability; but the former had three more CP arcs (13 vs. 10), which helped the classification rate of the training data, but slightly harmed the prediction in the testing data. When the threshold increased to 15 times the system default value, the classification accuracy rate on both the training and the testing data sets were the lowest among the five models, a clear indication of underfitting model structure.

Table 4.5

BBN Models with Different Threshold Values Specified

			Prediction Accuracy			
Threshold	No. of variables	CP	Training data (total: 6652)		Testing data (total: 3311)	
			No. of correct classification	Percent % (std.)	No. of correct classification	Percent % (std.)
8	8	21	3116	46.84 (0.60)	350	10.57 (0.53)
10	6	13	2066	31.06 (0.55)	372	11.24 (0.54)
12.5	6	10	1707	25.66 (0.53)	383	11.57 (0.55)
15	4	7	1297	19.50 (0.48)	340	10.27 (0.57)

Note. Number of variables does not include the predicted variable (SALARY). Numbers in "Threshold" column are times of the system default value.

Because generalizability to new data sets is an important characteristic of prediction models, the comparison of model parameters suggested that the model with the threshold value 12.5 times the system default value as the best BBN model learned: six variables connected by 10 CP arcs as shown in Figure 4.3.

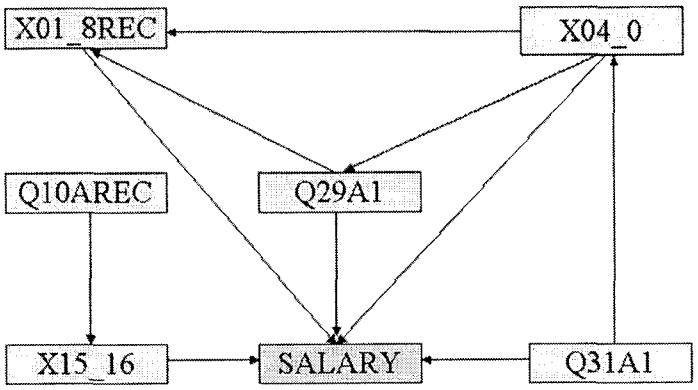


Figure 4.3. The BBN model of salary prediction. The CP tables are not included to avoid complexity. The definitions of the seven variables are

a. SALARY: Basic salary of the academic year.
b. Q29A1: Career creative works, juried media
c. Q31A1: Percentage of time actually spent teaching undergrads
d. X15_16: Years since highest degree
e. X01_8REC: Academic rank
f. X04_0: Carnegie classification of institution
g. Q10AREC: Years on achieved tenure

Model IV: Combination Model

The final prediction model produced by the Belief Network Powersoft had six predictor variables as shown in Figure 4.3. However, one of six, number of years since achieved tenure (Q10AREC), was only connected to another predictor variable (i.e., years since the highest degree), a strong relationship substantiated by their Pearson correlation (r = .64). Q10AREC also had a strong correlation with academic rank (r =

.43), another variable in the model. After a test confirmed that Q10AREC was not a suppressor variable, it was excluded from the regression model. Therefore, the last regression model had only five independent variables. Among them, the Carnegie classification of institutions as the only categorical measure was recoded into binary variables.

With log-transformed salary as the dependent variable, the process of building Model IV was simple because all five variables were significant at $p < .0001$ with both forced entry and stepwise variable selections (Table E1 in Appendix E). All-possible-subset evaluation procedures were not necessary because all variables were included. The model summary information is presented in Tables 4.6 and 4.7.

Table 4.6

Parameter Estimates of Model IV

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
Intercept	Intercept	10.5410	0.0272	387.28	<.0001
Q29A1	Career creative works, juried media	0.0024	0.0002	15.34	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0030	0.0002	-20.06	<.0001
X01_8REC	Academic rank	0.0664	0.0032	21.01	<.0001
X15_16	Years since highest degree	0.0088	0.0004	19.97	<.0001
Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	-0.0385	0.0250	-1.54	0.1236
STRATA2	Private comprehensive	-0.0645	0.0281	-2.29	0.0218

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
STRATA3	Public liberal arts	-0.0315	0.0363	-0.87	0.3853
STRATA4	Private liberal arts	-0.1221	0.0276	-4.42	<.0001
STRATA5	Public medical	0.2933	0.0339	8.66	<.0001
STRATA6	Private Medical	0.2915	0.0471	6.20	<.0001
STRATA7	Private religious	-0.2095	0.0551	-3.80	0.0001
STRATA8	Public 2-year	-0.0403	0.0258	-1.56	0.1179
STRATA9	Private 2-year	-0.0371	0.0611	-0.61	0.544
STRATA10	Public other	-0.0245	0.0594	-0.41	0.6802
STRATA11	Private other	-0.0871	0.0456	-1.91	0.0563
STRATA12	Public research	0.0479	0.0242	1.98	0.0472
STRATA13	Private research	0.1543	0.0276	5.60	<.0001
STRATA14	Public other Ph.D.	-0.0496	0.0268	-1.85	0.0648

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Table 4.7

The ANOVA Table of Model IV

Source	*df*	Sum of Squares	Mean Square	*F*	Pr > F
Model	18	520.2949	28.90527	268.4	<.0001
Error	6632	714.3279	0.10769		
Corrected Total	6651	1234.6228			

Note. Model R^2 = .4214 and adjusted R^2 = .4199. Root MSE (standard error of estimate) = 0.3282.

Model Comparison

In this section, the three multiple regression models--Models I, II, and IV--were compared and evaluated with model fit, model simplicity, and model generalizability. The data mining model--Model III--was evaluated in light of the statistical regression models on characteristics including input, algorithms, output, model interpretations, prediction accuracy, and so on.

Comparison of Multiple Regression Models

The first difference among the three multiple regression models were the variable reduction methods used. The theoretical model, Model I, started with 13 variables that were suggested as most relevant in faculty salary equation by previous studies, and produced a final model of 11 independent variables with a R^2 of 0.4213 (df = 26 and adjusted R^2 = .4190). The statistical model, Model II, started with all 90 variables in the pool, and identified 17 variables from the 70 variables that could be clustered with statistical techniques. Along with the ungrouped 20 variables, a total of 37 independent variables were input in the initial model, and 16 of them stayed in the a final model with a R^2 of .5036 (df = 47 and adjusted R^2 = .5001). The combination model, Model IV, had the fewest number of predictor variables to start with, and the five independent variables identified by the BBN data mining model were all significant at $p < 0.0001$, and resulted in a final model with a R^2 = .4214 (df = 18 and adjusted R^2 = .4199). Meanwhile, a greater R^2 of Model II also made its standard error of estimates the smallest among the three (root MSE = 0.305 vs. 0.328 of Model I and IV).

Although Model II had the greatest R^2 of the three regression models, it also had the largest model degrees of freedom. Its R^2 was higher than that of Model I by about .0823 with 21 more variables, which indicated that each additional variable added only about .0039 to the model R^2. Similarly, Model II had a R^2 about .0822 higher than Model IV at the expense of 29 more variables in the model. Each additional variable only increased the model R^2 by .0028 on average.

One of the negative effects associated with large numbers of independent variables in a multiple regression model is the threat of multicollinearity caused by possible strong correlations among the predictors. Model R^2 always increases with the number of predictor variables, but if the variables bring along multicollinearity, estimated model parameters can have large standard errors, leading to unreliable models. For the three regression models, Model I had 15 variables with a VIF > 1.5 out of the 26 total (58%); Model II had 31 out of 47 variable with a VIF > 1.5 (66%). Model IV had 10 out of 18 variables with a VIF > 1.5 (55%), and most of high VIF values were associated with the binary variables recoded from categorical variables.

Table 4.8

Residual Statistics of the Three Regression Models

Residuals	Mean	Standard deviation	Range	Skewness	Test for normality Kolmogorov-Smirnov
Model I	0.0000028	1.00014	13.761	-0.104	0.063 ($p < .01$)
Model II	-0.0000169	1.00033	14.857	-0.368	0.070 ($p < .01$)
Model IV	-0.0000022	1.00020	13.300	-0.054	0.057 ($p < .01$)

Another evaluation criterion of model appropriateness is the distribution of residuals. An unbiased model should produce residuals of normal distribution and scatter randomly around zero without specific patterns. The residual statistics in Table 4.8 suggested little differences among the three models. Although Model IV seemed better on most the measures, none of the differences was substantial. The scatter plots of the standardized residuals against the predicted values of the three models, although lacking perfect randomness, did not have any strong differences from each other (as shown Figure F1, F2 and F3 in Appendix F).

Because the OLS method in prediction analysis produces a regression equation that is optimized for the training data, model generalizability should be considered as another important index for good prediction models. Model generalizability was examined by cross validating the three models with the holdout testing data set. The three regression equations were applied to the 3,311 records in the testing data to obtain their predicted values, and the R^2s of Model I, II, and IV on the testing data set was found to be .4147, .5055, and .4489, respectively, as compared with .4213, .5036, and .4214 in the original data set. It is unusually for R^2 for the cross-validation set to exceed that of the data set from which the parameter estimates were obtained, but in point of fact, the R^2s for the testing data set were even larger when the OLS estimates of the parameters for the variables were directly obtained from the cross-validation set (.4453, .5301, and .4514; Models I and IV included the same sets of variables as those based on the training data, but Model II had one less variable than that of the training data). The greater R^2 for the cross-validation data set may also indicate the training data set could be idiosyncratic, an

aspect of the data analysis that will be discussed later in this chapter. Given the current findings, Model IV with only five variables seemed to work better with new data than Model I by a difference about .0342 in the R^2. Model II again had the greatest R^2 of all.

To further understand the differences, the variables with which the three models started with were compared. To build Model II, 70 dichotomous and continuous variables were grouped into 17 clusters based on variable similarity. From each of the clusters one variable was selected that had the greatest R^2 with the predicted variable. With a clear goal of prediction, it was an exploratory process without theoretical considerations from variable reduction through model building. The underlying theme was to find the variables that helped salary determination.

Comparing to Model I that started with compensation theories, Model II had more variables and a greater R^2 value. However, if the stepwise analysis results of the two models were laid side by side (Table B1 in Appendix B and Table C8 in Appendix C), the similarities of the two models became clear. Nine out of the 11 variables in Model I had close or equivalent measures in Model II (shown in Table 4.9), and these measures were among the first twelve variables that entered Model II. However, Model I had a R^2 of .4213 with 11 variables (adjusted R^2 = .4190), whereas the first 12 variables in Model II accounted for a R^2 of 0.4989. Was the .0776 increase in the R^2 due to different choices of variables on measures of publication (Q29A1 vs. TOTPUB), administration (Q31A1 vs. Q13), experience (Q25A5REC vs. X15_16), and classification of institution (X04_0 vs. X08_0D), or mainly because predictor variables increased from 26 to 43?

Table 4.9

The Comparable Variables in Model I and Model II

Model I		Model II	
Variable (*df*)	Definition	Variable (*df*)	Definition
X01_8REC (1)	Academic rank	X01_8REC (1)	Academic rank
X08_0D (1)	Institution type - doc, 4-year,or 2-year	X04_0 (14)	Carnegie classification of institution
DISCIPLINE (10)	Principal field of teaching/research	DISCIPLINE (10)	Principal field of teaching/research
Q24A1REC (1)	Years since first job in higher education	X15_16 (1)	Years since highest degree
Q16A1REC (1)	Highest degree type	Q16A1REC (1)	Highest degree type
TOTPUB (1)	Total number of publications	Q29A1 (1)	Career creative works, juried media
Q81 (1)	Gender	Q81 (1)	Gender
X37_0 (7)	Bureau of Economic Analysis (BEA) region code	X37_0 (8)	Bureau of Economic Analysis (BEA) region code
Q13 (1)	Chair of a department	Q31A5 (1)	Time actually spent at administration (percentage)

The answer became clear when Model IV was compared with Model I and II. Model IV had only five variables selected by the BBN model, and they were among the

top six variables in the stepwise selection of Model II. Model IV was also the product of a data-driven process; both models captured variables that share strong covariance with the predicted variable. The overlap of Models II and IV was more than a coincidence.

Table 4.10

The Comparable Variables in Model I and Model IV

Model I		Model IV	
Variable (*df*)	Definition	Variable (*df*)	Definition
X01_8REC (1)	Academic rank	X01_8REC (1)	Academic rank
TOTPUB (1)	Total number of publications in career	Q29A1 (1)	Career creative works, juried media
X08_0D (1)	Institutional type: doctoral, 4-yr, or 2-yr	X04_0 (14)	Carnegie classification of institution
Q24A1REC (1)	Years since first job in higher education	X15_16 (1)	Years since highest degree
X04_41 (1)	Total classroom credit hours taught	Q31A1 (1)	Percentage of time actually spent teaching undergrads

Note. X04_0 in Model IV is a categorical variable of 15 values. It was recoded into 14 binary variables in the regression model.

There were similarities between Model I and Model IV (Table 4.10). Although academic rank (X01_8REC) was the only variable that appeared in both; the two models both had variables related to publication, experience, and institutional type. The pairs of

comparable variables on the same constructs were part of similarities between Model I and Model II because Model II and Model IV shared those variables. In Model IV, there was a significant teaching-related variable: the percentage of time spent teaching undergraduate students, which had a negative impact on the basic salary. The final model of Model I did not have any teaching related variables because teaching responsibility quantified as total classroom credit hours taught during 1998-99 academic year was excluded due to nonsignificance.

To summarize the major variable differences, Model I used the total number of publications (TOTPUB) that included all the juried and non-juried creative works, reviews, books, presentations, and other types of publications in one's career, whereas Models II and IV had only publications in juried media. Institutions were grouped into three levels in Model I: doctoral, 4-year college, and 2-year college, but Models II and IV had the Carnegie classification that specified 15 categories of postsecondary institutions. Model I had the years working in higher education as a measure of experience, comparing to the number of years since highest degree in Models II and IV.

To find out whether the different choices of variables or the greater number of predictors made a difference in the model R^2, another regression model, Model V, was built to check whether Model I would be improved when the variables on publication, teaching, administrative responsibility, and experience being replaced by those in Models II or IV. The measure of institution type stayed unchanged to avoid a big leap in model degrees of freedom; five respondents from the U.S. service schools were excluded. Again, ethnicity and the interactions between gender and academic rank and between

gender and tenure status were nonsignificant. The interaction between teaching and publication was excluded as well because it had a partial R^2 of only .0007 in stepwise selection (Table F1 in Appendix F). With 12 significant variables and $df = 27$, the final Model V had a $R^2 = 0.4703$, adjusted $R^2 = 0.4682$, and the standard error of estimates = 0.314 (Table F2 in Appendix F). The approximate .05 increase in the R^2 of Model V from Model I was evident that the variable choices made a strong difference.

Because all three models had some strengths and weakness, it is difficult to choose one as the best. Model II had the greatest R^2, but its advantage was offset by the large number of variables and so a big loss in the degrees of freedom. The simplicity of Model IV was a strong plus; it had the same R^2 as of Model I, but many fewer predictor variables. Model I could be improved with different choices of variables, but as a model with theory-driven variable selection it did not originally make an impressive R^2.

The Data Mining Model in Contrast with Multiple Regression Models

In this section, the data mining model of salary prediction is discussed in contrast to regression Model IV because the same group of variables they shared help to concentrate the discussion on their fundamental differences.

Theoretically, the learning algorithms of a BBN data mining model has no limit on the number of input variables; but onerous computational resources are required when the number of variables is large. Computationally, the number of observations in a data set does not have as strong an influence on the learning algorithms as they do in traditional statistics because binning of variables into categories and use of variable association measures minimize the role of sample size in model learning.

Variable Transformation

In contrast to regression models that explicitly or implicitly recode categorical data, Bayesian algorithms keep the categorical variables untouched, but usually bin continuous variables into intervals. The loss of information associated with this process is a threat to model accuracy, but it helps to relax model assumptions and so BBN requires no linear relationships among variables. The assumptions of linearity (higher model R^2) and normality (assumed distribution of the residuals) were the reason for the log-transformation of SALARY as the dependent variable in the multiple regression models; whereas in the BBN model, the original SALARY measure was used because neither linearity nor normality is required.

As a matter of fact, the binning of continuous variables makes BBN a robust nonmetric procedure that is not influenced by any monotonic variable transformation. During the binning, a continuous variable is divided into a number of intervals of various widths so that a same number of subjects are included in each interval. Thus, monotonic variable transformations make no difference in binning as long as the same number of intervals is specified. Furthermore, the bins are treated as categorical variables rather than sequential values. As a result, monotonic transformations would not make any difference in the output model given the same binning schema and the fact that variable relationships are measured as associations.

Learning Variable Relationships and Finding the Optimal Model

In multiple regression analysis, variable relationships are measured by linear correlations. In contrast, Bayesian network discovery uses some statistical tests (e.g., χ^2

test of statistical independence) to compare how frequently different values of two variables are associated with how likely they happen to be together by random chance for building conditional probability statistics among variables (Chen et al., 2001). Such measures of variable associations do not assume any probabilistic forms of variable distributions.

With a large number of input variables, a BBN learning process can be computationally overwhelming given the back-tracking processes in the network discovery. Two types of techniques help to alleviate this problem. First, the assumption of conditional independence, which states that each variable is independent of its nondescendants in the network given the state of its immediate parents, simplifies the quantification of variable relevance and the network structure. That is, conditional independence helps to reduce the number of parameters needed to characterize the probability distribution and therefore helps to compute efficiently the posterior probabilities with given evidence (Friedman, Geiger, & Goldszmidt, 1997). Bayesian probability and the conditional independence assumption are the theoretical grounds for constructing BBNs and presenting variable relationships in graphs (e.g., Figure 4.5) with CP tables attached to the variables.

Secondly, users are allowed to specify a threshold value to exclude weak dependency relationships between variables from the model learning to speed up the process. With a specified threshold value, not only can the complexity of learning process be decreased, but also sensitivity of statistical tests to minor differences is overridden and problems caused by large sample sizes are no longer an issue.

In the automated model learning process, numerous candidate models are evaluated with certain evaluation criteria called *score functions*, and the best one is output as the optimal choice with the best prediction accuracy. In other words, the discovery of an optimal BBN is a search in model space which may consist of candidate models of substantially different structures. In the current study, a second phase of model selection happened when different threshold values were specified to output five optimal models for manual evaluation. In multiple regression analysis, model comparison is also part of the analysis procedure, but the structures of the candidate models are mostly similar in a nested schema, and human intervention in model selection is usually necessary.

Model Presentation

As a result of different approaches summarizing data and different algorithms analyzing data, the outputs of the BBN model and the multiple regression models were different. The final result of multiple regression can be presented as a mathematical equation. For example, Model IV can be written as

$$\text{Log (Salary)} = 10.5410 + 0.0024 \times \text{Q29A1} - 0.0030 \times \text{Q31A1} + 0.0664 \times \text{X01\_8REC} + 0.0088 \times \text{X15\_16} - 0.0385 \times \text{STRATA1} - 0.0645 \times \text{STRATA2} - 0.0315 \times \text{STRATA3} - 0.1221 \times \text{STRATA4} + 0.2933 \times \text{STRATA5} + 0.2915 \times \text{STRATA6} - 0.2095 \times \text{STRATA7} - 0.0403 \times \text{STRATA8} - 0.0371 \times \text{STRATA9} - 0.0245 \times \text{STRATA10} - 0.0871 \times \text{STRATA11} + 0.0479 \times \text{STRATA12} + 0.1543 \times \text{STRATA13} - 0.0496 \times \text{STRATA14} + \text{error}. \quad (4.1)$$

If a respondent received the highest degree three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of the work time teaching

undergraduate classes (Q31A1 = 20) as an assistant professor (X01_8REC = 4) in a public research institution (STRATA12 = 1 and all other STRATA variables were 0), the predicted value of this individual's log-transformed salary should be 10.83 according to Equation 4.1 (about $50,418), with an estimated standard error as uncertainty measure.

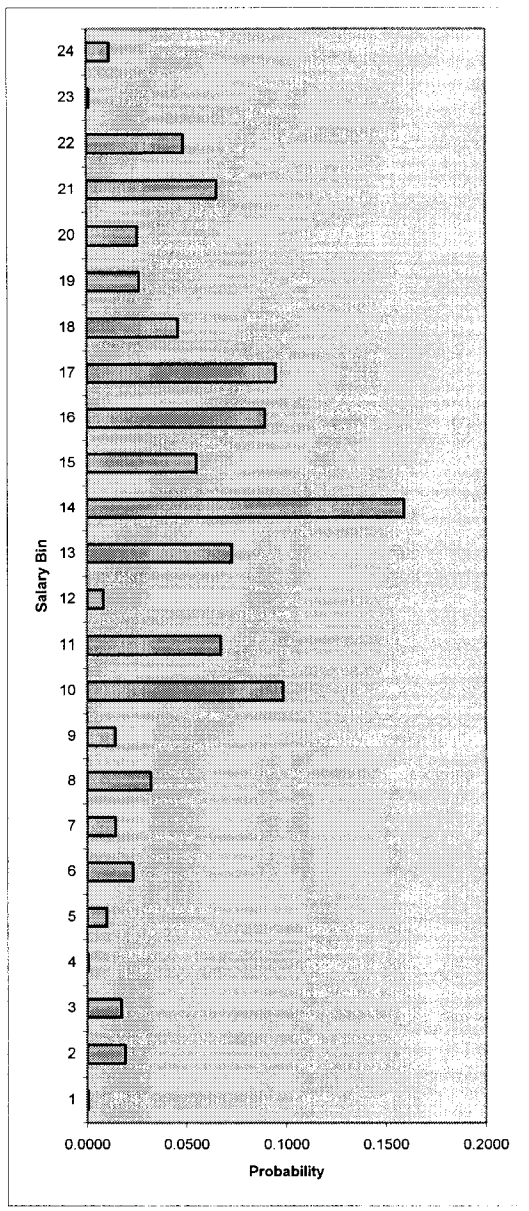
The result of the BBN model is presented in a quite different way. For the above case, the BBN model would make a prediction of salary for such faculty with a salary conditional probability table as shown in Table 4.11. The predicted salary fell in a range between $48,325 and $50,035 because it had the highest probability (p= 15.9%) in the CP table for this particular combination of variable values. A CP table like this was available for every unique combination of variable values (variable product state space).

The point estimation of a conditional mean in most statistical predictions implicitly expresses the prediction uncertainty with a standard error of estimation based on the assumption of normal distribution. In contrast, the BBN model makes predictions based on the distributional mode of the posterior probability of the predicted variable. The presentation of posterior probability as a random variable explicitly expresses the prediction uncertainty in term of probability. Without the assumption of normality, the conditional probability of the predicted values is the outcome of binning continuous variables and treating all variables as categorical class labels in the computation. The prediction based on the mode of a probabilistic distribution is another robust feature of the BBN: the mode is not sensitive to outliers or skewed distribution as the arithmetic mean is. However, one problem of the classification approach is that it is difficult to tell how far the predicted value missed the observed value when a case was misclassified.

Table 4.11

An Example of the BBN Conditional Probability Tables

Bin #	Salary range	Probability
1	Salary < 29600	0.0114
2	29600 < Salary < 32615	0.0012
3	32615 < Salary < 35015	0.0487
4	35015 < Salary < 37455	0.0655
5	37455 < Salary < 39025	0.0254
6	39025 < Salary < 40015	0.0263
7	40015 < Salary < 42010	0.0460
8	42010 < Salary < 44150	0.0950
9	44150 < Salary < 46025	0.0894
10	46025 < Salary < 48325	0.0552
11	48325 < Salary < 50035	0.1590
12	50035 < Salary < 53040	0.0728
13	53040 < Salary < 55080	0.0081
14	55080 < Salary < 58525	0.0672
15	58525 < Salary < 60010	0.0985
16	60010 < Salary < 64040	0.0140
17	64040 < Salary < 68010	0.0321
18	68010 < Salary < 72050	0.0142
19	72050 < Salary < 78250	0.0228
20	78250 < Salary < 85030	0.0098
21	85030 < Salary < 97320	0.0005
22	97320 < Salary < 116600	0.0170
23	116600 < Salary < 175090	0.0190
24	175090 < Salary	0.0005



Note. Salary was binned into 24 intervals. For this particular case, the product state is that the highest degree was obtained three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of the time teaching undergraduate classes (Q31A1 = .2) as an assistant professor (X01_8REC = 5) in a public research institution (STRATA = 12 and all other binary variables were 0).

The predication accuracy of the model was the ratio of the number of correct classifications to the total number of predictions. In multiple regression, predication accuracy is usually quantified with residuals or studentized residuals; also, the model R^2 is an index of how well the model fits the data. For example, Model IV had a R^2 of .4214 and .4489 on training and testing data, respectively, which was considered a decent level in regression based on such a complex data set. However, the prediction accuracy of the BBN model was only 25.66% and 11.57% on training and testing data, respectively.

Several possible explanations for this low prediction accuracy: First was the loss of information when continuous variables are binned: four of the five predictors were on an interval or ratio scale. Second, the final class identity of an individual case was only determined to be the salary bin that had the highest probability, which might not be substantially strong when the predictor variable was divided into many narrow bins (as in the above example $p = .16$). Also, when the bin widths are relatively narrow, misclassification may increase due to weakened differences among the levels of a variable. Finally, scoring functions used for model evaluation in the Bayesian network learning could be another factor. According to Friedman et al. (1997), when the structure of the network is not constrained with any prior knowledge as in the current case, nonspecialized scoring functions may result in a poor classifier function when there are many attributes.

A point estimate of the predicted value is possible in the BBN prediction: it can be calculated as the weighted sum of the mid-point of each interval; the probabilities are used as the weights. The validity of such an approach is beyond the scope of this study.

Prediction Accuracy

It is difficult to compare the prediction accuracies of the BBN model and the multiple regression models because the former defines accuracy as *correct* vs. *incorrect* (dichotomous scale), whereas the latter quantify the accuracy as how far away the predicted and observed values are (ratio scale). Because it is difficult to tell how far the predicted salary is from the observed values in BBN classification, measures of prediction accuracy for the BBN and multiple regression are not directly comparable.

One possible way to overcome this difficulty is to rank order the salary bins. If the observed value is in bin #A and the predicted values in bin #B, the absolute value of A-B can provide some knowledge about the distance between the observed and predicted values. The information is quite rough because that the bin widths are uneven, but it provides an opportunity to compare the multiple regression models with the BBN model with regard to their prediction accuracy.

To compare the prediction accuracies of the BBN model and the multiple regression models with this method, the observed and the predicted salary values in the regression models were binned into 24 intervals according to the same schema used in the BBN model building. The bins were rank ordered as shown in Table 4.11. For each regression model, the bin difference of the predicted and observed values was calculated for individual cases. If the predicted and the observed salary values of a respondent fell into the same bin, the bin difference is 0; otherwise, the difference is the absolute value of the bin difference between the predicted and the observed values. For example, if the observed salary was $38,000, its bin number was 5 ($37,455 ~ $39,025); and the

predicted value was $41,000, the bin number was 7 ($40,015 ~ $42,010). Then the difference is 2 (i.e., 7-5), indicating the predicted value was off by two bins from the observed value.

As shown in Tables F3 (for training data) and F4 (for testing data), the BBN Model III had a larger number of correct classifications in the training data than the three regression models (1707 vs. 637, 688, and 566 for Models I, II, and IV, respectively), but the differences among the four models substantially decreased if all the cases were taken as "accurate prediction" as long as the predicted value stayed within three-bin distance from the observed value (for Models I to IV respectively, the number of "accurate" prediction were 3758, 56.5%; 4036, 60.7%; 4053, 60.9%; and 3714, 55.8%).

The analysis results reviewed that the prediction accuracies of the regression models was stable in the training and testing data sets either by the model R^2s or by the classification accuracies in terms of binned values, indicating the linear relationships between the predictor and predicted variables stayed essentially constant across samples. However, the prediction accuracy of the BBN model decreased substantially from 25.66% for the training data to 11.57% for the testing data, a sign of model instability in generalization. However, the BBN model was relatively stable if all observations having predicted values within three-bin distance from the actual salary value were considered as "accurate prediction" (60.9% for the training data and 54.8% for the testing data), a piece of strong evidence that the poor stability of prediction accuracy of the BBN model was a negative outcome of binning the continuous variables in the network discovery process.

Robustness against Outliers

Bayesian networks are credited with their robustness against outliers. *Outlier* is a concept associated commonly with continuous variables, indicating cases that are far different from others. Bayesian networks technically turn continuous measures into categorical values through binning. For example, the distribution of SALARY suggested all cases greater than $190,000, a value more than four standard deviations beyond the mean, should be considered as outliers (Figure A1 in Appendix A). However, since all cases with a salary higher than $175,090 were grouped together in BBN, they shared the same class label even though the highest salary in this group was $250,000, 43% greater than the lower limit. The network learning was to induce salary rules from the binned data. The tradeoff of robustness with the binned data was that the prediction of salary might not be able to meet some application requirements when bin sizes are very wide. In addition, the prediction of output values in BBN is based on the distributional mode, which is not as strongly influenced by outliers as the predicted conditional mean.

In multiple regression models, in addition to the poor predictive accuracy at extreme values, outliers can also have strong influence on the prediction equation and model accuracy. Thus, a number of techniques are available for identifying outliers. For example, when a simple and clean data set is analyzed, cases with studentized residuals more than three standard deviations away from the mean are generally taken as outliers. When the problem turns more complex with data of high dimensionality, the multivariate outliers are peculiar combinations of measures. More sophisticated techniques, such as Cooks' distance and leverage statistics, are necessary to identify cases that unduly

influence the outcome and that have unusual predicator values. A plot of Cook's distance against the predicted values is shown in Figure F4 in Appendix F; a number of cases stood out in the plot. Those cases might have negative influence on the model accuracy; however, their effect could be hardly noticeable given the large sample size.

Conditional Dependency in BBN vs. Variable Correlation in Regression

Another aspect of the BBN model worth studying is the conditional dependency relationships among some of the predictor variables, as shown in Figure 4.3. Because the relationships were identified with mutual information tests, they were directional, suggesting possible causality.

Comparing to the correlation measures in the regression analysis, it is clear that variables with identified conditional relationships had relatively strong correlations. For instance, number of publications in juried media (Q29A1) and academic rank (X01_8REC) were correlated at $r = .32$; years since tenured (Q10AREC) and years since highest degree (X15_16) had $r = 0.64$. However, not all the correlations as strong as these were captured in the BBN as conditional dependency. The $r = .43$ between X01_8REC and Q10AREC was not shown in the network; neither were the correlations between X01_8REC and X15_16 ($r = .35$) and between Q29A1 and X15_16 ($r = .32$). Again, loss of information in binning could be one of the major causes for missing some of the high correlations in the network. Another reason was that conditional dependencies in BBN were not measured by linear relationships.

There were also conditional dependencies between the categorical variable Carnegie classification of institutions (X04_0) and three of the measures X01_8REC,

Q29A1, and Q31A1 (percentage time spent teaching undergrads) that were on ordinal, interval, or ratio scales. Statistically, relationships between nominal and other types of variables cannot be measured by linear correlation, but a simple ANOVA indicated that the means of X01_8REC, Q29A1, and Q31A1 in different X04_0 groups were significantly different at $p < 0.001$ (the estimated effect size $\eta^2 = .094$, .160, and .337, respectively). When the relationships identified by BBN are worth pursuing, more designed studies can be conducted to follow up and to substantiate related theories.

Conditional dependencies between predictor variables help to improve the prediction accuracy of a BBN model, although they may cause problems of overfitting when there are too many such relations included in the network. However, strong correlations among predictor variables in multiple regressions are unwelcome because they can lead to multicollinearity, a problem causing model instability and prediction inaccuracy.

Missing Data

The current data set had few missing data after the preprocessing. How to deal with missing data is a research topic in all data analysis fields. In Bayesian network learning, missing data can be estimated from related known variables and their conditional probabilities. Replacing missing data through Bayesian framework has been long recognized in statistics; studies (e.g., Kontkanen, Myllymäki, Silander, & Tirri, 1997) have been conducted to examine the performance of Bayesian probability in estimating missing elements in data matrices. Although the computational overhead remains expensive if all variable dependencies are considered in estimating missing data

elements, the current study showed that Bayesian network with feature selection makes it is possible to construct optimal probabilistic models using only a moderate number of parameters, indicating that using Bayesian probabilistic modeling to deal with missing data can be as well feasible even for massive data sets. In most traditional statistical techniques, cases with missing data are often excluded from analysis. Techniques of replacing missing data exist, but they may not be appropriate in small samples.

Generalizability of the Findings

The multiple regression models had greater cross-validation R^2s on the testing data than the model R^2s obtained from the data with which the parameters were estimated. The uncommon phenomenon was an indication of possible idiosyncrasy of the training data set and called into question whether the findings from the four prediction models were generalizable to different data sets. To make certain that the conclusion of the study was not based on idiosyncratic sample data, one more step was taken in the data analysis: the entire data set of 9,963 cases was randomly divided into four samples and each of the subsets was used to build the four models independently. The results of the analyses are presented in Table F5 to F8 for Models I, II, III, and IV in Appendix F.

Four Model Is were built from the four subsets. Although the models were different with respect to the inclusion of some predictor variables with small partial R^2s (e.g., Q81 and Q13), all models were consistent with regard to the major predictor variables. The variation of the model R^2s was evidence that the four subsets consisted of somewhat different samples; however, the model structures (e.g., major predictor variables and the estimated model parameters) appeared to be stable across samples.

Model I based on the original training data had a R^2 roughly equivalent to the average of the four subset models and retained more predictor variables (related to its much larger sample size). Given the large sample size and diverse subjects in the samples, the stability of the model structure supports the generalizability of the findings of the current study.

Similar results were found in Model IIs, Model IIIs, and Model IVs constructed from the four subsets. One interesting difference between the BBN models and the multiple regression models was also uncovered: With a smaller sample size, fewer predictor variables were found significant when building multiple regression models. In contrast, the BBN models contained more variables when the sample size was smaller. This difference is related to the statistics used for measuring variable relationships and the methods for determining significance. In multiple regression, the significance of linear correlations between variables are determined by statistical tests, which are sensitive to minor differences when the sample size is large. Therefore, more variables have significant relationships with the predicted variables given a larger sample size.

When building a BBN model, the variable associations are measured with χ^2 statistic, but the significance of a relationship is not determined with statistical tests; rather, the users can specify different threshold values and search for the optimal results through multiple trials. With a smaller sample size, more information is necessary to stabilize the model structure and provide accurate prediction. Thus, a smaller sample size often results in a model with more predictor variables. In general, the results of the subset models confirmed the stability of the models generated from the original training data and the generalizability of the findings and conclusions of this study.

CHAPTER 5 CONCLUSIONS

The objective of this study was to explore the potential benefits of using data mining techniques in studying large data sets or databases in educational research. With data analysis procedures and results presented in the previous chapters, it is time to answer the questions this research started with.

Out of the rich collection of different data mining techniques, the Bayesian network was chosen as a prediction (classification) model in contrast to multiple regression models for salary prediction with the same faculty compensation data set. Therefore, conclusion and answers to the research questions are based on the analyses of the BBN and the multiple regression models.

Research Question I: Comparison of Data Mining and Statistics

The first research question was about the similarities, differences, strengths, and weaknesses of data mining and traditional statistics when doing predictions with large data sets. Theoretical discussion on the similarities and differences of the two approaches were provided in the literature review, and the points were further illustrated in the data analysis and the result presentation. Here, the strengths and weakness of statistics and data mining--to be specific, the BBN model--are discussed in the context of prediction problems.

Large Sample Size

With a large volume of data, both the large numbers of observations and the large number of variables can cause trouble in traditional statistical analysis. The analysis in this study confirmed that statistical significance tests are oversensitive to minor

differences when the sample size is large. For example, as shown in Table C5 in Appendix C, a few variables with extremely small partial R^2s had significant p values in the stepwise selection of Model II. For example, union status had a partial $R^2 = .0009$, given a sample size of 6652, the variable was still added at a significant $p = 0.0073$.

Data mining models usually respond to large samples positively due to their inductive learning nature. Abundant information in large samples can help to improve the accuracy of the rules (descriptions of data structure) summarized from the data. Also, more data are needed to validate the models and to avoid optimistic bias (overfit), as shown in the study when five different BBN models were created and cross-validated with the testing data set in order to find the optimal Model III.

Data mining algorithms use statistics but do not rely on significance tests. In the BBN learning, although χ^2 and the mutual information tests are used for measuring variable relationships, a system default or user-specified threshold value replaces the tests of significance to separate meaningful variable relations from those too weak to be useful. The use of a flexible threshold value provides the solution to the problem of sensitivity of statistical tests to minor differences when the sample size is large. As a matter of fact, the data mining BBN model works better with large samples than traditional statistics. A small sample size can lead to instable network structure because the strengths of major variable relationships could be strongly impaired by noisy data; identification of major patterns in large data sets reduces the chance of drawing conclusions from random structures. For example, when the entire data set of 9,963 records was divided into four random subsets, the BBN learning curve in each subset was

unstable. Also, although the resulting BBN models based on the four subsets had prediction accuracies around 11.57% (the accuracy rate of Model III on the testing data set) when applying to new data sets (see Table F5 in Appendix F), they all needed more variables in the networks than the original Model III to reach the same prediction accuracy.

Large Number of Variables

Data analysis methods that can effectively handle a large number of variables were the major concern of this study with 91 variables (one was salary, the predicted variable) kept in the analysis. The variables were on different measurement scales; some of them were highly correlated measures of a same underlying construct. Three variable reduction methods were tried and compared: theoretical selection, statistical extraction, and data mining BBN feature selection.

The results of three multiple regression models with the three groups of selected variables showed that the theoretical Model I had the smallest model R^2. However, when variables on administrative responsibility, teaching activity, publication, and experience were replaced by the comparable measures identified by statistical Model II and the combination Model IV, more variance of the basic faculty salary was accounted for. It was reasonable to suspect that the variables selected for Model I were not good enough, but the variables were chosen so that they best conveyed the major concepts presented in the compensation theory. Therefore, when a large number of variables is available, Model I exposed a problem of theory-driven variable selection: the theories provide general directions in choosing the major concepts; but they offer little help to pick a

single *best* measure when a number of different but highly correlated variables are available on the same construct.

It became clear in the process of this study that the ability to identify the most important variable from a group of highly correlated measures is an important criterion for evaluating applied data analysis methods when handling a large number of variables because redundant measures on the same constructs are common in large data sets and databases. The findings of this study indicate that BBN is able to perform such a function because Model III identified five variables from groups of measures on teaching, publication, experience, academic seniority, and institution parameter, the same five as those selected by Model II for reasons that the five variables accounted for more variances of the predicted variables than other alternatives on those five constructs.

Model II and Model IV are comparable in many ways. First, both are models resulting from data-driven procedures; second, theoretically, they both started the variable selection from the entire pool of 90 predictors; and third, they share the same group of major variables even though Model II had a much larger group. With the common ground they shared, the differences between the two models provide good insight to the differences between the data mining BBN and the traditional statistics in analyzing large-scale data.

Dimensional Simplification

Regarding the ability of identifying important predictor variables, the same group of major variables for salary prediction identified by the BBN and the statistical reduction indicates that they both can serve the purposes of dimensional simplification. However,

the statistical approach was quite laborious; it started with several multivariate analysis techniques to understand the data dimensions, and their results were compared manually to reach a final variable clustering schema. Because some of the variables in the same cluster had nonlinear relationships (e.g., similarities identified by KMC), extraction of a single measure from each cluster was also difficult. In this study, individual multiple regression procedures were conducted in each variable cluster. Overall, constant human intervention is required during the process.

Another problem with the statistical approach is the difficulty of handling categorical variables. Categorical variables are common in large data sets, but most multivariate techniques for structural simplification rely on linear correlation or geometric distance as measures of variable relationships that do not work with measures on nominal scales. Thus, categorical variables have to be handled separately when the data structure is examined, which makes the statistical data analysis even more complicated. For instance, Model II started with 37 variables, which took 75 degrees of freedom when all the multilevel categorical variables were recoded in to binary variables. Such a loss of degrees of freedom may not be a concern if the sample size is large, but data sets having a large number of variables but only few hundreds of cases do exist.

Finally, as for any data-driven process, the analysis may capitalize on chance and some of the variables in the final model may not make conceptual sense. A good example in Model II was the inclusion of respondent's satisfaction of work aspects other than salary (X01_66) and the freelance income from consulting (Q76G) as predictor variables for basic salary.

In comparison, the BBN model as a tool for variable selection and dimensional simplification is much easier. It is basically an automated process once the variables are ready in the software. BBN measures variable relationships as associations among categorical variables. By binning interval and ratio variables, BBN can accommodate all types of variables and is able to detect nonlinear relationships. When the number of variables is large, BBN reduces the computational task and the network complexity by randomly selecting and evaluating subsets from a large pool of variables, an approach proved to be effective by the analysis in this study.

The identified dependency relationships among variables are marked as directional connections. Some of them may be counterintuitive or even nonsensical because the learning of a BBN is completely data-driven when prior knowledge is excluded. However, the learning of a BBN allows previous domain knowledge to be specified and combined with information in the objective data set for model construction. For instance, it does not make conceptual sense in Model II that the free-lance income (Q76G) was used to predict the basic salary; such irregularities can be avoided in the BBN by specifying rules to prevent this type of conclusions. Also, BBN tries to look for dependencies and even causal relations within network structure; the entire network is modified every time when any one of the relationships is updated qualitatively or quantitatively. When used properly, the sound mathematic reasoning helps model validity and reduces the possibility of data-driven biases.

Unfortunately, simplicity is not always good; BBN has problems as well. First, the binning of continuous variables is an arbitrary process. If the bin width is set too

large, information loss in downgrading measurement scales can be severe, which may impact the model accuracy negatively. On the other hand, if the bin width is narrow, too many levels of variables increase their product state space and thus the computational complexity. In addition, the distinctions among adjacent intervals of the variable decrease with the bin width; the blurred contrasts can jeopardize the model prediction accuracy as well. For example, the BBN model had close to 40% correct classification rate when salary was divided into 14 bins, but the rate decreased to about 26% when there were 24 bins. Different binning schemas of the predictor variables may change the measured strength of their mutual relationships. The appropriate binning schema can only be learned through trial and error. Along with the disregarding of the ordering of the continuous data, the data-driven model learning is also extremely sample dependent and leading to problematic outcomes, such as the predicted probability in Table 4.11 in previous chapter: the probability of making a salary between ($50,035, $50,040) and between ($55,080, $58,525) was 7.28% and 6.72%, respectively, whereas the probability for ($55,040, $55,080) was only 0.81%.

The automated process of BBN learning also blinds researchers from having a detailed picture of variable relationships. In the statistical variable reduction, the clustering structure of variables was clear, and so were the variables that were similar or dissimilar to each other. For example, the 15 publication variables seemed to be measuring the same construct, but they had enough differences that separated them into 5 factors by EFA. This type of information is not available in data mining process. Users see the conceptually related 15 publication variables, and see only one of them (the career

publication in juried media) in the final model. The high automation is useful when the underlying variable relationships are not a concern, or when the number of variables is extremely large.

Prediction Accuracy

In the term of prediction model accuracy, Model IV started with only five variables that were selected by the BBN model from the pool of 90, and all five variables were significant and at the top of the 16-variable list produced by Model II. Model II started with all the variables and, after the laborious process of dimensional analysis and variable extraction, identified 16 variables that were significant in explaining the variance of the dependent variable. With fewer variables, Model IV had a R^2 of 0.4214 that was lower than that of Model II by .0822. However, a big difference in model degrees of freedom revealed that there was only about .0028 average increase in R^2 for each additional variable in Model II, and more predictor variables also worsened the multicollinearity index. The call for the best model comes down to an appraisal of a higher model R^2 or lower model complexity and multicollinearity.

In general, the analysis result of Model IV indicates that BBN should be valued for its ability to identify important variables related to a core construct and to select the most pertinent variable when a group of different but highly correlated variables are available on the same underlying concept. However, the BBN model had unstable classification accuracies across training and testing data sets. In spite of its weaknesses, as a different and robust approach, BBN should be used when appropriate to bring more insights into the data structure from different point of view.

Research Question II: Understand Data Structure

The second research question was to find out which approach was better adapted to uncover useful information and to understand data structure of large data sets. The three available choices were traditional statistics, data mining, or a combination of the two at different stages of the analysis.

With the traditional inferential approach, the pseudo research question about salary prediction was answered through selecting important variables from a large collection based on previous studies or educated guesses, and the research questions and model structures are predetermined; as Wegman pointed out in his paper (1988), this approach allows little room for exploring unknown patterns and gaining insight into the data structure and the functional relationships among the variables that were not considered. Another potential problem is that, if the questions to be answered with a large data set are ground breaking without theoretical backup from previous studies and literature, the traditional approach leaves researchers with the risk of leaving out important variables and selecting ones with strong bias.

In the absence of any domain knowledge, statistical multivariate techniques including EFA, PCA, KMC, and MDS can be used to understand data structure and simplify variable dimensions. However, each of the techniques has strong limitations given the complexity of large-scale data. For example, EFA and PCA require linear relationships among variables, which exclude nonlinear relationships and categorical variables out of the analysis. KMC does not assume linearity, but it cannot accommodate multilevel categorical variables either because it calculates geometric distances of

variables. Nonparametric MDS can work with variables on all measurement scales, but the interpretation of the result is difficult when the number of dimensions goes beyond six. For these techniques, manual interpretations of the output also substantially limit their usefulness when the number of variables becomes large.

BBN can select important variables without any previous domain knowledge. The five variables selected by the BBN model, when used in regression Model IV, fit the data almost as well as the theory-supported Model I. In addition, all five variables had comparable measures in the theoretical model; it is an indication that the BBN model may improve not only the statistical modeling, but also the theoretical understanding of faculty salary structure. However, for the purpose of understanding a construct and the underlying variable relationships, a parsimonious model with good prediction accuracy may not be enough because chances exist that important conceptual variables may be omitted and so the result may be biased or at least incomplete. For example, the BBN model did not include academic discipline as an important variable for salary prediction, which made the regression Model IV theoretically incomplete in comparison to Model I and II.

In contrast to the traditional confirmatory approach, BBN is open to any potential patterns in the data set without a predetermined model structure. The exploratory nature of BBN shows special values in studying data sets with a large number of variables and in detecting the hidden information in loosely structured large data sets. Exploratory statistics, when the number of variables is small, can serve the same purpose, but BBN extends the exploration to a high dimension. If any of the relationships uncovered by

BBN models show some theoretical value, designed studies can be conducted to follow them up.

Another advantage of BBN is that it can also accommodate prior knowledge in network discovery. For example, the BBN model was completely data driven in this study. Lack of domain knowledge led to some directional relationships that were counterintuitive, such as the percentage of time spent teaching undergrads leading to the institution type as shown in Figure 4.3. In such case, the invalid local structure can be avoided by inputting the known facts (e.g., highest degree partly determines number of publications, salary partly determines job satisfaction, gender cannot be determined by any other variables) into the software as background knowledge. By including the known information, the computational task can be reduced, and most importantly, the network validity can be greatly improved.

The flexibility of BBN in adapting domain knowledge as well as staying open to unknown patterns is the most attractive feature that makes the network structure popular. However, it does not mean that BBN models are always better than multiple regression. Just as different statistical techniques are appropriate for different data types and for answering different types of research questions, BBN is just another type of data analysis technique that has some advantages in working with large-scale data sets. If used out of proper context, a network can fail and lead to bad models. For instance, the classification accuracy of the BBN Model III was only 26%, a result of binning the 58 interval and ratio variables and of the narrow bin widths of the predicted variable (24 bins). Although the regression model using variables selected by the BBN suggested an encouraging

application of this new technique, comparing the classification accuracy of Model III in this study to the results published by the software authors on other data sets (Chen et al., 2001), it becomes clear that BBN works better with true categorical variables. When there are many continuous variables as in the current study, the results indicate that the combination approach, using a BBN to select variables and a multiple regression to perform prediction, is more efficient and parsimonious than any of the individual techniques alone.

Research Question III: Usefulness of Data Mining

The last question was about the special features of data mining, which are not readily available in statistical software, that can be beneficial to educational researchers in processing data of very large volume.

The Bayesian network is a good example of data mining techniques and their relationships with statistics and machine learning. It originated from Bayes's Theorem in classical probability theory, but the application of Bayesian probability remained difficult until modern computing technology made it possible to perform the enormous computation for learning a network from a collection of variables. The flexibility and mathematical rigidity of BBN in identifying variable relationships and making predictions attracted strong interest from the community of machine learning and data mining. Researchers developed algorithms to realize the complicated network structure in data analysis. In short, the Bayesian network is a data mining model strongly empowered by its statistical foundation.

Like many data mining techniques, an outstanding feature of BBN is their exploratory nature in studying a large amount of data and looking for hidden information. Designed to work with large-scale data, BBN is performed with the assumption that manual cleaning of irregularities is not feasible. Without making restrictions regarding the functional form of the variable relationships or the probabilistic distribution of the errors, data cleaning is only to get ready for the automated algorithms rather than confirming probabilistic distributions. Therefore, although it is known that data of large volume are often non-experimental and have many irregularities, BBN has no choice but to treat all variables equally as worthy candidates in the modeling process without checking for their reliability or validity. With relaxed model assumptions (e.g., no requirement for linearity and normality) and making prediction based on distributional mode rather than mean, the robustness of BBN against unreliable data quality (missing data and outliers) makes it a special technique for analyzing large scale data; on the other hand, it also determines that BBN is more appropriate for suggesting directions rather than confirming assumptions when the process is completely data-driven.

Moreover, large data sets may consist of diverse samples that are best understood by being divided into homogenous subgroups based on some important categorical variables. With algorithms adaptive to handling categorical variables, BBN can be used to identify the categorical variables that have the strongest differentiation power in defining the subgroups and to direct the research to a finer level. For example, the BBN Model III had five variables related to basic salary; they described the general structure of the entire data set. One of the variables, the Carnegie classification of institutions,

indicated that different types of postsecondary institutions had their unique salary structures below the general schema. The next step of the research could be to study these 15 subgroups of institutions separately and obtain a better understanding of each group.

BBN should not be completely ruled out as a choice for confirmatory analysis. BBN sometimes produces problematic findings because of the quality of the input data, not the procedure itself. No matter how perfect a modeling technique can be, when the input data have poor quality, it always follows the rule of GIGO (garbage in garbage out). BBN studies variable relationships with strict probabilistic inferences. It should be able to help to understand theoretical phenomena when the data contain qualified information.

The special combination of robustness and data-driven nature of most data mining techniques make them powerful in performing mechanical operations in data analysis, for example, making prediction without understanding a construct. Actually, a majority of data mining techniques are designed to fulfill such as operations including prediction and classification with no concerns about theoretical understanding of the conclusions. For instance, the type of cars owned can be used to predict whether a customer is going to respond to a cruise promotion. In such cases, model accuracy is the priority. However, the ability to include known domain knowledge in learning the network structure makes BBN quite special in prediction problems. Given a large group of input variables, inclusion of prior valid knowledge not only simplifies the computational task, but also helps to improve the prediction accuracy.

In short, data mining techniques, such as BBN and ANN, have unique advantages in analyzing large amounts of data. Data mining is not intended to replace statistics; it is a

collection of special tools that helps to probe through a data mountain, to look for unknown patterns and suggest future research directions, and to provide a more complete picture of data structure.

Limitations and Future Research Direction

The uniqueness of BBN in selecting critical variables and simplifying variable dimension was only demonstrated with prediction models. More studies are necessary for further understanding of Bayesian networks and other data mining techniques and their usefulness in educational research by applying them to more data sets and to answer different statistical questions.

Readers need to be aware that the intention of this study was to explore new data analysis methods for understanding large volumes of data in comparison with traditional statistical approaches. The use of faculty compensation data was only a choice of convenience rather than trying to make contributions to the compensation theory in postsecondary institutions. Any findings related to variable importance in faculty salary determination need further validation because the author does not have any expertise in this area. Also, the survey data were from a sample stratified by the Carnegie institution classification and by race and gender factors, the disproportionate selection and use of unequal clusters require the probability weight to be included to correct unequal representation of each observation in the sample for unbiased conclusions regarding faculty salary. In spite of the fact that this study took a model-based prediction approach in the analysis and based the conclusions on a sample of substantial size, risks of misspecified models and biased conclusions regarding the salary structure do exist.

Finally, with the recognition of BBN as a useful tool in examining variable relationships and detecting unknown patterns from a large volume of data, the binning of variables on interval and ratio scales is a major drawback in cases where many continuous variables are mixed with many categorical variables because information loss in the downgrade of measurement scales may lead to invalid learning results. This limitation may be overcome in the future. Some studies (e.g., Friedman, Goldszmidt, & Lee, 1998) are available discussing the possibility and technical details of including continuous predictor variables and their parametric distribution functions in estimating posterior probability in Bayesian networks. Once the study results can be realized in BBN software, the flexibility of Bayesian probability and the validity of network outputs can both benefit.

APPENDIX A

VARIABLE INFORMATION

Table A1

Name, Definition, and Measurement Scale of the 91 Variables from NSOPF:99

Variable name	Variable definition	Scale
Q1	Instructional duties	Categorical
Q10AREC	Years achieved tenure	Interval
Q10REC	Tenure status	Ordinal
Q12A	Appointments: Acting	Categorical
Q12E	Appointments: Clinical	Categorical
Q12F	Appointments: Research	Categorical
Q13	Chair of a department	Categorical
Q16A1REC	Highest degree type	Ordinal
Q16A2REC	2nd highest degree type	Ordinal
Q16B2REC	Years since 2nd highest degree	Interval
Q19	Current position as primary employment	Categorical
Q20	Outside consulting	Categorical
Q21	Other employment, fall 1998, non-consulting	Categorical
Q23	Positions in higher education during career	Interval
Q24A1REC	Years since 1st job in higher education	Interval
Q24A3	Employment status for 1st job in higher education	Categorical
Q24A5REC	Rank at hire for 1st job in higher education	Ordinal
Q25	Years teaching in higher education institution	Interval
Q26	Positions outside higher education during career	Interval

Variable name	Variable definition	Scale
Q29A1	Career creative works, juried media	Interval
Q29A2	Career creative works, non-juried media	Interval
Q29A3	Career reviews of books, creative works	Interval
Q29A4	Career books, textbooks, reports	Interval
Q29A5	Career presentations, performances	Interval
Q29B1	Recent sole creative works, juried media	Interval
Q29B2	Recent sole creative works, non-juried media	Interval
Q29B3	Recent sole reviews of books, creative works	Interval
Q29B4	Recent sole books, textbooks, reports	Interval
Q29B5	Recent sole presentations, performances	Interval
Q29C1	Recent joint creative works, juried media	Interval
Q29C2	Recent joint creative works, non-juried media	Interval
Q29C3	Recent joint reviews of books, creative works	Interval
Q29C4	Recent joint books, textbooks, reports	Interval
Q29C5	Recent joint presentations, performances	Interval
Q2REC	Teaching credit or noncredit courses	Ordinal
Q30B	Hours/week unpaid activities at the institution	Interval
Q30C	Hours/week paid activities not at the institution	Interval
Q30D	Hours/week unpaid activities not at the institution	Interval
Q31A1	Time actually spent teaching undergraduates (percentage)	Ratio
Q31A2	Time actually spent at teaching graduates (percentage)	Ratio

Variable name	Variable definition	Scale
Q31A3	Time actually spent at research (percentage)	Ratio
Q31A4	Time actually spent on professional growth (percentage)	Ratio
Q31A5	Time actually spent at administration (percentage)	Ratio
Q31A6	Time actually spent on service activity (percentage)	Ratio
Q31A7	Time actually spent on consulting (percentage)	Ratio
Q32A1	Number of undergraduate committees served on	Interval
Q32A2	Number of graduate committees served on	Interval
Q32B1	Number of undergraduate committees chaired	Interval
Q32B2	Number of graduate committees chaired	Interval
Q33	Total classes taught	Interval
Q40	Total credit classes taught	Interval
Q50	Total contact hours/week with students	Interval
Q51	Total office hours/week	Interval
Q52	Any creative work/writing/research	Categorical
Q54_55RE	PI / Co-PI on grants or contracts	Ordinal
Q58	Total number of grants or contracts	Interval
Q59A	Total funds from all sources	Ratio
Q61SREC	Work environment index	Ordinal
Q64	Union status	Categorical
Q76G	Consulting/freelance income	Ratio
Q7REC	Years on current job	Interval

Variable name	Variable definition	Scale
Q80	Number of dependents	Interval
Q81	Gender	Categorical
Q85	Disability	Categorical
Q87	Marital status	Categorical
Q90	Citizenship status	Categorical
Q9REC	Years on achieved rank	Interval
X01_3	Principal activity	Categorical
X01_60	Overall quality of research index	Ordinal
X01_66	Job satisfaction: other aspects of job	Ordinal
X01_82	Age	Interval
X01_8REC	Academic rank	Ordinal
X01_91RE	Highest educational level of parents	Ordinal
DISCIPLINE	Principal field of teaching/research	Categorical
X02_49	Individual instruction w/grad &1st professional students	Interval
X03_49	Number of students receiving individual instruction	Interval
X04_0	Carnegie classification of institution	Categorical
X04_41	Total classroom credit hours	Interval
X04_84	Ethnicity in single category	Categorical
X08_0D	Doctoral, 4-year, or 2-year institution	Ordinal
X08_0P	Private or public institution	Categorical
X09_0RE	Degree of urbanization of location city	Ordinal

Variable name	Variable definition	Scale
X09_76	Total income not from the institution	Ratio
X10_0	Ratio: FTE enrollment / FTE faculty	Ratio
X15_16	Years since highest degree	Interval
X21_0	Institution size: FTE graduate enrollment	Interval
X25_0	Institution size: Total FTE enrollment	Interval
X37_0	Bureau of Economic Analysis (BEA) region code	Categorical
X46_41	Undergraduate classroom credit hours	Interval
X47_41	Graduate and 1st professional classroom credit hours	Interval
SALARY	Basic academic year salary	Ratio

Note. All data were based on respondent' reported status during 1998-99 academic year.

Table A2

Frequency Table of Major Nominal Variables

Variable	Value label	Values	Frequency	Percent	Valid percent
Chair of a department	No	0	5703	85.7	85.7
(Q13)	Yes	1	949	14.3	14.3
		Total	6652	100.0	100.0
Gender	Male	1	3927	59.0	59.0
(Q81)	Female	2	2725	41.0	41.0
		Total	6652	100.0	100.0
Principal field of teaching/research	Legitimate skip	0	592	8.9	8.9
(DISCIPLINE)	Agriculture & home economics	1	139	2.1	2.1
	Business	2	383	5.8	5.8
	Education	3	445	6.7	6.7
	Engineering	4	302	4.5	4.5
	Fine arts	5	305	4.6	4.6
	Health sciences	6	920	13.8	13.8
	Humanities	7	864	13.0	13.0
	Natural sciences	8	1312	19.7	19.7
	Social sciences	9	684	10.3	10.3
	All other programs	10	706	10.6	10.6
		Total	6652	100.0	100.0
Carnegie classification of institutions	Private other PhD	1	208	3.1	3.1
	Public comprehensive	2	1123	16.9	16.9
(X04_0)	Private comprehensive	3	415	6.2	6.2

Variable	Value label	Values	Frequency	Percent	Valid Percent
	Public liberal arts	4	141	2.1	2.1
	Private liberal arts	5	485	7.3	7.3
	Public medical	6	175	2.6	2.6
	Private Medical	7	64	1.0	1.0
	Private religious	8	43	0.6	0.6
	Public 2-year	9	1129	17.0	17.0
	Private 2-year	10	34	0.5	0.5
	Public other	11	36	0.5	0.5
	Private other	12	69	1.0	1.0
	Public research	13	1731	26.0	26.0
	Private research	14	465	7.0	7.0
	Public other Ph.D.	15	534	8.0	8.0
		Total	6652	100.0	100.0
BEA region code	U.S. service school	0	5	0.1	0.1
(X37_0)	New England	1	370	5.6	5.6
	Mid East	2	1024	15.4	15.4
	Great Lakes	3	1190	17.9	17.9
	Plains	4	628	9.4	9.4
	Southeast	5	1737	26.1	26.1
	Southwest	6	611	9.2	9.2
	Rocky Mountain	7	306	4.6	4.6
	Far West	8	776	11.7	11.7
		Total	6647	99.9	100.0
	Missing	System	5	0.1	

Table A3

Descriptive Information of Variables on Interval/Ratio Scales

Variable	Minimum	Maximum	Mean	Std. deviation
LNSALR	8.6482	12.4292	10.9002	0.4308
SALARY	5700	250000	59867.97	30914.08
Q10AREC	0	40	6.69	8.91
Q16B2REC	0	51	16.65	13.09
Q23	1	17	2.16	1.72
Q24A1REC	1	46	17.18	10.51
Q25	0	50	16.06	10.28
Q26	0	30	1.41	2.10
Q29A1	0	200	15.85	30.30
Q29A2	0	175	7.57	19.96
Q29A3	0	75	4.12	10.02
Q29A4	0	64	2.62	7.53
Q29A5	0	750	38.54	82.68
Q29B1	0	40	1.63	4.53
Q29B2	0	50	1.61	4.96
Q29B3	0	25	0.98	2.95
Q29B4	0	22	0.56	2.13
Q29B5	0	160	7.70	19.17
Q29C1	0	41	2.32	5.51
Q29C2	0	25	0.71	2.65
Q29C3	0	10	0.34	1.19

Variable	Minimum	Maximum	Mean	Std. deviation
Q29C4	0	15	0.44	1.62
Q29C5	0	85	3.45	9.30
Q30B	0	40	2.58	5.28
Q30C	0	47	1.96	4.53
Q30D	0	37	1.57	3.15
Q31A1	0	100	41.80	32.98
Q31A2	0	100	12.24	18.74
Q31A3	0	100	16.09	20.20
Q31A4	0	100	4.32	6.70
Q31A5	0	100	15.46	20.91
Q31A6	0	100	6.99	14.19
Q31A7	0	100	3.12	8.75
Q32A1	0	10	0.57	1.49
Q32A2	0	20	1.90	3.52
Q32B1	0	10	0.26	0.97
Q32B2	0	20	0.73	1.84
Q33	0	20	3.17	3.02
Q40	0	20	2.88	2.68
Q50	0	40	2.91	5.20
Q51	0	40	6.84	8.23
Q58	0	10	0.81	1.58
Q59A	0	2000000	58921.19	200003.95
Q61SREC	6	24	16.30	3.52
Q76G	0	75000	1581.66	6104.72

Variable	Minimum	Maximum	Mean	Std. deviation
Q7REC	1	40	12.11	9.69
Q80	0	10	1.32	1.37
Q9REC	0	40	7.34	6.98
X01_82	24	82	49.06	9.49
X02_49	0	85	3.35	8.85
X03_49	0	185	11.20	22.27
X04_41	0	200	8.77	10.56
X09_76	0	208600	6322.71	15862.82
X10_0	0.4	83.8	14.14	6.01
X15_16	0	47	16.80	10.11
X21_0	0	10298.2	1890.38	2274.95
X25_0	18	45263.4	11481.93	10358.34
X46_41	0	200	7.05	10.19
X47_41	0	133	1.71	4.83

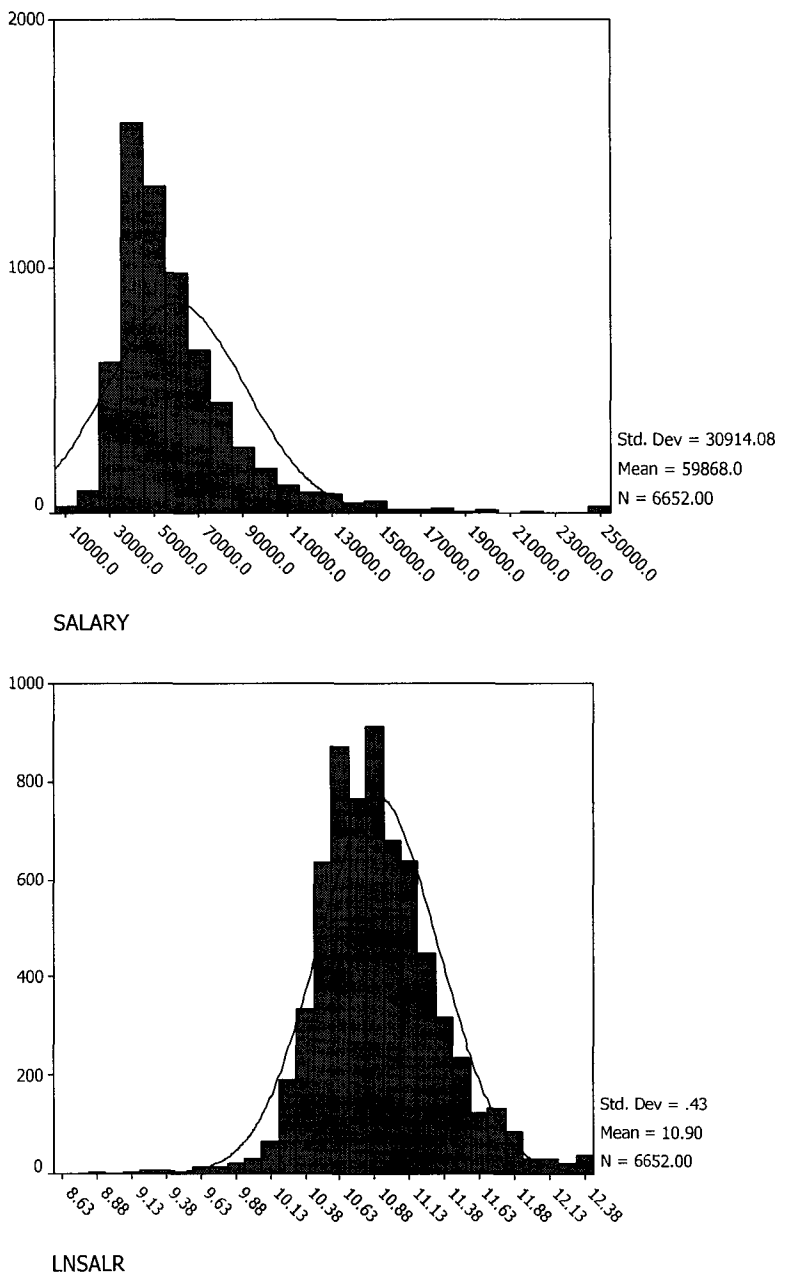


Figure A1. Distribution of SALARY and log-transformed SALARY (LOGSAL) in the training data set.

APPENDIX B

BUILDING THE THEORETICAL MODEL: REGRESSION MODEL I

Table B1

Summary of the Stepwise Selection of Regression Model I: Binary Variables from the Same Categorical Variable Evaluated as a Group

Step	Variables entered	Label	*df*	Partial R^2	Model R^2	*F*	$p > F$
1	X01_8REC	Academic rank	1	0.1679	0.1679	1340.27	<.0001
2	X08_0D	Inst type: doctoral, 4-yr, or 2-yr	2	0.0754	0.2434	661.91	<.0001
3	DISCIPLINE	Principal field of teaching/research	12	0.0505	0.2939	47.44	<.0001
4	Q24A1REC	Years working in higher ed	13	0.0440	0.3379	440.13	<.0001
5	Q16A1REC	Highest degree type	14	0.0288	0.3667	301.89	<.0001
6	TOTPUB	Total number of publications	15	0.0201	0.3869	217.72	<.0001
7	Q81	Gender	16	0.0111	0.3980	122.25	<.0001
8	Q59A	Total funds from all sources	17	0.0091	0.4071	102.16	<.0001
9	BEA	BEA region codes	24	0.0106	0.4177	17.15	<.0001
10	Q10REC	Tenure status	25	0.0026	0.4203	29.76	<.0001
11	Q13	Department chair	26	0.0010	0.4213	10.87	0.001

Note. A total of 35 independent variables. The dependent variable was log-transformed SALARY.

APPENDIX C

BUILDING THE STATISTICAL MODEL: REGRESSION MODEL II

Table C1

Model II Variable Reduction: Factors after VARIMAX Rotation (82 Variables)

Factor	Variable name	Definition	Factor loading
1	Q7REC	Years on current job	0.901
Experience	Q25	Years teaching in higher education institution	0.899
	Q24A1REC	Years since 1st job in higher education	0.895
	Q10AREC	Years achieved tenure	0.861
	X15_16	Years since highest degree	0.839
	Q9REC	Years on achieved rank	0.819
	X01_82	Age	0.810
	Q16B2REC	Years since 2nd highest degree	0.473
	Q10REC	Tenure status	0.431
	X01_8REC	Academic rank	0.426
2	X21_0	Institution size: FTE graduate enrollment	0.859
Institution	X25_0	Institution size: Total FTE enrollment	0.836
parameter	X08_0D	Doctoral, 4-year, or 2-year institution	0.659
	X09_0RE	Degree of urbanization of location city	0.525
	Q31A1	Time actually spent teaching undergraduates (percentage)	-0.417
3	Q58	Total number of grants or contracts	0.832
Research	Q54_55RE	PI / Co-PI on grants or contracts	0.782
activity	Q59A	Total funds from all sources	0.702
	Q31A3	Time actually spent at research (percentage)	0.490
	Q52	Any creative work/writing/research	0.449
	Q29A1	Career creative works, juried media	0.396
4	Q29B5	Recent sole presentations, performances	0.775
Publication	Q29A5	Career presentations, performances	0.683
(presentations, etc.)	Q29C5	Recent joint presentations, performances	0.610
	Q29B1	Recent sole creative works, juried media	0.494
5	Q29A4	Career books, textbooks, reports	0.864
Publication	Q29B4	Recent sole books, textbooks, reports	0.791
(books)	Q29C4	Recent joint books, textbooks, reports	0.771

Factor	Variable name	Definition	Factor loading
6	Q32B2	Number of graduate committees chaired	0.900
Teaching	Q32A2	Number of graduate committees served on	0.876
(graduate)	Q31A2	Time actually spent at teaching graduates (percentage)	0.421
7	X46_41	Undergraduate classroom credit hours	0.900
Teaching (credits)	X04_41	Total classroom credit hours	0.864
8	Q1	Instructional duties	0.891
Teaching (general)	Q2REC	Teaching credit or noncredit courses	0.868
9	Q33	Total classes taught	0.837
Teaching	Q40	Total credit classes taught	0.828
(classes)	X47_41	Graduate and 1st professional classroom credit hours	0.529
	X04_41	Total classroom credit hours	0.369
10	Q29A2	Career creative works, non-juried media	0.820
Publication	Q29B2	Recent sole creative works, non-juried media	0.740
(non-juried)	Q29C2	Recent joint creative works, non-juried media	0.702
11	Q29B3	Recent sole reviews of books, creative works	0.810
Publication	Q29A3	Career reviews of books, creative works	0.808
(reviews, etc.)	Q29C3	Recent joint reviews of books, creative works	0.444
12	X01_8REC	Academic rank	0.533
Academic	Q16A1REC	Highest degree type	0.495
seniority	Q10REC	Tenure status	0.492
	Q52	Any creative work/writing/research	0.415
	Q24A5REC	Rank at hire for 1st job in higher education	0.380
13	Q32B1	Number of undergraduate committees chaired	0.936
Teaching (undergrad)	Q32A1	Number of undergraduate committees served on	0.934
14	Q16A2REC	2nd highest degree type	0.884
2nd highest degree	Q16B2REC	Years since 2nd highest degree	0.749
15	Q76G	Consulting/freelance income	0.744
Outside income	X09_76	Total income not from the institution	0.741
	Q30C	Hours/week paid activities not at the institution	0.383
16	Q30C	Hours/week paid activities not at the institution	0.598
Other employment	Q20	Outside consulting	0.544
	Q31A7	Time actually spent on consulting (percentage)	0.522

Factor	Variable name	Definition	Factor loading
	Q21	Other employment, fall 1998, non-consulting	0.472
	Q30D	Hours/week unpaid activities not at the institution	0.422
17	X08_0P	Private or public institution	0.747
Institution parameter	X10_0	Ratio: FTE enrollment / FTE faculty	-0.613
18	Q31A5	Time actually spent at administration (percentage)	0.765
Administration	Q13	Chair of a department	0.689
19	Q31A6	Time actually spent on service activity (percentage)	0.761
(Not interpretable)	Q12E	Appointments: Clinical	-0.641
20	Q24A3	Employment status for 1st job in higher education	0.681
Beginning status	Q24A5REC	Rank at hire for 1st job in higher education	0.565
	Q23	Positions in higher education during career	-0.692
21	X03_49	Number of students receiving individual instruction	0.833
Teaching (indv.)	X02_49	Individual instruction w/grad &1st professional students	0.730
22	Q29C1	Recent joint creative works, juried media	0.631
Publication	Q29C3	Recent joint reviews of books, creative works	0.549
(joint authorship)	Q29C5	Recent joint presentations, performances	0.374
23	X01_66	Job satisfaction: other aspects of job	0.746
Job satisfaction	X01_60	Overall quality of research index	0.730
	Q61SREC	Work environment index	0.387
24	Q80	Number of dependents	0.700
(Not interpretable)	Q81	Gender	-0.680
25	Q26	Positions outside higher education during career	0.535
(Not interpretable)	Q85	Disability	0.475
	X01_91RE	Highest educational level of parents	-0.454
26	Q51	Total office hours/week	0.614
Teaching	Q50	Total contact hours/week with students	0.548
(contact hours)	Q12A	Appointments: Acting	-0.500

Note: Only variables with a loading of 0.35 or higher are listed.

Table C2

Model II Variable Reduction: Factors after OBLIMIN Rotation (82 Variables)

Factor	Variable name	Definition	Factor loading
1	Q58	Total number of grants or contracts	0.851
Research activity	Q54_55RE	PI / Co-PI on grants or contracts	0.804
	Q59A	Total funds from all sources	0.720
	Q31A3	Time actually spent at research (percentage)	0.531
	Q29A1	Career creative works, juried media	0.487
	Q52	Any creative work/writing/research	0.475
	Q29C1	Recent joint creative works, juried media	0.424
2	Q25	Years teaching in higher education institution	0.907
Experience	Q7REC	Years on current job	0.902
	Q24A1REC	Years since 1st job in higher education	0.899
	Q10AREC	Years achieved tenure	0.870
	X15_16	Years since highest degree	0.842
	Q9REC	Years on achieved rank	0.817
	X01_82	Age	0.816
	Q16B2REC	Years since 2nd highest degree	0.488
	Q10REC	Tenure status	0.459
	X01_8REC	Academic rank	0.459
3	Q33	Total classes taught	0.854
Teaching (classes)	Q40	Total credit classes taught	0.851
	X47_41	Graduate and 1st professional classroom credit hours	0.530
	X04_41	Total classroom credit hours	0.427
4	Q31A1	Time actually spent teaching undergraduates (percentage)	-0.353
Teaching (general)	Q1	Instructional duties	-0.888
	Q2REC	Teaching credit or noncredit courses	-0.890
5	X08_0P	Private or public institution	0.728
Institution parameter	X10_0	Ratio: FTE enrollment / FTE faculty	-0.648

Factor	Variable name	Definition	Factor loading
6	Q16A2REC	2nd highest degree type	0.902
Highest degree	Q16B2REC	Years since 2nd highest degree	0.782
	Q16A1REC	Highest degree type	0.370
7	Q31A7	Time actually spent on consulting (percentage)	-0.400
Other employment	Q20	Outside consulting	-0.612
	Q30C	Hours/week paid activities not at the institution	-0.664
	X09_76	Total income not from the institution	-0.685
	Q76G	Consulting/freelance income	-0.735
8	Q32B1	Number of undergraduate committees chaired	-0.931
Teaching (undergrad)	Q32A1	Number of undergraduate committees served on	-0.935
9	Q31A1	Time actually spent teaching undergraduates (percentage)	0.498
Institution parameter	Q31A3	Time actually spent at research (percentage)	-0.397
	X09_0RE	Degree of urbanization of location city	-0.545
	X08_0D	Doctoral, 4-year, or 2-year institution	-0.729
	X25_0	Institution size: Total FTE enrollment	-0.847
	X21_0	Institution size: FTE graduate enrollment	-0.890
10	X03_49	Number of students receiving individual instruction	0.832
Teaching (indv)	X02_49	Individual instruction w/grad &1st professional students	0.747
11	Q29C4	Recent joint books, textbooks, reports	-0.794
Publication (books)	Q29B4	Recent sole books, textbooks, reports	-0.801
	Q29A4	Career books, textbooks, reports	-0.884
12	Q29B5	Recent sole presentations, performances	0.787
Publication	Q29A5	Career presentations, performances	0.704
(presentations, etc.)	Q29C5	Recent joint presentations, performances	0.663
	Q29B1	Recent sole creative works, juried media	0.558
	Q29C1	Recent joint creative works, juried media	0.476
	Q29A1	Career creative works, juried media	0.416
13	Q31A5	Time actually spent at administration (percentage)	0.761
Administration	Q13	Chair of a department	0.692
	Q31A3	Time actually spent at research (percentage)	-0.368

Factor	Variable name	Definition	Factor loading
14	Q29A3	Career reviews of books, creative works	0.845
Publication	Q29B3	Recent sole reviews of books, creative works	0.828
(reviews)	Q29C3	Recent joint reviews of books, creative works	0.495
	Q29B1	Recent sole creative works, juried media	0.396
	Q29A1	Career creative works, juried media	0.374
15	Q31A1	Time actually spent teaching undergraduates (percentage)	0.389
Teaching (graduate)	X08_0D	Doctoral, 4-year, or 2-year institution	-0.358
	Q31A2	Time actually spent at teaching graduates (percentage)	-0.521
	Q32B2	Number of graduate committees chaired	-0.903
	Q32A2	Number of graduate committees served on	-0.915
16	X01_66	Job satisfaction: other aspects of job	0.754
Job satisfaction	X01_60	Overall quality of research index	0.734
	Q61SREC	Work environment index	0.372
17	Q30D	Hours/week unpaid activities not at the institution	0.534
Professional growth	Q31A4	Time actually spent on professional growth (percentage)	0.403
18	Q23	Positions in higher education during career	0.672
Beginning status	Q24A5REC	Rank at hire for 1st job in higher education	-0.614
	Q24A3	Employment status for 1st job in higher education	-0.675
19	Q29C2	Recent joint creative works, non-juried media	-0.734
Publication	Q29B2	Recent sole creative works, non-juried media	-0.769
(non-juried media)	Q29A2	Career creative works, non-juried media	-0.849
20	Q31A6	Time actually spent on service activity (percentage)	0.767
(Not interpretable)	Q12E	Appointments: Clinical	-0.654
21	Q81	Gender	0.705
(Not interpretable)	Q80	Number of dependents	-0.674
22	X46_41	Undergraduate classroom credit hours	0.947
Teaching load	X04_41	Total classroom credit hours	0.883
	Q31A1	Time actually spent teaching undergraduates (percentage)	0.517
	Q40	Total credit classes taught	0.394
	Q33	Total classes taught	0.356

Factor	Variable name	Definition	Factor loading
	X08_0D	Doctoral, 4-year, or 2-year institution	-0.403
23	X01_91RE	Highest educational level of parents	0.517
(Not interpretable)	Q10REC	Tenure status	0.401
	Q16A1REC	Highest degree type	0.359
	Q26	Positions outside higher education during career	-0.541
24	Q12A	Appointments: Acting	0.367
Teaching	Q50	Total contact hours/week with students	-0.567
(contact hours)	Q51	Total office hours/week	-0.626
25	Q29C3	Recent joint reviews of books, creative works	0.518
Publication	Q29C1	Recent joint creative works, juried media	0.496
26	Q85	Disability	0.537
(Not interpretable)	X47_41	Graduate and 1st professional classroom credit hours	0.427
	Q31A2	Time actually spent at teaching graduates (percentage)	0.407

Note. Only variables with a loading of 0.35 or higher are listed.

Table C3

Model II Variable Reduction: Factors after VARIMAX Rotation (70 Variables)

Factor	Variable name	Definition	Factor loading
1	Q25	Years teaching in higher education institution	0.901
Experience	Q7REC	Years on current job	0.901
	Q24A1REC	Years since 1st job in higher education	0.898
	Q10AREC	Years achieved tenure	0.860
	X15_16	Years since highest degree	0.845
	Q9REC	Years on achieved rank	0.819
	X01_82	Age	0.815
	Q16B2REC	Years since 2nd highest degree	0.472
	Q10REC	Tenure status	0.430
	X01_8REC	Academic rank	0.426
2	X21_0	Institution size: FTE graduate enrollment	0.868
Institution	X25_0	Institution size: Total FTE enrollment	0.835
parameter	X08_0D	Doctoral, 4-year, or 2-year institution	0.673
	X09_0RE	Degree of urbanization of location city	0.530
	Q31A1	Time actually spent teaching undergraduates (percentage)	-0.444
3	Q58	Total number of grants or contracts	0.843
Research activities	Q54_55RE	PI / Co-PI on grants or contracts	0.788
	Q59A	Total funds from all sources	0.718
	Q31A3	Time actually spent at research (percentage)	0.477
	Q52	Any creative work/writing/research	0.465
	Q29A1	Career creative works, juried media	0.375
4	Q29A4	Career books, textbooks, reports	0.865
Publication (books)	Q29B4	Recent sole books, textbooks, reports	0.795
	Q29C4	Recent joint books, textbooks, reports	0.768
5	Q32B2	Number of graduate committees chaired	0.896
Teaching (graduate)	Q32A2	Number of graduate committees served on	0.874
	Q31A2	Time actually spent at teaching graduates (percentage)	0.444

Factor	Variable name	Definition	Factor loading
6	Q29B5	Recent sole presentations, performances	0.787
Publication	Q29A5	Career presentations, performances	0.697
(performance etc.)	Q29C5	Recent joint presentations, performances	0.589
	Q29B1	Recent sole creative works, juried media	0.467
7	Q33	Total classes taught	0.844
Teaching load	Q40	Total credit classes taught	0.836
	X47_41	Graduate and 1st professional classroom credit hours	0.527
	X04_41	Total classroom credit hours	0.359
8	X46_41	Undergraduate classroom credit hours	0.906
Teaching (credit hrs)	X04_41	Total classroom credit hours	0.888
9	Q1	Instructional duties	0.903
Teaching (general)	Q2REC	Teaching credit or noncredit courses	0.873
10	X01_8REC	Academic rank	0.583
Academic seniority	Q16A1REC	Highest degree type	0.538
	Q10REC	Tenure status	0.530
	Q52	Any creative work/writing/research	0.449
	Q24A5REC	Rank at hire for 1st job in higher education	0.387
11	X09_76	Total income not from the institution	0.776
Outside income	Q76G	Consulting/freelance income	0.769
	Q30C	Hours/week paid activities not at the institution	0.502
	Q20	Outside consulting	0.470
12	Q29A2	Career creative works, non-juried media	0.826
Publication	Q29B2	Recent sole creative works, non-juried media	0.728
(non-juried media)	Q29C2	Recent joint creative works, non-juried media	0.703
13	Q32B1	Number of undergraduate committees chaired	0.934
Teaching (undergrad)	Q32A1	Number of undergraduate committees served on	0.933
14	Q29B3	Recent sole reviews of books, creative works	0.815
Publication (reviews)	Q29A3	Career reviews of books, creative works	0.804
	Q29C3	Recent joint reviews of books, creative works	0.403

Factor	Variable name	Definition	Factor loading
15	Q29C1	Recent joint creative works, juried media	0.714
Publication	Q29C3	Recent joint reviews of books, creative works	0.623
(juried media)	Q29A1	Career creative works, juried media	0.418
	Q29C5	Recent joint presentations, performances	0.406
	Q29C2	Recent joint creative works, non-juried media	0.402
16	Q16A2REC	2nd highest degree type	0.913
2nd highest degree	Q16B2REC	Years since 2nd highest degree	0.784
17	X08_0P	Private or public institution	0.746
Institution parameter	X25_0	Institution size: Total FTE enrollment	-0.359
	X10_0	Ratio: FTE enrollment / FTE faculty	-0.633
18	Q31A5	Time actually spent at administration (percentage)	0.770
Admin responsibility	Q13	Chair of a department	0.748
19	Q24A3	Employment status for 1st job in higher education	0.680
Beginning status	Q24A5REC	Rank at hire for 1st job in higher education	0.562
	Q23	Positions in higher education during career	-0.709
20	X03_49	Number of students receiving individual instruction	0.824
Teaching (indv instr)	X02_49	Individual instruction w/grad &1st professional students	0.771
21	X01_66	Job satisfaction: other aspects of job	0.772
Job satisfaction	X01_60	Overall quality of research index	0.759
	Q61SREC	Work support availability	0.399
22	Q31A7	Time actually spent on consulting (percentage)	0.618
Other employment	Q30C	Hours/week paid activities not at the institution	0.508
	Q21	Other employment, fall 1998, non-consulting	0.439
	Q20	Outside consulting	0.416
	Q12A	Appointments: Acting	0.385
23	Q51	Total office hours/week	0.676
Teaching (contact)	Q50	Total contact hours/week with students	0.665

Note. Only variables with a loading of 0.35 or higher are listed.

Table C4

Model II Variable Reduction: Variable Clusters by KMC Analysis

Cluster ID	Variable name	Variable label	Distance to centroid
1	Q40	Total credit classes taught	53.73
	X04_41	Total classroom credit hours	55.71
	X46_41	Undergraduate classroom credit hours	55.83
	Q33	Total classes taught	57.09
	Q2REC	Teaching credit or noncredit courses	61.93
	Q31A1	Time actually spent teaching undergraduates (percentage)	63.16
	Q1	Instructional duties	68.88
2	Q10REC	Tenure status	69.73
	Q12E	Appointments: Clinical	70.90
	Q16B2REC	Years since 2nd highest degree	72.63
	X10_0	Ratio: FTE enrollment / FTE faculty	72.82
	Q12F	Appointments: Research	74.98
	Q19	Current position as primary employment	75.46
	Q85	Disability	76.27
3	Q13	Chair of a department	61.76
	Q31A5	Time actually spent at administration (percentage)	62.53
	Q61SREC	Work environment index	67.48
	X08_0P	Private or public institution	71.03
4	Q25	Years teaching in higher education institution	40.35
	Q24A1REC	Years since 1st job in higher education	40.72
	Q7REC	Years on current job	41.32
	Q10AREC	Years achieved tenure	45.16
	X15_16	Years since highest degree	46.52
	X01_82	Age	49.02
	Q9REC	Years on achieved rank	49.86
	X01_66	Job satisfaction: other aspects of job	82.08
	X01_60	Overall quality of research index	87.33

Cluster ID	Variable name	Variable label	Distance to cluster seed
5	Q21	Other employment, fall 1998, non-consulting	63.66
	Q23	Positions in higher education during career	68.23
	Q30C	Hours/week paid activities not at the institution	68.64
	Q16A2REC	2nd highest degree type	70.27
6	Q24A3	Employment status for 1st job in higher education	51.16
	Q24A5REC	Rank at hire for 1st job in higher education	51.16
7	X08_0D	Doctoral, 4-year, or 2-year institution	59.70
	X21_0	Institution size: FTE graduate enrollment	61.08
	X25_0	Institution size: Total FTE enrollment	64.61
	Q32A2	Number of graduate committees served on	65.74
	Q31A2	Time actually spent at teaching graduates (percentage)	65.79
	Q16A1REC	Highest degree type	68.54
	Q32B2	Number of graduate committees chaired	69.43
	X02_49	Individual instruction w/grad &1st professional students	72.53
	X47_41	Graduate and 1st professional classroom credit hours	73.51
	X01_8REC	Academic rank	74.01
	X09_0RE	Degree of urbanization of location city	74.13
	X01_91RE	Highest educational level of parents	81.58
	Q12A	Appointments: Acting	84.21
8	Q26	Positions outside higher education during career	53.79
	Q81	Gender	53.79
9	Q30B	Hours/week unpaid activities at the institution	62.68
	Q30D	Hours/week unpaid activities not at the institution	64.00
	Q31A4	Time actually spent on professional growth (percentage)	64.63
10	Q50	Total contact hours/week with students	60.66
	Q51	Total office hours/week	63.65
	X03_49	Number of students receiving individual instruction	64.27
11	Q32A1	Number of undergraduate committees served on	26.66
	Q32B1	Number of undergraduate committees chaired	26.66
12	Q31A6	Time actually spent on service activity (percentage)	55.36
	Q80	Number of dependents	55.36

Cluster ID	Variable name	Variable label	Distance to cluster seed
13	Q20	Outside consulting	55.65
	Q76G	Consulting/freelance income	56.73
	Q31A7	Time actually spent on consulting (percentage)	62.68
14	Q29B4	Recent sole books, textbooks, reports	62.90
	Q29A4	Career books, textbooks, reports	63.09
	Q29C4	Recent joint books, textbooks, reports	64.48
	Q29B5	Recent sole presentations, performances	66.22
	Q29B1	Recent sole creative works, juried media	67.75
	Q29B2	Recent sole creative works, non-juried media	68.04
	Q29B3	Recent sole reviews of books, creative works	70.06
	X09_76	Total income not from the institution	77.39
15	Q29A1	Career creative works, juried media	60.23
	Q54_55RE	PI / Co-PI on grants or contracts	60.49
	Q58	Total number of grants or contracts	60.55
	Q29C1	Recent joint creative works, juried media	62.78
	Q59A	Total funds from all sources	68.37
	Q31A3	Time actually spent at research (percentage)	68.38
	Q52	Any creative work/writing/research	69.42
	Q29A5	Career presentations, performances	69.96
	Q29C3	Recent joint reviews of books, creative works	71.07
	Q29C2	Recent joint creative works, non-juried media	71.32
	Q29C5	Recent joint presentations, performances	71.41
	Q29A2	Career creative works, non-juried media	71.94
	Q29A3	Career reviews of books, creative works	73.69

Table C5

Model II Variable Reduction: Final Variable Clusters

Cluster definition	Variable	Definition
1. Research	Q58	Total number of grants or contracts
	Q54_55RE	PI / Co-PI on grants or contracts
	Q59A	Total funds from all sources
	Q31A3	Time actually spent at research (percentage)
	Q29A1	Career creative works, juried media
	Q52	Any creative work/writing/research
	Q29C1	Recent joint creative works, juried media
2. Experience	Q25	Years teaching in higher education institution
	Q7REC	Years on current job
	Q24A1REC	Years since 1st job in higher education
	Q10AREC	Years achieved tenure
	X15_16	Years since highest degree
	Q9REC	Years on achieved rank
	X01_82	Age
3. Teaching	Q33	Total classes taught
	Q40	Total credit classes taught
	X47_41	Graduate and 1st professional classroom credit hours
	X04_41	Total classroom credit hours
	Q31A1	Time actually spent teaching undergrads (percentage)
	Q1	Instructional duties
	Q2REC	Teaching credit or noncredit courses
	X46_41	Undergraduate classroom credit hours

Cluster definition	Variable	Definition
4. Teaching: Graduate	Q32B2	Number of graduate committees chaired
	Q32A2	Number of graduate committees served on
	Q31A2	Time actually spent at teaching graduates (percentage)
5. Teaching: Individual	X03_49	Number of students receiving individual instruction
Instruction	X02_49	Individual instruction w/grad &1st prof students
	Q50	Total contact hours/week with students
	Q51	Total office hours/week
6. Teaching: Undergrad	Q32B1	Number of undergraduate committees chaired
Committee	Q32A1	Number of undergraduate committees served on
7. Administrative responsibility	Q31A5	Time actually spent at administration (percentage)
	Q13	Chair of a department
8. Highest degree	Q16A2REC	2nd highest degree type
	Q16B2REC	Years since 2nd highest degree
	Q16A1REC	Highest degree type
9. Beginning status	Q23 (-)	Positions in higher education during career
	Q24A5REC	Rank at hire for 1st job in higher education
	Q24A3	Employment status for 1st job in higher education
10. Publication: Reviews	Q29A3	Career reviews of books, creative works
	Q29B3	Recent sole reviews of books, creative works
	Q29C3	Recent joint reviews of books, creative works
11. Publication: Creative works	Q29C2	Recent joint creative works, non-juried media
	Q29B2	Recent sole creative works, non-juried media
	Q29A2	Career creative works, non-juried media
	Q29B5	Recent sole presentations, performances

Cluster definition	Variable	Definition
	Q29A5	Career presentations, performances
	Q29C5	Recent joint presentations, performances
	Q29B1	Recent sole creative works, juried media
12. Publication: Books	Q29C4	Recent joint books, textbooks, reports
	Q29B4	Recent sole books, textbooks, reports
	Q29A4	Career books, textbooks, reports
13. Institution parameter	X09_0RE	Degree of urbanization of location city
	X08_0D	Doctoral, 4-year, or 2-year institution
	X25_0	Institution size: Total FTE enrollment
	X21_0	Institution size: FTE graduate enrollment
14. Institution parameter: Other	X08_0P (-)	Private or public institution
	X10_0	Ratio: FTE enrollment / FTE faculty
15. Other employment/income	Q31A7	Time actually spent on consulting (percentage)
	Q20	Outside consulting
	Q30C	Hours/week paid activities not at the institution
	X09_76	Total income not from the institution
	Q76G	Consulting/freelance income
	Q21	Other employment, fall 1998, non-consulting
	Q30D	Hours/week unpaid activities not at the institution
16. Work environment	X01_66	Job satisfaction: other aspects of job
	X01_60	Overall quality of research index
	Q61SREC	Work environment index
17. Academic rank	Q10REC	Tenure status
	X01_8REC	Academic rank

Table C6

Candidate Independent Variables of Model II

Variable name	Variable definition	*df*
	Variables from the clusters	
Q29A1	Career creative works, juried media	1
X15_16	Years since highest degree	1
Q31A1	Time actually spent teaching undergraduates (percentage)	1
Q31A2	Time actually spent at teaching graduates (percentage)	1
X02_49	Individual instruction w/grad &1st professional students	1
Q32B1	Number of undergraduate committees chaired	1
Q31A5	Time actually spent at administration (percentage)	1
Q16A1REC	Highest degree type	1
Q24A5REC	Rank at hire for 1st job in higher education	1
Q29A3	Career reviews of books, creative works	1
Q29A5	Career presentations, performances	1
X08_0D	Doctoral, 4-year, or 2-year institution	1
Q29A4	Career books, textbooks, reports	1
X10_0	Ratio: FTE enrollment / FTE faculty	1
Q76G	Consulting/freelance income	1
X01_66	Job satisfaction: other aspects of job	1
X01_8REC	Academic rank	1
	Variables from the original set	
DISCIPLINE	Principal field of teaching/research	10
Q12A	Appointments: Acting	1

Variable name	Variable definition	*df*
Q12E	Appointments: Clinical	1
Q12F	Appointments: Research	1
Q19	Current position as primary employment	1
Q26	Positions outside higher education during career	1
Q30B	Hours/week unpaid activities at the institution	1
Q31A4	Time actually spent on professional growth (percentage)	1
Q31A6	Time actually spent on service activity (percentage)	1
Q64	Union status	3
Q80	Number of dependents	1
Q81	Gender	1
Q85	Disability	1
Q87	Marital status	3
Q90	Citizenship status	3
X01_3	Principal activity	1
X01_91RE	Highest educational level of parents	1
X04_0	Carnegie classification of institution	14
X04_84	Ethnicity in single category	3
X37_0	Bureau of Economic Analysis (BEA) region code	8

Table C7

Initial Variable Evaluation of Model II (Forced Entry)

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$	Tolerance
Intercept	Intercept	9.9808	0.0935	106.70	<.0001	.
Q12A	Appointments: Acting	0.0125	0.0221	0.56	0.5735	0.9839
Q12E	Appointments: Clinical	-0.0346	0.0162	-2.14	0.0323	0.7945
Q12F	Appointments: Research	0.0352	0.0176	2.00	0.0458	0.8840
Q19	Current position as prim employment	0.0126	0.0470	0.27	0.7882	0.9782
Q26	Positions outside hi-ed during career	-0.0038	0.0019	-1.99	0.0462	0.8914
Q30B	Hrs/week unpaid activities at the inst.	-0.0003	0.0007	-0.47	0.6381	0.9397
Q31A4	Time actually spent on prof. growth (%)	-0.0023	0.0006	-3.54	0.0004	0.8042
Q31A6	Time actually spent on service (%)	0.0012	0.0004	3.16	0.0016	0.5042
X01_91RE	Highest education level of parents	-0.0013	0.0030	-0.41	0.6800	0.9013
Q80	Number of dependents	0.0051	0.0031	1.64	0.1010	0.7770
Q81	Gender	-0.0621	0.0088	-7.06	<.0001	0.7549
Q85	Disability	-0.0230	0.0213	-1.08	0.2802	0.9752
Q29A1	Career creative works, juried media	0.0019	0.0002	9.50	<.0001	0.3751
X15_16	Years since highest degree	0.0076	0.0005	16.96	<.0001	0.6822
Q31A1	Time spent teaching undergrads (%)	-0.0010	0.0003	-3.95	<.0001	0.1873
Q31A2	Time spent at teaching graduates (%)	0.0001	0.0003	0.23	0.8156	0.3778
X02_49	Indv. instruction w/grad &1st prof stud	0.0007	0.0005	1.50	0.1339	0.7521
Q32B1	No. of undergrad committees chaired	0.0055	0.0039	1.40	0.1616	0.9551
Q31A5	Time spent at administration (%)	0.0018	0.0003	5.66	<.0001	0.3157
Q16A1REC	Highest degree type	0.0842	0.0052	16.19	<.0001	0.6213

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$	Tolerance
Q24A5REC	Rank at hire for 1st job in higher ed	0.0077	0.0031	2.44	0.0146	0.8676
Q29A3	Career reviews of books/creative wks	0.0017	0.0004	4.11	<.0001	0.7750
Q29A5	Career presentations, performances	0.0001	0.0001	2.15	0.0313	0.7843
Q29A4	Career books, textbooks, reports	-0.0001	0.0005	-0.11	0.9095	0.8815
X10_0	Inst size: FTE graduate enrollment	-0.0019	0.0008	-2.46	0.0139	0.6353
Q76G	Consulting/freelance income	0.0000	0.0000	5.43	<.0001	0.9031
X01_66	Job satisfaction: other aspects of job	0.0523	0.0060	8.79	<.0001	0.9244
X01_8REC	Academic rank	0.0492	0.0033	14.96	<.0001	0.6169
TCHPUP	Interaction of Q31A1 and Q29A1	0.0000	0.0000	-1.35	0.1766	0.5368
BEA1	New England	-0.0548	0.0200	-2.74	0.0061	0.6670
BEA2	Mid East	0.0039	0.0150	0.26	0.7943	0.4754
BEA3	Great Lakes	-0.0529	0.0145	-3.64	0.0003	0.4560
BEA4	Plains	-0.0834	0.0170	-4.91	<.0001	0.5765
BEA5	Southeast	-0.0781	0.0140	-5.57	<.0001	0.3727
BEA6	Southwest	-0.0791	0.0176	-4.49	<.0001	0.5408
BEA7	Rocky Mountain	-0.0879	0.0220	-3.99	<.0001	0.6873
BEA8	U.S. service school	0.1276	0.1396	0.91	0.3607	0.9488
DSCPL1	Agriculture & home economics	-0.0281	0.0311	-0.90	0.3655	0.7144
DSCPL2	Business	0.1164	0.0231	5.03	<.0001	0.4907
DSCPL3	Education	-0.0651	0.0218	-2.98	0.0029	0.4723
DSCPL4	Engineering	0.0679	0.0251	2.71	0.0068	0.5125
DSCPL5	Fine arts	-0.0447	0.0248	-1.80	0.0717	0.5258
DSCPL6	Health sciences	0.0843	0.0186	4.52	<.0001	0.3421

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$	Tolerance
DSCPL7	Humanities	-0.0628	0.0200	-3.14	0.0017	0.3138
DSCPL8	Natural sciences	-0.0221	0.0195	-1.13	0.257	0.2344
DSCPL9	Social sciences	-0.0252	0.0205	-1.23	0.2201	0.3609
DSCPL10	All other programs	0.0144	0.0198	0.73	0.4667	0.3840
ETHNIC1	Native American	0.0418	0.0382	1.09	0.2749	0.9827
ETHNIC2	Asian American	0.0394	0.0170	2.32	0.0202	0.7231
ETHNIC3	Africa American	0.0117	0.0155	0.75	0.4516	0.9169
PRIMACT1	Primary activity: teaching	-0.0467	0.0174	-2.69	0.0072	0.2072
PRIMACT2	Primary activity: research	-0.0011	0.0205	-0.06	0.956	0.3358
PRIMACT3	Primary activity: administration	0.0564	0.0206	2.74	0.0062	0.2961
MARITAL1	Married	-0.0057	0.0131	-0.44	0.6626	0.4381
MARITAL2	Living with someone	-0.0125	0.0242	-0.52	0.6061	0.7974
MARITAL3	Separated, divorced, or widowed	0.0086	0.0163	0.53	0.5957	0.5535
CITIZEN1	US citizen, naturalized	-0.0068	0.0150	-0.45	0.6515	0.7963
CITIZEN2	Permanent resident	-0.0159	0.0172	-0.92	0.3577	0.8074
CITIZEN3	Non-immigrant visa	-0.0289	0.0358	-0.81	0.4205	0.9348
UNION1	Union availlabe, not eligible	-0.0437	0.0255	-1.71	0.0867	0.9534
UNION2	Eligible, but not a member	-0.0085	0.0112	-0.76	0.4486	0.8663
UNION3	Union member	0.0314	0.0114	2.75	0.0060	0.6732
STRATA1	Public comprehensive	0.0092	0.0242	0.38	0.7032	0.1710
STRATA2	Private comprehensive	-0.0303	0.0266	-1.14	0.2538	0.3498
STRATA3	Public liberal arts	-0.0038	0.0343	-0.11	0.9118	0.5691
STRATA4	Private liberal arts	-0.0895	0.0262	-3.42	0.0006	0.3035

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$	Tolerance
STRATA5	Public medical	0.2335	0.0334	7.00	<.0001	0.4879
STRATA6	Private Medical	0.2331	0.0451	5.17	<.0001	0.7167
STRATA7	Private religious	-0.1529	0.0524	-2.92	0.0036	0.7871
STRATA8	Public 2-year	0.0353	0.0255	1.38	0.1671	0.1599
STRATA9	Private 2-year	0.0117	0.0591	0.20	0.8433	0.8306
STRATA10	Public other	-0.0106	0.0564	-0.19	0.8516	0.8116
STRATA11	Private other	-0.0983	0.0439	-2.24	0.0251	0.7460
STRATA12	Public research	0.0765	0.0230	3.33	0.0009	0.1375
STRATA13	Private research	0.1329	0.0262	5.06	<.0001	0.3107
STRATA14	Public other Ph.D.	0.0059	0.0256	0.23	0.8182	0.2866

Table C8

Stepwise Selection of Model II: Binary Variables from the Same Categorical Variable Evaluated as a Group

Step	Variable entered	Label	*df*	Partial R^2	Model R^2	*F* value	$p > F$
1	Q29A1	Career creative works, juried media	1	0.1945	0.1945	1580.11	<.0001
2	Q31A1	Time actually spent teaching undergraduates (percentage)	2	0.0917	0.2862	840.57	<.0001
3	X01_8REC	Academic rank	3	0.0681	0.3543	689.94	<.0001
4	X15_16	Years since highest degree	4	0.0350	0.3893	374.68	<.0001
5	Q16A1REC	Highest degree type	5	0.0311	0.4204	351.18	<.0001
6	STRATA	Carnegie institution classification	19	0.0274	0.4478	23.13	<.0001
7	DISCIPLINE	Principal field of teaching/research	29	0.0182	0.4660	22.21	<.0001
8	PRIMACT	Primary activity	32	0.0102	0.4762	42.10	<.0001
9	BEA	BEA region codes	40	0.0088	0.4850	13.90	<.0001
10	X01_66	Job satisfaction: other aspects of job	41	0.0061	0.4911	78.39	<.0001
11	Q81	Gender	42	0.0052	0.4963	67.55	<.0001
12	Q31A5	Time actually spent at administration (percentage)	43	0.0025	0.4989	32.77	<.0001
13	Q76G	Consulting/freelance income	44	0.0024	0.5012	30.91	<.0001
14	Q31A4	Time actually spent on professional growth (percentage)	45	0.0013	0.5025	17.29	<.0001
15	Q29A3	Career reviews of books, creative works	46	0.0013	0.5038	16.83	<.0001
16	Q31A6	Time spent on service activity (percentage)	47	0.0011	0.5049	14.11	0.0002
17	Q24A5REC	Rank at hire for 1st job in higher education	48	0.0006	0.5055	7.31	0.0069
18	UNION	Union status	51	0.0009	0.5064	4.01	0.0073
19	X10_0	Appointments: Clinical	52	0.0005	0.5069	7.07	0.0079

APPENDIX D

BUILDING THE DATA MINING MODEL: BBN MODEL III

Table D1

Binning Schema of Continuous Variables in BBN Learning

Variable name	Variable definition	Number of bins
Q10AREC	Years achieved tenure	6
Q16B2REC	Years since 2nd highest degree	6
Q23	Positions in higher education during career	3
Q24A1REC	Years since 1st job in higher education	7
Q25	Years teaching in higher education institution	8
Q26	Positions outside higher education during career	3
Q29A1	Career creative works, juried media	8
Q29A2	Career creative works, non-juried media	6
Q29A3	Career reviews of books, creative works	4
Q29A4	Career books, textbooks, reports	4
Q29A5	Career presentations, performances	7
Q29B1	Recent sole creative works, juried media	5
Q29B2	Recent sole creative works, non-juried media	4
Q29B3	Recent sole reviews of books, creative works	3
Q29B4	Recent sole books, textbooks, reports	2
Q29B5	Recent sole presentations, performances	5
Q29C1	Recent joint creative works, juried media	4
Q29C2	Recent joint creative works, non-juried media	2
Q29C3	Recent joint reviews of books, creative works	2
Q29C4	Recent joint books, textbooks, reports	2
Q29C5	Recent joint presentations, performances	4

Variable name	Variable definition	Number of bins
Q30B	Hours/week unpaid activities at the institution	3
Q30C	Hours/week paid activities not at the institution	3
Q30D	Hours/week unpaid activities not at the institution	4
Q31A1	Time actually spent teaching undergraduates (percentage)	6
Q31A2	Time actually spent at teaching graduates (percentage)	4
Q31A3	Time actually spent at research (percentage)	6
Q31A4	Time actually spent on professional growth (percentage)	4
Q31A5	Time actually spent at administration (percentage)	6
Q31A6	Time actually spent on service activity (percentage)	4
Q31A7	Time actually spent on consulting (percentage)	3
Q32A1	Number of undergraduate committees served on	2
Q32A2	Number of graduate committees served on	3
Q32B1	Number of undergraduate committees chaired	2
Q32B2	Number of graduate committees chaired	3
Q33	Total classes taught	5
Q40	Total credit classes taught	5
Q50	Total contact hours/week with students	5
Q51	Total office hours/week	5
Q58	Total number of grants or contracts	3
Q59A	Total funds from all sources	4
Q61SREC	Work support availability	6
Q76G	Consulting/freelance income	3
Q7REC	Years on current job	7
Q80	Number of dependents	3

Variable name	Variable definition	Number of bins
Q9REC	Years on achieved rank	7
SALARY	Basic salary of the 1998-99 academic year	24
X01_82	Age	8
X02_49	Individual instruction w/grad &1st professional students	4
X03_49	Number of students receiving individual instruction	5
X04_41	Total classroom credit hours	6
X09_76	Total income not from the institution	6
X10_0	Ratio: FTE enrollment / FTE faculty	7
X15_16	Years since highest degree	7
X21_0	Institution size: FTE graduate enrollment	6
X25_0	Institution size: Total FTE enrollment	7
X46_41	Undergraduate classroom credit hours	5
X47_41	Graduate and 1st professional classroom credit hours	3

APPENDIX E

BUILDING THE COMBINATION MODEL: REGRESSION MODEL IV

Table E1

Summary of the Stepwise Selection of Model IV: Binary Variables from the Same Categorical Variable Evaluated as a Group

Step	Variable entered	Label	No. of vars	Partial R^2	Model R^2	F Value	$p > F$
1	Q29A1	Career creative works, juried media	1	0.1951	0.1951	1611.45	<.0001
2	Q31A1	Time actually spent teaching undergrads (percentage)	2	0.0920	0.2870	857.90	<.0001
3	X01_8REC	Academic rank	3	0.0655	0.3525	672.16	<.0001
4	X15_16	Years since highest degree	4	0.0359	0.3884	390.03	<.0001
5	STRATA	Carnegie institution classification	18	0.0330	0.4214	27.04	<.0001

Note: All independent variables were entered. The dependent variable was log-transformed SALARY.

APPENDIX F

MODEL COMPARISONS

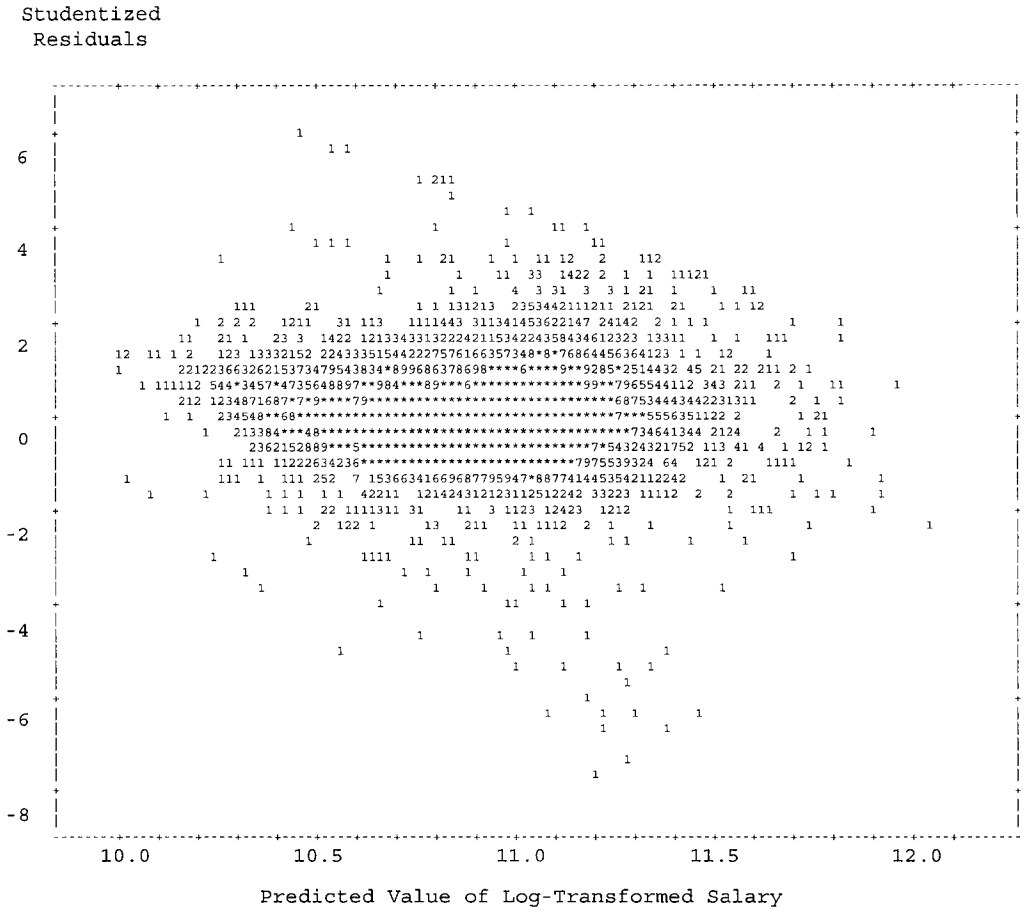


Figure F1. The scatter plot of the studentized residuals against the predicted values for Model I.

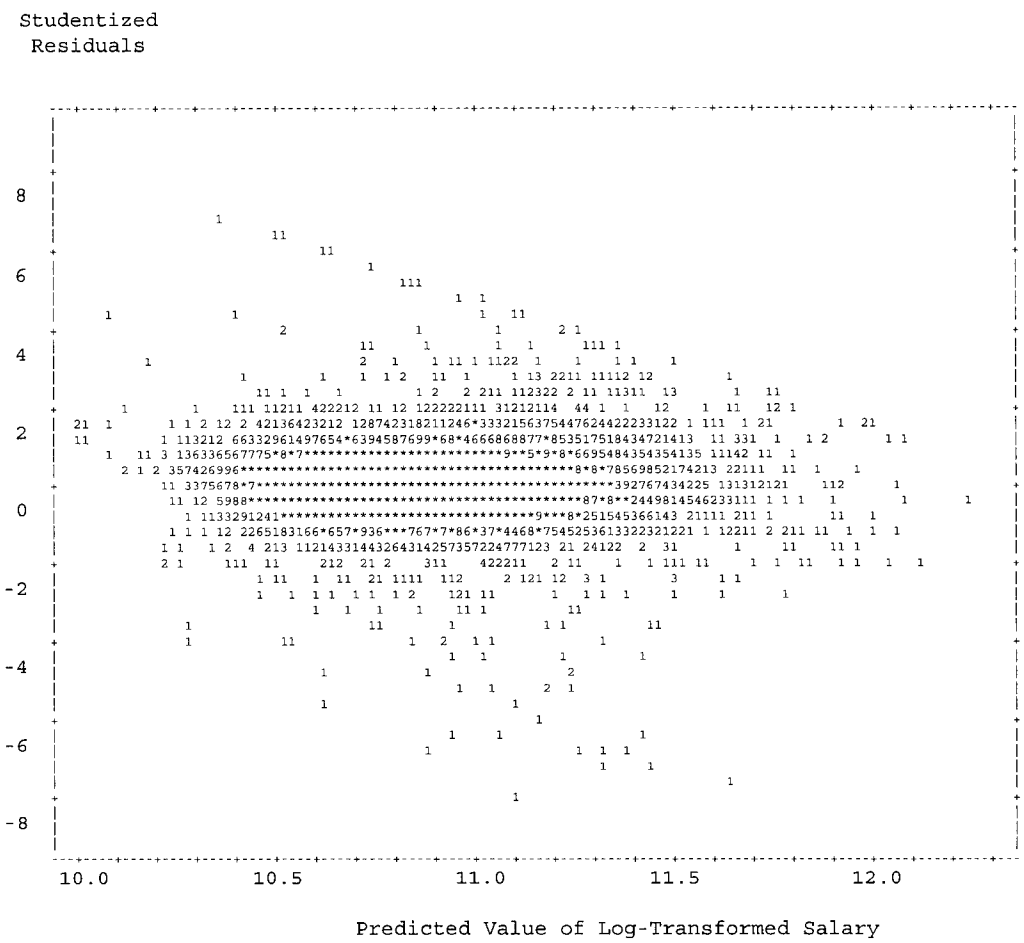


Figure F2. The scatter plot of the studentized residuals against the predicted values for Model II.

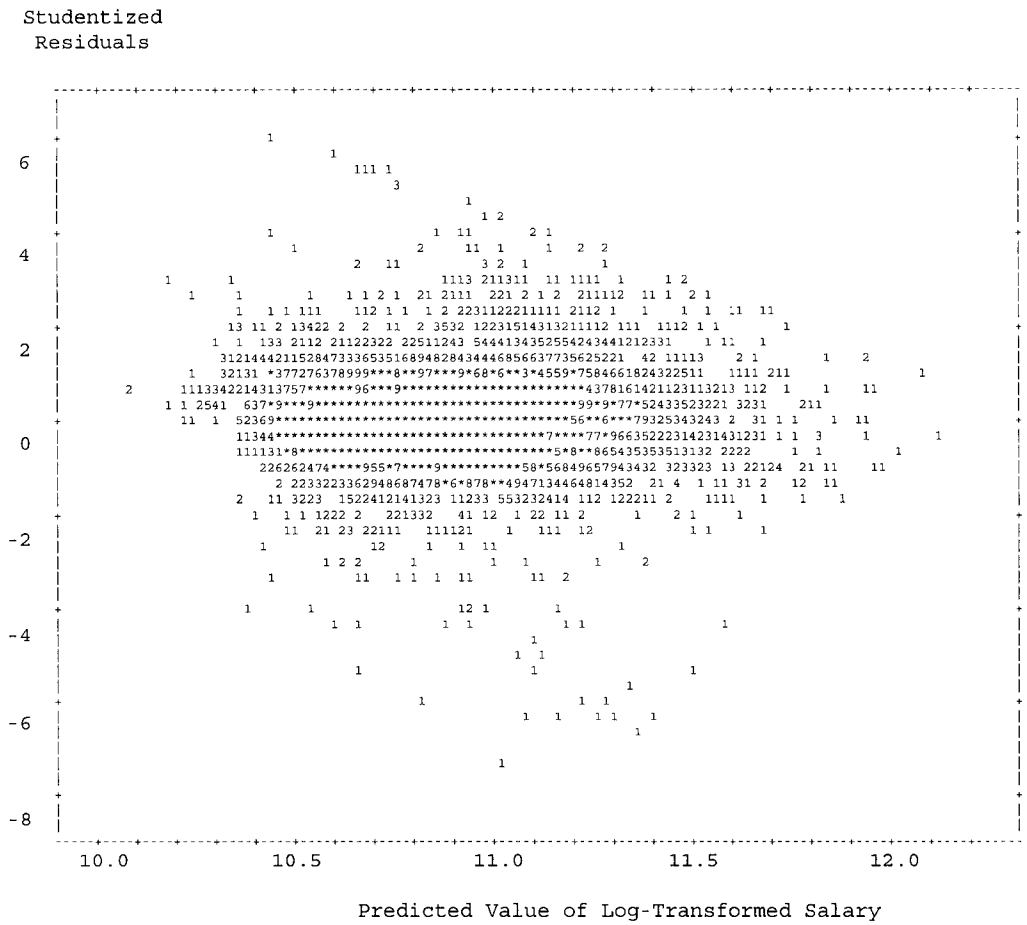


Figure F3. The scatter plots of the studentized residuals against the predicted values for Model IV.

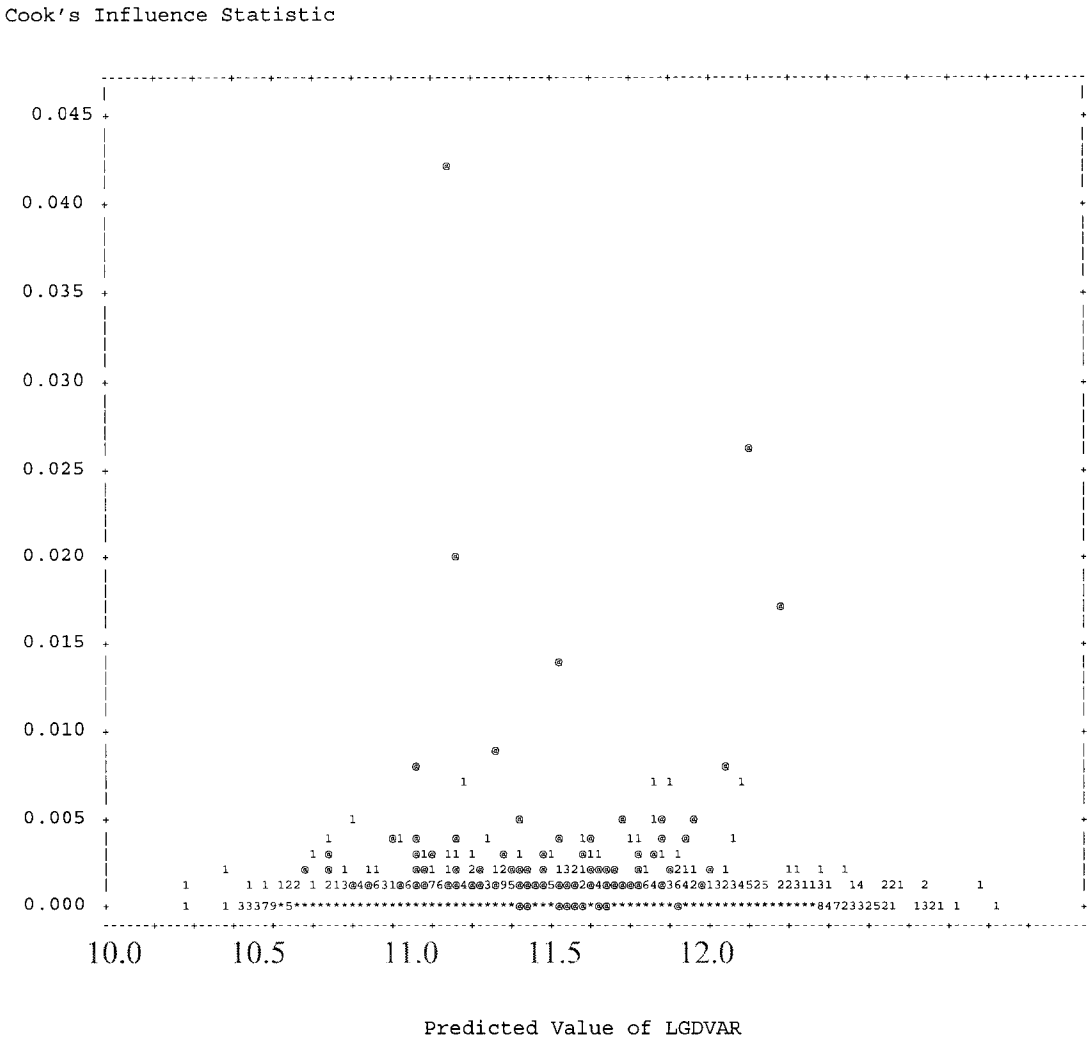


Figure F4. An example of Cook's distance plot showing outliers

Table F1

Stepwise Selection of Model V: Binary Variables from the Same Categorical Variable Evaluated as a Group

Step	Variable entered	Label	No. of vars	Partial R^2	Model R^2	F Value	$p > F$
1	Q29A1	Career creative works, juried media	1	0.1950	0.1950	1608.81	<.0001
2	Q31A1	Time actually spent teaching undergrads (%)	2	0.0916	0.2867	852.74	<.0001
3	X01_8REC	Academic rank	3	0.0662	0.3529	679.49	<.0001
4	X15_16	Years since highest degree	4	0.0357	0.3886	387.39	<.0001
5	Q16A1REC	Highest degree type	5	0.0300	0.4186	342.77	<.0001
6	DISCIPLINE	Principal field of teaching/research	15	0.0236	0.4422	28.05	<.0001
7	BEA	BEA region codes	22	0.0088	0.4511	15.2	<.0001
8	Q81	Gender	23	0.0052	0.4563	63.73	<.0001
9	Q31A5	Time actually spent at admin (%)	24	0.0053	0.4616	65.47	<.0001
10	Q59A	Total funds from all sources	25	0.0036	0.4652	44.26	<.0001
11	Q10REC	Tenure status	26	0.0032	0.4684	39.32	<.0001
12	X08_0D	Inst type: doctoral, 4-yr, or 2-yr	27	0.0020	0.4703	24.57	<.0001

Note. A total of 35 independent variables. The dependent variable was log-transformed SALARY.

Table F2

The ANOVA Table of Multiple Regression Model V

Source	*df*	Sum of squares	Mean square	F	$p > F$
Model	27	579.6364	21.46801	217.51	<.0001
Error	6614	652.7971	0.0987		
Corrected Total	6641	1232.433			

Note. Model R^2 = .4703, adjusted R^2 = .4682, and Root MSE (standard error of estimate) = 0.314.

Table F3

Comparison of Model Prediction Accuracy by Salary Bins (the Training Data Set)

Distance by No. of bins	Model I		Model II		Model III		Model IV	
	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
0	637	9.58%	688	10.34%	1707	25.66%	566	8.51%
1	1140	17.14%	1254	18.85%	930	13.98%	1132	17.02%
2	1056	15.87%	1184	17.80%	777	11.68%	1027	15.44%
3	925	13.91%	910	13.68%	639	9.61%	989	14.87%
4	701	10.54%	724	10.88%	493	7.41%	766	11.52%
5	610	9.17%	546	8.21%	413	6.21%	589	8.85%
6	476	7.16%	418	6.28%	340	5.11%	446	6.70%
7	344	5.17%	298	4.48%	266	4.00%	361	5.43%
8	248	3.73%	200	3.01%	226	3.40%	231	3.47%
9	177	2.66%	133	2.00%	189	2.84%	193	2.90%
10	99	1.49%	61	0.92%	155	2.33%	118	1.77%
11	76	1.14%	70	1.05%	119	1.79%	66	0.99%
12	43	0.65%	44	0.66%	102	1.53%	44	0.66%
13	27	0.41%	30	0.45%	90	1.35%	46	0.69%
14	30	0.45%	24	0.36%	44	0.66%	21	0.32%
15	18	0.27%	14	0.21%	34	0.51%	16	0.24%
16	10	0.15%	14	0.21%	35	0.53%	16	0.24%
17	7	0.11%	11	0.17%	24	0.36%	8	0.12%
18	8	0.12%	9	0.14%	25	0.38%	5	0.08%
19	8	0.12%	6	0.09%	12	0.18%	6	0.09%
20	7	0.11%	7	0.11%	11	0.17%	3	0.05%
21			2	0.03%	10	0.15%	3	0.05%
22					9	0.14%		
23					2	0.03%		
Missing	5	0.08%	5	0.08%	0		0	
Total	6652	100%	6652	100%	6652	100%	6652	100%

Table F4

Comparison of Model Prediction Accuracy by Salary Bins (the Testing Data Set)

Distance by No. of bins	Model I		Model II		Model III		Model IV	
	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
0	293	8.85%	347	10.48%	383	11.57%	310	9.36%
1	603	18.21%	659	19.90%	615	18.57%	551	16.64%
2	494	14.92%	528	15.95%	446	13.47%	519	15.68%
3	448	13.53%	471	14.23%	370	11.17%	443	13.38%
4	367	11.08%	367	11.08%	312	9.42%	402	12.14%
5	292	8.82%	279	8.43%	270	8.15%	319	9.63%
6	237	7.16%	202	6.10%	186	5.62%	206	6.22%
7	191	5.77%	157	4.74%	155	4.68%	157	4.74%
8	104	3.14%	98	2.96%	130	3.93%	137	4.14%
9	90	2.72%	52	1.57%	99	2.99%	77	2.33%
10	56	1.69%	34	1.03%	87	2.63%	63	1.90%
11	44	1.33%	25	0.76%	58	1.75%	39	1.18%
12	22	0.66%	20	0.60%	42	1.27%	29	0.88%
13	17	0.51%	17	0.51%	39	1.18%	19	0.57%
14	11	0.33%	10	0.30%	35	1.06%	10	0.30%
15	17	0.51%	15	0.45%	22	0.66%	8	0.24%
16	5	0.15%	5	0.15%	19	0.57%	7	0.21%
17	5	0.15%	7	0.21%	11	0.33%	6	0.18%
18	3	0.09%	1	0.03%	8	0.24%	2	0.06%
19	4	0.12%	2	0.06%	7	0.21%	3	0.09%
20	0	0.00%	7	0.21%	7	0.21%	2	0.06%
21	2	0.06%	0	0.00%	6	0.18%	2	0.06%
22			2	0.06%	3	0.09%		
23					1	0.03%		
Missing	6	0.18%	6	0.18%	0		0	
Total	3311	100%	3311	100%	3311	100%	3311	100%

Table F5

Model I Rebuilt from Four Subsets in Comparison to the Original

Data set	Subset 1	Subset 2	Subset 3	Subset 4	Training
Size	2491	2491	2491	2490	6652
	Final model				
Model *df*	23	25	23	25	26
R^2	0.4001	0.4480	0.3960	0.4537	0.4213
Adjusted R^2	0.3945	0.4424	0.3904	0.4482	0.4190
	Predictor variables in the model				
	X01_8REC	X01_8REC	X01_8REC	X01_8REC	X01_8REC
	X08_0D	X08_0D	X08_0D	X08_0D	X08_0D
	Q24A1REC	Q24A1REC	Q24A1REC	Q24A1REC	Q24A1REC
	DISCIPLINE	DISCIPLINE	DISCIPLINE	DISCIPLINE	DISCIPLINE
	Q16A1REC	Q16A1REC	Q16A1REC	Q16A1REC	Q16A1REC
	Q59A	Q59A	Q59A	Q59A	Q59A
	TOTPUB	TOTPUB	TOTPUB	TOTPUB	TOTPUB
	BEA	BEA	BEA	BEA	BEA
		Q81		Q81	Q81
		Q13		Q13	Q13
					Q10AREC

Table F6

Model II Rebuilt from Four Subsets in Comparison to the Original

Data set	Subset 1	Subset 2	Subset 3	Subset 4	Training
Size	2491	2491	2491	2490	6652
			Final model		
Model *df*	42	44	41	42	47
R^2	0.5013	0.5347	0.4934	0.5307	0.5036
Adjusted R^2	0.4927	0.5263	0.4849	0.5226	0.5001
			Predictor variables in the model		
	Q31A1	Q31A1	Q31A1	Q31A1	Q31A1
	X01_8REC	X01_8REC	X01_8REC	X01_8REC	X01_8REC
	X15_16	X15_16	X15_16	X15_16	X15_16
	X01_66	X01_66	X01_66	X01_66	X01_66
	STRATA	STRATA	STRATA	STRATA	STRATA
	DISCIPLINE	DISCIPLINE	DISCIPLINE	DISCIPLINE	DISCIPLINE
	Q16A1REC	Q16A1REC	Q16A1REC	Q16A1REC	Q16A1REC
	Q29A1	Q29A1	Q29A1	Q29A1	Q29A1
	BEA	BEA	BEA	BEA	BEA
		Q31A5	Q31A5	Q31A5	Q31A5
		Q81	Q81	Q81	Q81
	Q31A4	Q31A4			Q31A4
	Q76G		Q76G		Q76G
	PRIMACT			PRIMACT	PRIMACT
			Q31A6		Q31A6
		Q26			
		UNION			
					Q29A3

Table F7

Model III (the BBN Model) Rebuilt from the Four Subsets in Comparison to the Original

Training data	Subset 1	Subset 2	Subset 3	Subset 4	Training
Sample size	2491	2491	2491	2490	6652
			Final model		
Threshold	10	12.5	8	8	12.5
Conditional dependency	21	12	24	20	10
			Prediction accuracy		
Subset 1	**1627 (65.32 ± 1.87%)**	275 (11.04 ± 1.23%)	274 (11.00 ± 1.24%)	311 (12.48 ± 1.3%)	509 (20.43 ± 1.58%)
Subset 2	270 (10.84 ± 1.22%)	**1050 (42.15 ± 1.94%)**	278 (11.16 ± 1.24%)	303 (12.16 ± 1.28%)	542 (20.76 ± 1.62%)
Subset 3	259 (10.40 ± 1.20%)	278 (11.16 ± 1.24%)	**1386 (55.64 ± 1.95%)**	298 (11.96 ± 1.27%)	501 (20.11 ± 1.57%)
Subset 4	267 (10.72 ± 1.22%)	261 (10.48 ± 1.20%)	271 (10.88 ± 1.22%)	**1355 (54.42 ± 1.96%)**	538 (21.61 ± 1.62%)
Variables	X04_0 X01_8REC Q31A1 Q29A1 X15_16 Q10AREC DISCIPLINE Q29A2 Q25	X04_0 X01_8REC Q31A1 Q29A1 X15_16 DISCIPLINE	X04_0 X01_8REC Q31A1 Q29A1 Q10AREC Q25 Q16A1REC X46_41 Q32A2 Q29A3	X04_0 X01_8REC Q31A1 Q29A1 X15_16 Q10AREC Q29A2 Q16A1REC Q24A1REC	X04_0 X01_8REC Q31A1 Q29A1 X15_16 Q10AREC

Table F8

Model IV Rebuilt from Four Subsets in Comparison to the Original

Data set	Subset 1	Subset 2	Subset 3	Subset 4	Training
Size	2491	2491	2491	2490	6652
	Final model				
Model *df*	31	30	30	30	18
R^2	0.4797	0.5057	0.4722	0.4938	0.4214
Adjusted R^2	0.4732	0.4996	0.4658	0.4876	0.4199
	Predictor variables in the model				
	Q31A1	Q31A1	Q31A1	Q31A1	Q31A1
	Q29A1	Q29A1	Q29A1	Q29A1	Q29A1
	X01_8REC	X01_8REC	X01_8REC	X01_8REC	X01_8REC
	X15_16	X15_16	X15_16	X15_16	X15_16
	STRATA	STRATA	STRATA	STRATA	STRATA
	DISCIPLINE	DISCIPLINE	DISCIPLINE	DISCIPLINE	
	Q16A1REC	Q16A1REC	Q16A1REC	Q16A1REC	
	Q25		Q25	Q25	
	Q10AREC	Q10AREC			

REFERENCES

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Beverly Hills, CA: Sage Publications.

Allinson, N. M. (1999, February 26). Lie, damned lies and... [Letter to the editor]. *Computing and Control Engineering Journal*, April 1999, p.70.

American Association of Community Colleges, Washington D.C. (1998). *AACC annual, 1998-99: State-by-state analysis of community college trends and statistics.* Washington, D. C. (ERIC Document Reproduction Service No. ED422994).

American Heritage Dictionary of the English Language (4th Ed.). (2000). Dell Publishing Company.

Aptė, C. (1997). Data mining: An industrial research perspective. *IEEE Computational Science and Engineering, 4*(2), 6-9.

Bella, M. L., Ritchey, P. N., & Parmer, P. (2001). Gender differences in the salaries and salary growth rates of university faculty: An exploratory study. *Sociological Perspectives, 44*(2), 163-187.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bursteinas, B., & Long, J. A. (2001). Transforming supervised classifiers for feature extraction (extended version). *International Journal on Artificial Intelligence Tools, 10*(4), 663-674.

Chen, J., & Greiner, R. (1999). Comparing Bayesian network classifiers. *Proceedings of the Fifteenth Conference on Uncertainty In Artificial Intelligence (UAI), Sweden,* 101-108.

Chen, J. (2001). The Belief Network Powersoft (Version 1.1 beta) [Computer software]. Shareware available at http://www.cs.ualberta.ca/~jcheng/bnpp.htm .

Chen, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2001). Learning Bayesian networks from data: An information-theory based approach. *Artificially Intelligence, 137*(1-2), 43-100.

Chou, Y. (1972). *Probability and statistics for decision making.* New York: Holt, Rinehart, and Winston.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Lawrence Erlbaum Associates.

Cumming, M. (2003). *Bayesian belief networks.* Retrieved on March 20, 2003 from http://murrayc.com/learning/AI/bbn.shtml#Introduction

Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Representative subset selection. *Analytical Chimica Acta, 468*(1), 91-103.

Datta, P. (1999). Business focused evaluation methods: A case study. In J. M. Zytkow, & J. Rauch (Eds.), *Principles of Data Mining and Knowledge Discovery: Proceedings at the Third European Conference, PKDD'99.* Prague, Czech Republic.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithms (with discussion). *Journal of the Royal Statistical Society, B39*, 1-38.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications.* New York: John Wiley & Sons.

Elder, J. F., & Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, R. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 83-113). Menlo Park, CA: AAAI/MIT Press.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). New York : Halsted Press.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, R. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1-37). Menlo Park, CA: AAAI/MIT Press.

Fayyad, U.M. (Ed.). (1997a). Editorial in: *Data Mining and Knowledge Discovery: An International Journal, 1,* 1.

Fayyad, U. M. (1997b, August). *Data mining and knowledge discovery in databases: Implications for scientific databases.* Paper presented at 9th International Conference on Scientific and Statistical Database Management (SSDBM'97), Olympia, WA.

Fisher, R.A. (1966). *The design of experiment* (6th ed.). Edinburgh, London: Oliver & Boyd.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheu, C. J. (1991). Knowledge discovery in database: An overview. In G. Piatetsky-Shapiro, & W. J. Frawley (Eds.), *Knowledge Discovery in Databases* (pp. 1-27). MIT: AAAI Press.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, 23*, 881-889.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association, 82,* 249-266.

Friedman, J. H. (1997). Data mining and statistics: What's the connection? In D. W. Scott (Ed.), *Computing Science and Statistics: Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 3-9). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29,* 131-161.

Friedman, N., Goldszmidt, M., & Lee, T. J. (1998). Bayesian network classification with continuous attributes: Getting the best of both discretization and parametric fitting. In J. Shavlik (Ed.), *Proceedings of the 15th International Conference on Machine Learning (ICML)* (pp. 179-187). San Francisco, CA: Morgan Kaufmann.

Freund, R. J., & Wilson, W. J. (1998). *Regression analysis: Statistical modeling of a response variable.* San Diego, CA: Academic Press.

Gelman, A., & Meng, X. (1996). Model Checking and model improvement. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 189-202). London: Chapman & Hall.

Gelman, A., Meng, X., & Stern, H. (1995). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*(4), 733-807.

Gillies, D. (2001). Bayesianism and the fixity of the theoretical framework. In D. Corfield, & J. Williamson (Eds.), *Foundations of Bayesianism.* Dordrecht, Boston, London: Kluwer Academic.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1996). Statistical inference and data mining. *Communications of the ACM, 3*(11), 35-41.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery, 1,* 11-28.

Groth, R. (1999). *Data mining: Building competitive advantage.* New Jersey: Prentice Hall.

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930.* New York: Wiley.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52,* 112-118.

Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *SIGKDD Exploration, 1,* 16-19.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discover, 1*, 79-119.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (2000). *Understanding robust and exploratory data analysis.* New York: John Wiley & Sons.

Huang, G. (1995). *National data for studying rural education: Elementary and secondary education applications.* ERIC Digest: ERIC Clearinghouse on Rural Education and Small Schools, Charleston, WV. (ERIC Document Reproduction Service No. ED 383518).

Huber, P. J. (1994). Huge data sets. In R. Dutter, & W. Grossmann (Eds.), *Compstat 1994 Proceedings* (pp. 3-13). Heidelberg: Physica Verlag.

Huber, P. J. (1999). Massive datasets workshop: Four years after. *Journal of Computational and Graphical Statistics, 8*(3), 635 –652.

Ibrahim, J.G., Chen, M., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer-Verlag.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice Hall.

Johnson, R. J. (1999). Female/male salary inequities: The role of promotion. *On Campus with Women, 28*(2), 2,15.

Jones, P. B., & Sabers, D. L. (1992). Examining test data using multivariate procedures. In M. Zeidner, & R. Most (Eds.), *Psychological Testing*. Palo Alto, CA: Consulting Psychologists Press.

Kira, K., & Rendell, L. (1992). A practical approach to feature selection. In D. Sleeman, and P. Edwards (Eds.), *Proceedings of the Ninth International Conference on Machine Learning* (ICML-92) (pp. 249-256). San Mateo, CA: Morgan Kaufmann.

Kohavi, R., & John, G. H. (1998). The Wrapper approach. In H. Liu, & H. Motoda (Eds.), *Feature Extraction, Construction, and Selection: A Data Mining Perspective* (pp. 33-50). Boston: Kluwer Academic Publishers.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1997). Comparing stochastic complexity minimization algorithms in estimating missing data. *Proceedings of WUPES'97, the 4th Workshop on Uncertainty Processing*. Prague, Czech Republic, 81-90.

Krzanowski, W. J., & Marriott, F. H. C. (1994). *Multivariate analysis part I: Distributions, ordination and inference.* Great Britain: Edward Arnold.

Kubat, M., Bratko, I., & Michalski, R. S. (1998). A review of machine learning methods. In R. S. Michalski, I. Bratko, & M. Kubat (Eds.), *Machine Learning and Data Mining: Methods and Applications* (pp. 3-69). Chichester: John Wiley & Sons.

Liu, H., & Motoda, H. (1998a). *Feature selection for knowledge discovery and data mining.* Boston: Kluwer Academic.

Liu, H., & Motoda, H. (Eds.). (1998b). *Feature extraction, construction, and selection: A data mining perspective.* Boston: Kluwer Academic.

Mannila, H. (1996). Data mining: Machine learning, statistics, and databases. *Proceedings of Eighth International Conference on Statistics and Scientific Database Management*, Stockholm, 2-8.

Maruyama, G. (1997). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage Publications.

McLaugulin, G. W., & McLaugulin, J. S. (2003). Conducting a salary-equity study: A consultant's view. *New Directions for Institutional Research, 117,* 97-114.

Michalski, R. S., & Kaufman, A. K. (1998). Data mining and knowledge discovery: A review of issues and a multistrategy approach. In R. S. Michalski, I. Bratko, & M. Kubat (Eds.), *Machine Learning and Data Mining: Methods and Applications* (pp. 71-112). Chichester: John Wiley & Sons.

Minium, E. W., & Clarke, R. B. (1982). *Elements of statistical reasoning.* New York: John Wiley & Sons.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Density-based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(4), 1-14.

Mitchell, T. M. (1997). *Machine learning.* New York, London: McGraw-Hill.

Moore, N. (1993). Faculty salary equity: Issues in regression model selection. *Research in Higher Education, 34*(1), 107-126.

National Survey of Postsecondary Faculty 1999 (NCES Publication No. 2002151) [Restricted-use data file, CD-ROM]. Washington, DC: National Center of Education Statistics.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical model (4^{th} ed.).* Chicago: Irwin.

Niedermayer, D. (1998). *An introduction to Bayesian networks and their contemporary applications*. Retrieved on September 24, 2003 from http://www.niedermayer.ca/papers/bayesian/

Parzen, E. (1997). Data mining, statistical methods mining, and history of statistics. In D. W. Scott (Ed.), *Computing Science and Statistics*: *Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 365-374). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Petersen, A. C. (1997). Using statistics: The perspective from funders and consumers. In D. W. Scott (Ed.), *Computing Science and Statistics*: *Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 375-378). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Pednault, E. P. D. (1999). Statistical learning theory. In R. A. Wilson, & F. Keil (Eds.) *MIT Encyclopedia of the Cognitive Sciences.* Cambridge, MA: MIT Press.

Piramuthu, S. (1998). Evaluating feature selection methods for learning in data mining application. *IEEE 31st Annual Hawaii International Conferences on System Sciences,* 294- 301.

Quenouille, M. H. (1958). *The fundamentals of statistical reasoning.* London: Charles Griffin.

Raiffa, H. (1968). *Decision analysis*: *Introductory lectures on choices under uncertainty.* Reading, MA: Addison-Wesley.

Rao, C. R. (1997). *Statistics and truth: Putting chance to work (2nd ed.).* Singapore, River Edge, NJ: World Scientific.

Rocke, D. (1998). A perspective on statistical tools for data mining applications. *Proceedings of the Second International Conference on Practical Application of Knowledge Discovery and Data Mining, USA,* 313-318.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377-400.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151-1172.

Russell, S. H. (1991). The status of women and minorities in higher education: Findings from the 1988 National Survey of Postsecondary Faculty. *CUPA Journal, 42*(1), 1-11.

Shafer, G., Pearl, J., & Kaufmann, M. (1990). *Readings in uncertain reasoning.* San Mateo, CA: Morgan Kaufmann.

Shapiro, S. S., & Gross, A. J. (1981). *Statistical modeling techniques.* New York: Marcel Dekker.

Savage, L. J. (1972). *The foundations of statistics.* New York: Dover Publications.

Simpson, W. A., & Sperber, W. E. (1988). Salary comparisons: New methods for correcting old fallacies. *Research in Higher Education, 28*(1), 49-66.

Sobel, M. E. (1995). The analysis of contingency tables. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum Press.

StatSoft Electronic Textbook. (n.d.). Online textbooks. Retrieved on July 7, 2003 from http://www.statsoftinc.com/textbook/stathome.html

Thompson, J. R. (1997). Has high speed computing caused a paradigm shift in statistical science? In D. W. Scott (Ed.), *Computing Science and Statistics: Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 185-192). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Ullman, J.B. (1996). Structural equation modeling. In B.G. Tabachnick, & L.S. Fidell (Eds.), *Using Multivariate Statistics* (3rd ed.) (pp. 709-819). New York: Harper Collins College Publishers.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Berlin: Springer-Verlag.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988-999.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation, 14*(10), 2339-2468.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis.* Boston: Duxbury Press.

Wegman, E., J. (1988). Computational statistics: A new agenda for statistical theory and practice. *Journal of the Washington Academy of Science, 78,* 310-322.

Wegman, E., J. (1995). Huge data sets and the frontiers of computational feasibility. *Journal of Computational and Graphical Statistics, 4*(4), 281-295.

Wegman, E., J. (2000). Visions: New techniques and technologies in statistics. *Computer Statistics, 15*(1), 133-144.

Wolfram, S. (2002). *A new kind of science.* Champaign, IL: Wolfram Media.

Winkler, R. L. (1972). *An introduction to Bayesian inference and decision.* New York: Holt, Rinehart, & Winston.

Yu, Y., & Johnson, B. W. (2002). *Bayesian belief network and its applications* (Tech. Rep. UVA-CSCS-BBN-001). Charlottesville, VA: University of Virginia, Center for Safety-Critical Systems.

Yuan, H., Tseng, S., Wu, G., & Zhang, F. (1999). A two-phase feature selection method using both filter and wrapper. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Nashville, TN, 2,* 132-136.

Zheng, H. Y. (1996). *School contexts, principal characteristics, and instructional leadership effectiveness: a statistical analysis.* Paper presented at the Annual Meeting of the American Educational Research Association. New York.