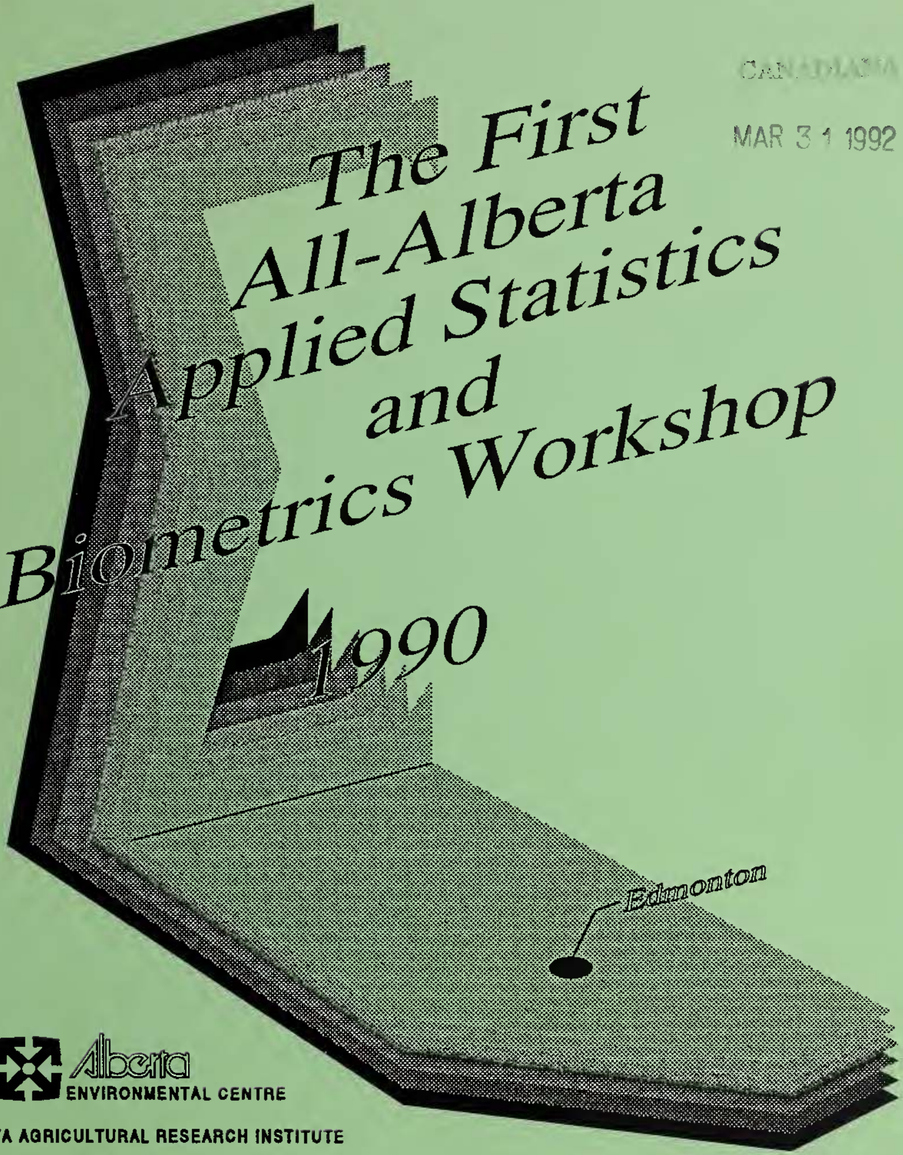
AECV91-P1

Proceedings

The First All-Alberta Applied Statistics and Biometrics Workshop

1990



Alberta
ENVIRONMENTAL CENTRE

ALBERTA AGRICULTURAL RESEARCH INSTITUTE

Alberta
AGRICULTURE

Proceedings of the First All-Alberta Applied Statistics and Biometrics Workshop

L.Z. Florence and L.A. Goonewardene (editors)

October 18-19, 1990

Alberta Research Council

250 Karl Clark Road

Edmonton, Alberta

Canada

Alberta ENVIRONMENTAL CENTRE

ALBERTA AGRICULTURAL RESEARCH INSTITUTE

Alberta AGRICULTURE

DISCLAIMER

The contents of these Proceedings reflect only the opinions of the individual authors and must not be construed in any manner to represent the opinions or policies, official or unofficial, of Alberta Agriculture, the Alberta Environmental Centre or Alberta Environment.

This publication may be cited as:
Florence, L.Z. and L.A.Goonewardene (editors). 1991. Proceedings of the First All-Alberta Applied Statistics and Biometrics Workshop, Oct. 18-19, 1990.
Alberta Environmental Centre, Vegreville, AB. AECV-91P1. 126 pp.

ISBN 0-7732-0636-1

COPYRIGHT STATEMENT

Copies of the complete proceedings may be obtained by contacting:
Alberta Environmental Centre
Postal Bag 4000
Vegreville, Alberta
Canada T0B 4L0

TABLE OF CONTENTS

INTRODUCTION

When we first agreed in mid-1989 that a need existed for an applied (practical) statistics workshop, it hardly occurred to us that as many as 63 other people in Alberta would concur! Yet, that was the final attendance at the First All-Alberta Applied Statistics and Biometrics Workshop, October 18-19, 1990, at the Alberta Research Council, Edmonton (Millwoods facility). We believe these Proceedings will reflect the fine participation and diversity of interests exhibited by the speakers and those who attended.

We are fortunate to have received the support of the Alberta Agricultural Research Institute (AARI) which granted us funds through the Research Coordination Grants program; our sincerest appreciation to the AARI Board for deeming this project worthy of support in 1990. We are most pleased to say that the Board has approved our application and agreed to renew our funding in 1991. Plans are well underway for the next workshop to be held in Edmonton, October 21-22, 1991.

In addition to the list of participants in Appendix B and the speakers featured in these Proceedings, we are indebted to many others who got behind this project and proceeded to help make it happen.

Thanks to Serge Dupuis, Bob Hardin and Keith Toogood for serving on the ad hoc Program Planning Committee.

Arhlene Hrynyk, her staff, and the late Don Klick, in the Animal Sciences Division, Alberta Environmental Centre (AEC), were most helpful during the early stages by preparing mailing lists, survey forms and correspondence. Special appreciation to Phil Henry who lent his ideas, setup the computer aided registration on the MacIntosh and created our "Σ -Alberta" logo, among other graphics. Sincere appreciation as well to Janet Smalley and the support staff of the Beef Cattle and Sheep Branch of Alberta Agriculture for their kind assistance during registration.

Many thanks go to Marilyn Florence who volunteered her assistance during preparation of the program packets and these Proceedings.

Robert Heimann, Alberta Research Council, suggested our having the workshop at the Edmonton (Millwoods) facility and kindly agreed to act as official local host.

Finally, we are particularly pleased to acknowledge the support we have each received from our respective directors and branch heads, and the Departments of Environment and Agriculture.

Notwithstanding the fine suggestions for topics from the Program Planning Committee, the workshop had no theme; our attempt to have a good cross section of applied statistical topics, we believe, is demonstrated by the contents of these Proceedings.

L. Zack Florence
Alberta Environmental Centre

Laki Goonewardene
Alberta Agriculture

APPLYING STATISTICS TO PRACTICAL PROBLEMS

Milton Weiss, Consultant
Pincher Creek, Alberta

Applied statistics in agriculture or biometrics had its start about the end of WWI and early 1920s when R.A. Fisher, Sewall Wright, and L. J. Lush first published on the application of mathematics and probability theory to animal populations. Various other people at about the same time or soon after began applying these same principles to designed situations in field trials. Yates, Fairfield Smith, Haldane, Student, Bartlett, K. Pearson, and others laid a solid foundation for further advances in designed experiments in agronomy as well as quantitative genetics.

The interruption by WWII did not slow this advance, it more aptly put it on simmer since, when hostilities ended, a large number of young men and women returned to universities to develop careers. These numbers were greatly enhanced by the aid of various G.I. bills which made it possible for many to attend university who would otherwise have been denied. This influence along with greatly increased government support for research and development had an immediate and profound influence on applied statistics.

There were many timely and extremely well-written texts published in the 15-20 years following WWII. More importantly, a great number of well-qualified statisticians turned to teaching which allowed many universities in North America to establish statistics as a valid course in either a mathematics department or more importantly to establish biometrics or applied statistics as a field of instruction in agriculture and/or biology faculties.

This has led to the position we are in today, where we can classify our applied statisticians in various ways, but, for simplicity, we will break them into two groups:

1. A statistician with some background and experience in the subject matter field he is serving.
2. A subject matter specialist with some interest and training in applied statistics, particularly with those techniques or tools he uses in his day- to-day work.

A third approach, and one that I have had a lot of experience in, and enjoyment and satisfaction from is where a statistician and a subject matter specialist work jointly or co-operatively on a project.

It is highly desirable (imperative ?) that the statistician be involved in a project from the start. Once a project is deemed necessary or worth carrying-out, the statistician should be made a part of the project team immediately. This will often ensure that the design has the capability, if carried out properly, of answering the questions posed by the project i.e. that the specified goals can be met. The statistician should help with the data collection protocol and definitely has valuable inputs with respect to data edits and validity checks. The analysis to be used should be laid out or specified at this time, complete with tests of hypotheses, linear comparisons, orthogonal contrasts, and so on.

Once the project has progressed to the stage of analysis and interpretation, the statistician again becomes vital particularly in the areas of interpretation and drawing of valid conclusions. The reporting of statistical design, models used, and analysis procedures as well as certain parts of the results section, are best handled by the statistician or co-operatively between the subject matter specialist and the statistician. An area which is becoming less of a problem with more people trained in and/or appreciative of statistics is the role of referee or arbitrator between the author, the editor, and/or the reviewers on statistical matters. The applied statistician also can provide a useful role in technology transfer - boiling the whole process down so that it is meaningful to the end users — not just to other specialists in the subject matter field.

When I first became interested in statistics in the mid 1950s there was one mechanical desk calculator for a whole class to use (and it usually didn't work). I soon became very proficient in hand calculations with a pencil, paper, sliderule, and a very large eraser. I was fortunate to have the opportunity to use the Mystic at Michigan State University and the Cyclone at Iowa State University before they became museum pieces. In 1961 I had my initiation to IBM and the then popular 650, soon to be replaced in our work by the 1620 in early 1962. Overlapping with this experience were the wired board (402, 407, 602, etc.) machines. Instead of programming them, you wired all the instructions one digit at a time into boards and then ran applications from punched cards.

With the 1620 we really started to move forward - every shop acquired a numerical analyst and one or more programmers. It was the "in thing" to learn programming; machine language, symbolic languages and then EUREKA!! we got compilers for Fortran and Cobol. However, computing was so slow and expensive that you spent days and even weeks desk checking, debugging and correcting machine language compiled card decks rather than go on the computer and chance blowing the budget. Every shop was developing programs and exchanging with each other, often with great duplication of effort. In short, there was little time for the real job of applying statistics to user problems.

The next big step was commercial statistical analysis packages - BMD, SPSS, and later SAS. This was fantasy land. Not only could we get the computer to work for us - the instructions involved were simple and easy to follow - too easy in some cases. Along with these new packages we were getting better access to bigger, faster, and less expensive computers. What used to take days and even weeks to accomplish can now be done in minutes, right from your own desk.

Now we not only have terminals but also micro computers that for a very modest price can do virtually everything we need right at hand. SAS and many other statistical packages are available (at a fairly modest cost) for micro-computers.

A note of caution - we can do what we want fast and efficiently in so far as statistics is concerned. However, many times we forget to apply our statistical knowledge to problems. You rarely hear mention of such topics as fixed, mixed, or random models, residual analysis, and so on. Yet these are easily accommodated, at most requiring only a little juggling of the degrees of freedom (df) and corresponding sums of squares (ss) terms in the ANOVA table. The tests of significance and comparisons may or may not require extra work beyond this first manipulation as many packages allow the user to specify the error term to be used.

Personally I like the present situation - I can now accomplish in one relaxing evening what used to take six months (with the help of 3 clerks on calculators) of hard work to achieve.

Time Series Models in Econometrics

Vic Adamowicz
Associate Professor
Department of Rural Economy
University of Alberta
Edmonton, Alberta

Economists have struggled with attempts to model the structure of the economy for decades. They have concentrated on so-called structural models of demand and supply, interest rates and money supply, exchange rates and export quantities, and a variety of other relations which arise from economic theory. The problem for the statistical modeler, however, is that these relationships cannot be isolated from all the other aspects of the economy that function around them. We cannot stop the world to examine the relationship between exports of wheat and the Canada-U.S. exchange rate. A myriad of other influences, including the influence of time, are creating a smoke screen.

The challenge for economists, or more particularly econometricians, is to find a fan to blow the smoke away, so that we can see the correct relationship. In most of the physical sciences, the "fan" is experimental design. Controls are put in place and unwanted smoke is kept away from the relationship in question. In the analysis of social systems, such designs are not possible. The search for a good "fan" is the attempt to resolve the "identification problem."

The identification problem is a statistical technicality that arises when the parameters of the underlying structural model cannot be uniquely "identified" from the data. The classic example is that of supply and demand. A supply curve traces out the response of sellers, in terms of the quantity they wish to sell, to changes in the market price. A demand curve traces out the response of buyers to these same market prices. The data we gather from a market, however, are only the equilibrium quantities and prices (the prices and quantities agreed on by the buyers and sellers in that period). An estimation of the relationship between price and quantity reveals a mix of supply and demand factors and not a unique identification of either set of underlying structural parameters. If we have data on a number of periods of equilibrium, we still cannot sort out whether we are watching demand, supply or some temporal influence on the price-quantity data.

Economists have chosen a number of approaches to modeling or forecasting. Three forms of modeling will be discussed in this paper; the traditional structural approach to modeling economic systems over time, the time series approach and the vector autoregressive approach. The three approaches are quite different in their statistical methods and their method.of addressing the identification problem.

Structural Econometric Modeling

Structural econometric models use time series data to estimate the relationships between economic variables. The statistical techniques are modifications of multivariate analysis. Using the supply-demand example, a structural model would theoretically assign a particular form to the demand relation and the supply relation. For example, both quantities would be a function of price but demand may also be a function of income while supply may be a function of weather factors. Typically there are few explicit dynamic elements in the equations: quantities today are expressed as functions of prices and other factors in the current time period.

The modeling process described above requires a strong theoretical base to provide the specification of the equations. The resulting statistical model has the merit that it is based on theory and should provide more than just "correlations". However, more works needs to be done to identify the parameters of each function. Additional restrictions must be placed on the model to be able to isolate the demand function from the supply function.

In order to "identify" the demand function from the supply function, a very particular set of restrictions is used. The restrictions must produce a model where there are enough exogenous variables (variables not determined within the system) in each equation to allow identification. For example, weather is an exogenous factor. If weather only affects the supply curve, these changes in the supply curve will allow us to trace out all the equilibrium prices and quantities that occur along a single demand curve. The exogenous variable in the supply curve identifies the demand curve. Similarly, there must be exogenous variables in the demand curve which help identify the supply curve. In more complex models the solution to this search for exogenous variables can become quite difficult. The implication for modeling is that if one wants to be able to identify the underlying structural equations in their model, they must exclude some factors from the equations. These restrictions, while solving the identification problem, lead to questions about appropriate specification of the models. The estimation of these models, even after the identification problems has been solved, involves one of many variants of multivariate regressions analysis. These estimation techniques include Two Stage Least Squares, Three Stage Least Squares and a variety of maximum likelihood based approaches (see Judge, et al, 1988).

The structural econometric approach to time series analysis has been criticized for a number of reasons. First, the models tend not to forecast very well. A variety of reasons have been suggested for the poor forecasting performance including the fact that little dynamic influence is included in these models. Nevertheless, structural modelers have maintained that they are attempting to adhere to theory and they suggest that empirical analysis without this formal theory is vacuous. A second criticism of structural models is that they tend to become rather large and expensive to run. An example of a large structural model is Agriculture Canada's FARM model which contains over 1,000 equations in a number of sectors.

One of the most scathing critiques of structural models came from Robert Lucas (1976). The so called "Lucas Critique" is based on the notion that the correct theoretical model depends, at least in part, on the current policy scenario. A change in the policy situation requires a reformulation of the parameters of the theoretical model. In forecasting the impact of a policy change, however, structural modelers leave the coefficients intact and adjust the exogenous variables. Lucas argued that these structural models will undoubtedly provide poor forecasts of policy shocks. Another criticism related to the coefficients of structural models is attributed to Sims (1980). Sims argues that in attempting to identify traditional models, overly restrictive assumptions are likely to be used, resulting in poor models.

The criticisms of structural models led to a wave of other models which were designed for forecasting purposes. These models, which focused on the time series elements in economic data, provide a contrast to the traditional models, not only by their emphasis on the temporal dimension, but also on their lack of explicit theoretical base.

Time Series Modeling

Simple time series models concentrate on explaining the data as a stochastic process. The emphasis is on the temporal structure of the data series. The data are first examined for their structure over time or degree of stationarity. Most simple time series techniques assume that the data are covariance stationary stochastic processes (see Judge, et al, 1988). Given covariance stationarity, the series can be modeled as an "autoregressive process" or a "moving average" process. An example of the former is a first order autoregressive process (AR1) or

[1] $$Q_t = \theta Q_{t-1} + \epsilon_t.$$

A first order moving average process (MA1) can be written as

[2] $$Q_t = \epsilon_t - \phi \epsilon_{t-1}.$$

These simple time series representations focus on past value of the series itself or the error process for explanations of the present observation. A more general time series representation can be specified as

[3] $$\theta(L)Q_t = \phi(L)\epsilon_t$$

where $\theta(L)$and $\phi(L)$are polynomials in the lag operator.[1]

Once the structure of the stochastic process is determined, a model can be estimated and the process can be used for forecasting. A common approach used to determine the structure of time series models is the Box-Jenkins approach. The Box-Jenkins approach involves three steps, (1) Identification, (2) Estimation and (3) Diagnostic checking (see Judge, et al, 1988, chapter 16). Identification includes determining if the process is covariance stationary and using plots of the autocorrelation and partial autocorrelation function to determine the order of the process. The process may be autoregressive, moving average or it may be a mixture of the two. If a model is not stationary, differencing is usually used to remove the trend from the data and reduce it to stationarity. This differencing process is also called integration. Thus a common name for these models, ARIMA, stems from the Autoregressive, Integrated, and/or Moving Average components of the data.

Estimation of these models can proceed in a number of ways. The autoregressive models can be estimated using OLS but a number of other methods are typically used to increase efficiency. The moving average models must be estimated using nonlinear least squares as the use of lagged values of the error term in the model requires iteration around an initial condition. There are two types of diagnostics used to check these models. First, the coefficients are examined to ensure that they do not produce explosive processes. For example, if a first order autoregressive model has a coefficient that exceeds unity, the process is not stationary. Second, the errors of the models are checked to determine if they correspond to white noise or a gaussian mean zero, identically distributed, independent process.

Additional issues in ARIMA modeling include modification of the process for seasonality effects and the use of transfer functions. Seasonality is modeled by including lags corresponding to the seasonal pattern in the data. Transfer functions are somewhat similar to dummy variables in regression analysis. They indicate a break point in the data and a change in coefficient values.

The emphasis in ARIMA models is the time series structure of the data. Cause and effect relationships are ignored in favour of the temporal dimension. This emphasis on data analysis leaves most researchers who believe in theory first and models second somewhat disappointed. One can forecast with ARIMA models but

[1] The symbol L is the lag operator, ie. $LY_t = Y_{t-1}$. The expression $\theta(L)X_t$describes $\theta_0 X_t + \theta_1 X_{t-1} + \theta_2 X_{t-2}\ldots$

the effect of a policy change, unless it is modeled with a transfer function, cannot be simulated. However, even though ARIMA models do not rely on theoretical specifications, they tend to forecast very well, especially in the short term. Numerous studies of forecasting performance find ARIMA models to be at least as good as simple econometric models and they are usually simpler to build and less expensive to estimate. The lack of theoretical base, however, leaves these models open to attack. The next set of models, VAR models, incorporate the time series elements of ARIMA models and the structural econometric elements of traditional models.

Vector Autoregression Models[2]

A Vector Autoregression Model (VAR) is essentially a dynamic simultaneous equation system. The dependent variables are, by definition, all endogenous variables and the independent variables are lagged observations of all variables in the system[3]. Each equation contains the time series structure of an ARIMA model with all variables interacting in the system. A VAR model imposes very few a priori restrictions on the parameters in the simultaneous equation system. This allows the data to provide a representation of the changes in the system without the "zero restrictions" required in traditional simultaneous equation techniques. One should note, however, that while a VAR model does not impose zero restrictions on the parameters in the traditional simultaneous equation fashion, the model does require identification restrictions to provide information on the response of system variables to shocks. The nature of these restrictions will be outlined below.

The VAR approach uses the set of lags of all of the endogenous variables in each behavioral equation as the reduced form or statistical model. The economic structure is identified using the variance matrix of the residuals to place identifying restrictions on the matrix of contemporaneous coefficients. In VAR models, the statistical model is developed first and then the structural model is identified. This is opposite to the approach followed in traditional econometrics and is favoured by some statistical theorists (Spanos, 1989).

While both the VAR approach and traditional econometric approaches require identification restrictions, the nature of these restrictions are quite different. The traditional approach uses zero restrictions on parameters for identification while the VAR approach uses the covariance matrix of the reduced form residuals and the assumption of orthogonal behavioral shocks to establish identification. Both approaches may be used to study responses to policy shocks (see Mount, 1989; Todd, 1989). The traditional approaches

[2] This section is based on the discussion in Jennings, et al, 1991.

[3] Where they are considered to be important, exogenous (or deterministic) variables may be included in the set of independent variables in the system.

tend to place little emphasis on lags in equations while the VAR approach emphasizes it. The traditional approach places strict interpretations on the parameters of each equation while the VAR approach interprets the system as a whole and the analyzes responses to the behavioral shocks (see Orden and Fackler, 1989).

The VAR approach begins with a dynamic equation system of the form

[4] $$\sum_{s=0}^{\infty} A(s) \quad Y(t-s) = \sum_{s=0}^{\infty} v(t-s)$$

where Y(t) and v(t) are k × 1 vectors and A(s) is a k × k matrix of coefficients for each time period (s) previous to current time (t). The model in (4) relates the observable data (Y) to sources of variation in the economy (v). The shocks in v(t) are assumed to "represent behaviorally distinct sources of variation that drive the economy over time" (Orden and Fackler, 1989, p 496). The vector v(t) has an expected value of zero and an assumed diagonal covariance matrix, Ω. The covariance matrix is assumed to be diagonal so that individual shocks (v(t)) apply to only one behavioral equation at a time. Thus we can evaluate the effect of shocks to each behavioral equation on each variable in the system.

Assuming that errors from previous lags do not affect the current values, equation (1) can be rewritten in autoregressive form as

[5] $$A(0)Y(t) = -\sum_{s=1}^{\infty} A(s)Y(t-s) + v(t)$$

The matrix A(0) is the set of contemporaneous parameters on Y(t). Multiplying through by A(0) inverse yields

[6] $$Y(t) = \sum_{s=1}^{\infty} D(s)Y(t-s) + u(t)$$

where $D(s) = -A(0)^{-1}A(s)$ and $u(t)=A(0)^{-1}v(t)$. The vector u(t) is the one step ahead prediction error in Y(t) and the covariance matrix of u(t) is Σ. Equation (6) is the autoregressive equation which is estimated given an assumption on the lag length. It is the reduced form model.

In attempting to identify the effect of a shock to a behavioral equation on the variables in the system we can use the coefficients estimated in (6) and the observed error to simulate the impact. Since all the variables are related in the system it is not possible to "untangle" the effects of one variable on another using the autoregressive representation. However, the autoregressive representation can be used to find the moving average representation which expresses the level of a particular variable as a function of the error process. From the moving average process the impact of the behavioral shocks in each equation on each other variable can be identified.

The moving average representation of (6) can be written as

[7] $$Y(t) = \sum_{s=0}^{\infty} C(s) A(0)^{-1} v(t)$$

This is the Impulse Response Function (IRF) which describes the effect of shocks to the behavioral relations on variables in the system[4]. The matrix A(0) contains the information required for the identification of the model. Restrictions on A(0) are analgous to restrictions on coefficients in structural modeling (see Orden and Fackler for a description of the identification of VAR models). The IRF summarizes the dynamic multipliers as implied by our identification. A shock may be represented by the placement of the value unity in one element of the vector v(t). The IRF provides the response of all variables in the system to this unit shock. The interpretation of these shocks is analgous to the interpretation of coefficients in a structural model.

Finally, much of the appeal of VAR modeling lies in the fact that restrictions on the parameters of reduced form do not need to be specified a priori. Often, however, unrestricted VAR models suffer from overparameterization, resulting in estimates which reflect purely random fluctuations in the data and not the systematic variation which we are interested in identifying. Consequently, estimated variances will be too large and will produce models with poor forecasting performance. One way to handle this problem is to use Bayesian prior estimation in which stochastic restrictions, in the form of prior distribution weights, are applied to VAR parameters (see Sims, 1986). Bayesian techniques are used to assign weights to certain lags in the system. For example, financial data often exhibit random walks, or coefficients near unity on the first lag and zeros elsewhere. A Bayesian scheme would impose a prior mean of unity on all first lags of the dependent variable and a prior mean of zero on all other lags. Mixed estimation is used to impose this prior information (Litterman, 1986).

The VAR approach incorporates the theoretical elements of traditional econometrics and the time series elements of ARIMA modeling. Other advantages of the VAR include the fact that model development and estimation is relatively inexpensive and the forecasting accuracy is quite good. Litterman (1986) compares forecasts from several large econometric models, simple time series models and a Bayesian VAR

[4] See Judge et al., 1988, p. 771-775 for an illustration of the derivation and use of impulse response functions or innovation accounting.

model. The results of a forecasting experiment with these models is presented in Table 1. The Bayesian VAR outperforms the other models in Real GNP Growth and Inflation prediction for most periods (as measured using mean squared error). The VAR model is also respectable in forecasts of Unemployment.

The forecasting results of VAR models and the fact that policy analysis can be performed with them makes these tools a viable alternative to structural econometric models and simple time series models. While a variety of other multivariate time series models exist (Aoki, 1990) the VAR approach is relatively simple to estimate, relying primarily on OLS regression results, and relatively inexpensive to forecast with. The application of Bayesian priors to VAR analysis allows for individual researchers to factor in their own beliefs in a systematic manner. Such beliefs have typically been imposed in an *ad hoc* manner in structural models.

Conclusions

This paper has described three approaches to time series modeling in econometrics. Traditional multivariate structural modeling, ARIMA modeling and VAR modeling have been outlined as competing techniques. The paper has concentrated on econometric examples but these models are applicable to a wide range of topics, including those in the physical sciences. This discussion has also been centered around the ability of these models to predict well and to fit with existing theory. The identification problem, or the "smoke" caused by non-experimental data, requires careful theoretical modeling as well as proper statistical analysis. Traditional econometrics has used theory to provide the functional relationships and then attempted to conquer the statistical modeling problem. Time series models place more emphasis on the statistical process with little explicit recognition of economic theory. VAR approaches try to combine theoretical principles with a time series statistical component. However, they address the identification problem after the estimation of a statistical reduced form model. While the VAR results seem promising one must emphasize that VARs are really an alternative rather than a successor to structural modeling. The VAR technique is another kind of fan used to blow away the statistical "smoke".

Table 1. Mean Squared Error Forecasts 1976:1 - 1979:4

	Forecast Horizon: Quarters Ahead				
Variable	2	3	4	5	6
Real GNP Growth					
DRI§	2.726	2.801	2.951	3.388	3.566
ARIMA	2.882	3.071	3.076	3.181	3.209
Univariate AR	3.192	3.401	3.405	3.656	3.391
Bayesian VAR	2.841	3.053	2.948	2.959	3.021
Inflation					
DRI§	1.605	1.929	2.277	2.894	2.727
ARIMA	1.674	1.907	1.755	2.211	2.327
Univariate AR	2.289	2.735	3.111	3.526	3.940
Bayesian VAR	1.426	1.624	1.441	1.710	1.640
Unemployment					
DRI§	.341	.449	.485	.494	.430
ARIMA	.466	.712	.915	1.073	1.236
Univariate AR	.362	.493	.541	.566	.576
Bayesian VAR	.383	.497	.559	.627	.738

§ DRI structural econometric models forecasts.

Source: Litterman, 1986

References

Aoki, M. 1987. State space modeling of time series. Springer-Verlag. New York.

Jennings, S., W.L. Adamowicz and L. Constantino. 1991. Macroeconomic impacts on Canada's lumber sector. in press Canadian Journal of Forest Research. March 1991.

Judge, G., Hill, R., Griffiths, W., Lutkepohl, H. and Lee, T. 1988. Introduction to the theory and practice of econometrics. 2nd Edition, John Wiley & Sons. New York.

Litterman, R.B. 1986. Forecasting with bayesian vector autoregressions - Five years of experience. Journal of Business and Economic Statistics 4:25-38.

Lucas, R.E. 1976. Econometric policy evaluation: a critique. The Phillips curve and labor markets. K. Brunner and A. Metzler editors. pp. 19-46. North Holland. Amsterdam.

Mount, T.D. 1989. Policy analysis with time-series econometric models:discussion. American Journal of Agricultural Economics. 71:507-508.

Orden, D. and P.L. Fackler. 1989. Identifying monetary impacts on agricultural prices in VAR models. American Journal of Agricultural Economics. 71:495-502.

Sims, C.A. 1980. Macroeconomics and reality. Econometrica. 48:1-48.

Sims, C.A. 1986. Are forecasting models useful for policy analysis? Federal Reserve Bank of Minnesota Quarterly Review. Winter:2-16.

Spanos, A. 1990. The simultaneous equations model revisited. Journal of Econometrics. 44:87-105.

Todd, R.M. 1989. Policy analysis with time-series econometric models:discussion. American Journal of Agricultural Economics. 71:509-510.

SOME MULTIVARIATE METHODS
FOR
CATEGORICAL DATA

J.D. Jobson
Professor, Faculty of Business
University of Alberta
Edmonton, Alberta

Taken from : *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*, by J.D. Jobson; to be published by Springer-Verlag.

ABSTRACT

This paper contains two parts. The first part summarizes the methodology for fitting loglinear models to three-dimensional contingency tables using the method of maximum likelihood. An example based on traffic accident data is used to illustrate the techniques. The second part of the paper outlines the logistic regression model for a dichotomous dependent variable. An example based on female labor force participation data is used to demonstrate the methodology.

OUTLINE

1. Loglinear Models for Three-Dimensional Contingency Tables

The three-dimensional contingency table arises from the cross-classification of the categories associated with three qualitative random variables. Geometrically the table may be viewed as having rows, columns and layers. The subscripts for the rows, columns and layers will be denoted by i, j and k respectively. The number of rows, columns and layers will be denoted by r, c and ℓ respectively. The probability density for cell (i, j, k) will be denoted by f_{ijk} and the theoretical cell frequency by $F_{ijk} = nf_{ijk}$ for a total table frequency of n. The allocation of a sample of size n to the total of $rc\ell$ cells yields cell frequencies n_{ijk}. Table 1 shows the n_{ijk} for a sample of size n.

Various marginal totals will be denoted using dots to indicate which subscripts have been summed. For the three possible two-dimensional tables, the cell frequencies are denoted by the marginals $n_{ij.}$, $n_{i.k}$ and $n_{.jk}$. For each of the three variables the univariate marginals are given by $n_{i..}$, $n_{.j.}$ and $n_{..k}$.

Table 1. A Three Dimensional Contingency Table

		Columns			
Layers	Rows	1	2	…	c
1	1	n_{111}	n_{121}	…	n_{1c1}
	2	n_{211}	n_{221}	…	n_{2c1}
	⋮	⋮	⋮		⋮
	r	n_{r11}	n_{r21}	…	n_{rc1}
2	1	n_{112}	n_{122}	…	n_{1c2}
	2	n_{212}	n_{222}	…	n_{2c2}
	⋮	⋮	⋮		⋮
	r	n_{r12}	n_{r22}	…	n_{rc2}
	⋮	⋮	⋮		⋮
ℓ	1	$n_{11\ell}$	$n_{12\ell}$	…	$n_{1c\ell}$
	2	$n_{21\ell}$	$n_{22\ell}$	…	$n_{2c\ell}$
	⋮	⋮	⋮		⋮
	r	$n_{r1\ell}$	$n_{r2\ell}$	…	$n_{rc\ell}$

Table 2. Frequency Table – Driver Injury Level vs. Seatbelt Usage and Driver Condition

		Driver Injury Level				
Driver Condition	Seatbelt Usage	None	Minimal	Minor	Major/ Fatal	Totals
Normal	Yes	12500 (11817.8)	604 (697.1)	344 (450.2)	38 (51.8)	13486
	No	61971 (62161.0)	3519 (3666.9)	2272 (2368.0)	237 (272.2)	67999
	Totals	74471	4123	2616	275	81485
Been Drinking	Yes	313 (766.3)	43 (45.2)	15 (29.2)	4 (3.4)	375
	No	3992 (4030.9)	481 (283.0)	370 (153.6)	66 (17.7)	4909
	Totals	4305	524	385	70	5284
Totals Both Conditions		78776	4647	3001	345	86769

Example

An example of a three-dimensional table is presented in Table 2. The three-way table shows the relationships between extent of injury, seatbelt usage and driver condition for a sample of 86,769 auto accidents.

Models for Three-Way Tables

We begin here with the independence model. The independence model requires that the joint density f_{ijk} in cell (i, j, k) be equal to the product of the three univariate marginal densities $f_{ijk} = f_{i..}f_{.j.}f_{..k}$. The theoretical frequency for a total frequency of n is given by

$$F_{ijk} = nf_{ijk} = F_{i..}F_{.j.}F_{..k}/n^2$$

where $F_{i..} = nf_{i..}$, $F_{.j.} = nf_{.j.}$ and $F_{..k} = nf_{..k}$.

Inference for the Independence Model

Given a sample of size n, the maximum likelihood estimators of the expected cell frequencies under the independence assumption are given by

$$E_{ijk} = n_{i..}n_{.j.}n_{..k}/n^2 \quad \begin{aligned} i &= 1,2,\ldots,r; \\ j &= 1,2,\ldots,c; \\ k &= 1,2,\ldots,\ell. \end{aligned}$$

The fitted cell frequencies depend only on the row, column and layer marginals. Using the estimated expected frequencies E_{ijk}, the χ^2 tests of goodness of fit for the independence model are carried out using either of two statistics

$$G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell} \frac{(E_{ijk} - n_{ijk})^2}{E_{ijk}}$$

or

$$L^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell} n_{ijk}[\ln n_{ijk} - \ln E_{ijk}],$$

both of which are asymptotically χ^2 with $(rc\ell - r - \ell - c + 2)$ degrees of freedom if the independence hypothesis holds. The G^2 statistic is the Pearson χ^2 statistic, while L^2 is derived from a likelihood ratio test.

Example

The χ^2 test of independence for Table 2 yields 1057.47 and 939.90 for the Pearson and likelihood ratio statistics respectively. Both of these χ^2 statistics have 10 degrees of freedom and are significant at the 0.000 level. The expected frequencies under the independence model are shown in Table 2 in brackets. A comparison of the observed and expected frequencies permits us to conclude the following:

(a) For seatbelt users who appeared normal, the number of accidents resulting in no injury was larger than expected, while the number who sustained any injury was smaller than expected under independence.

(b) For seatbelt users who had been drinking, the number of accidents resulting in no injury was less than half the number expected under independence. In the minor injury category there were fewer cases than expected.

(c) For non-users of seatbelts who appeared normal, the number of accidents in all injury categories was less than expected under independence.

(d) For non-users of seatbelts who had been drinking, the number of accidents resulting in no injury was less than expected. For the three injury categories, the number of accidents was much larger than expected under independence.

Drivers who wore seatbelts and appeared normal sustained fewer injuries than expected, while drivers who did not wear seatbelts and had been drinking suffered more injuries than expected under independence. For the remaining two categories, the difference between the observed and expected frequencies seems less obvious. A loglinear model representation for this table will be used below to identify the interactions among the three variables. Before attempting to model the variation in the table, a discussion of various model types is required.

For the remainder of this discussion the sampling model assumed is either multinomial or independent Poisson. The two distributions are equivalent if the sample size n is fixed. Because the product multinomial places additional restrictions on some marginals, additional requirements must be adhered to in order to obtain maximum likelihood estimates.

Partial Independence

Since there are three variables in the table, it is possible to have two variables related to each other but both be independent of a third variable. This model is called the ***partial independence model*** and is given by

$$f_{ijk} = (f_{ij.})(f_{..k}).$$

In this case, the third variable with subscript k is independent of the remaining two variables with subscripts i and j. The theoretical frequency is given by

$$F_{ijk} = F_{ij.}F_{..k}/n$$

and is estimated by

$$E_{ijk} = n_{ij.}n_{..k}/n.$$

The two-dimensional marginals $n_{ij.}$ are being fitted since $E_{ij.} = n_{ij.}$. The χ^2 goodness of fit statistic in this case has $(rc-1)(\ell-1)$ degrees of freedom.

An example of a partial independence relationship would exist if in Table 2 seatbelt usage were independent of both driver condition and driver injury level, but at the same time driver condition and injury level were related.

Conditional Independence

A ***conditional independence model*** permits two variables to be independent after controlling for a third variable. An example of such a model is provided by

$$f_{ijk} = f_{i.k}f_{.jk}/f_{..k}$$

where the variables with subscripts i and j are independent at every level of the variable with subscript k. The theoretical frequency is given by

$$F_{ijk} = F_{i.k}F_{.jk}/F_{..k}$$

and the maximum likelihood estimator is given by

$$E_{ijk} = n_{i.k}n_{.jk}/n_{..k}.$$

For this model the two-dimensional marginals $n_{i.k}$ and $n_{.jk}$ are being fitted since $E_{i.k} = n_{i.k}$ and $E_{.jk} = n_{.jk}$. The χ^2 goodness of fit statistic has $\ell(r-1)(c-1)$ degrees of freedom.

An example of a conditional independence model in Table 2 would occur if, for each of the two driver conditions, driver injury level is independent of seatbelt usage. In this case driver injury level is related to seatbelt usage, but if driver condition is held fixed then seatbelt usage and driver injury level are independent. In other words, any relationship between driver injury level and seatbelt usage is due to the relation between driver condition and the other two variables. This result is similar to obtaining a zero first-order partial correlation coefficient with three quantitative variables.

No Three-Way Interaction

The next step in moving to less restrictive models is to assume that each pair of variables is related, but that the relation between any pair of variables does not depend on the level of the third. This model is usually referred to as the ***no three-way interaction model.*** It is not possible to given an expression for f_{ijk} or for F_{ijk} that would permit us to determine the estimators E_{ijk} directly. For this model the E_{ijk} are obtained by a procedure known as ***iterative proportional fitting.***

Since the model to be fitted assumes that all possible pairs are related but that there is no three-way interaction, we need only fit a model which preserves the three two-dimensional marginal totals $n_{ij.}$, $n_{.jk}$ and $n_{i.k}$. The steps for iterative proportional fitting proceed as follows:

STEP 1: Compute the observed marginal totals $n_{ij.}$, $n_{.jk}$, $n_{i.k}$.

STEP 2: Assign the initial value 1 to every estimated cell frequency i.e., $E^{(0)}_{ijk} = 1$, for all i, j, k

STEP 3: Compute new estimates of the E_{ijk} so that they sum to the marginal totals $n_{ij.}$ using

$$E_{ijk}^{(1)} = E_{ijk}^{(0)} \left[\frac{n_{ij.}}{E_{ij.}^{(0)}}\right] \quad \text{for all } i, j, k.$$

STEP 4: Compute new estimates of the E_{ijk} so that they sum to the marginal totals $n_{i.k}$ using

$$E_{ijk}^{(2)} = E_{ijk}^{(1)} \left[\frac{n_{i.k}}{E_{i.k}^{(1)}}\right] \quad \text{for all } i, j, k.$$

STEP 5: Compute new estimates of the E_{ijk} so that they sum to the marginal totals $n_{.jk}$ using

$$E_{ijk}^{(3)} = E_{ijk}^{(2)} \left[\frac{n_{.jk}}{E_{.jk}^{(2)}}\right] \quad \text{for all } i, j, k.$$

STEP 6 and subsequent steps — repeat the cycle given by Steps 3, 4 and 5 until the changes in the E_{ijk} are smaller than some preassigned number.

For the fitted model the three two-dimensional marginals $E_{ij.}$, $E_{.jk}$ and $E_{i.k}$ will be very close to their observed counterparts $n_{ij.}$, $n_{.jk}$ and $n_{i.k}$. The number of degrees of freedom for a χ^2 goodness of fit test would in this case be $(r-1)(k-1)(c-1)$.

A no three-way interaction model implies that the interaction between any pair does not depend on the third variable. For the data in Table 2 a no three-way interaction model would imply that the interaction between seatbelt usage and driver injury level does not depend on driver condition. Similarly the interaction between driver injury level and driver condition does not depend on seatbelt usage, and the interaction between seatbelt usage and driver condition does not depend on driver injury level.

Saturated Model

As in the case of the two-way contingency table, the most general model for the three-way contingency table is the saturated model that fits the data perfectly. The *saturated model* for the three-way table includes a three-way interaction which allows the two-way interaction between any pair to vary at each level of the third variable. This model will be discussed further with the introduction of the loglinear model for three-way tables below.

Loglinear Models for Three-way Tables

The saturated model for the three-way table is given by

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)},$$

$$i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, c; \quad k = 1, 2, \ldots, \ell$$

where F_{ijk} = true frequency in cell (i, j, k) and

$$\mu = 1/rc\ell \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{\ell} \ln F_{ijk},$$

$$\mu_{1(i)} = 1/c\ell \sum_{j=1}^{c} \sum_{k=1}^{\ell} \ln F_{ijk} - \mu,$$

$$\mu_{2(j)} = 1/r\ell \sum_{i=1}^{r} \sum_{k=1}^{\ell} \ln F_{ijk} - \mu,$$

$$\mu_{3(k)} = 1/rc \sum_{i=1}^{r} \sum_{j=1}^{c} \ln F_{ijk} - \mu,$$

$$\mu_{12(ij)} = 1/\ell \sum_{k=1}^{\ell} \ln F_{ijk} - \mu_{1(i)} - \mu_{2(j)} - \mu,$$

$$\mu_{13(ik)} = 1/c \sum_{j=1}^{c} \ln F_{ijk} - \mu_{1(i)} - \mu_{3(k)} - \mu,$$

$$\mu_{23(jk)} = 1/r \sum_{i=1}^{r} \ln F_{ijk} - \mu_{2(j)} - \mu_{3(k)} - \mu,$$

$$\mu_{123(ijk)} = \ln F_{ijk} - \mu_{1(i)} - \mu_{2(j)} - \mu_{3(k)} - \mu_{12(ij)},$$

$$- \mu_{23(jk)} - \mu_{13(ik)} - \mu.$$

The following conditions follow from these definitions

$$\sum_{i=1}^{r}\mu_{1(i)} = \sum_{j=1}^{c}\mu_{2(j)} = \sum_{k=1}^{\ell}\mu_{3(k)} = 0,$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{12(ij)} = \sum_{i=1}^{r}\sum_{k=1}^{\ell}\mu_{13(ik)} = \sum_{j=1}^{c}\sum_{k=1}^{\ell}\mu_{23(jk)} = 0,$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell}\mu_{123(ijk)} = 0.$$

The μ parameters are functions of various marginal totals in the table of logarithms of the theoretical frequencies, $\ln F_{ijk}$. The μ parameters are functions of the logarithms of various geometric means of the frequencies. The expressions for the μ parameters may also be written as $\mu = \ln \widetilde{F}_{...}$,

$$\mu_{1(i)} = \ln \widetilde{F}_{i..} - \ln \widetilde{F}_{...},$$

$$\mu_{2(j)} = \ln \widetilde{F}_{.j.} - \ln \widetilde{F}_{...},$$

$$\mu_{3(k)} = \ln \widetilde{F}_{..k} - \ln \widetilde{F}_{...},$$

$$\mu_{12(ij)} = \ln \widetilde{F}_{ij.} - \ln \widetilde{F}_{i..} - \ln \widetilde{F}_{.j.} + \ln \widetilde{F}_{...},$$

$$\mu_{13(ik)} = \ln \widetilde{F}_{i.k} - \ln \widetilde{F}_{i..} - \ln \widetilde{F}_{..k} + \ln \widetilde{F}_{...},$$

$$\mu_{23(jk)} = \ln \widetilde{F}_{.jk} - \ln \widetilde{F}_{.j.} - \ln \widetilde{F}_{..k} + \ln \widetilde{F}_{...},$$

$$\begin{aligned}\mu_{123(ijk)} &= \ln \widetilde{F}_{ijk} - \ln \widetilde{F}_{ij.} - \ln \widetilde{F}_{i.k} - \ln \widetilde{F}_{.jk},\\ &\quad + \ln \widetilde{F}_{i..} + \ln \widetilde{F}_{.j.} + \ln \widetilde{F}_{..k} - \ln \widetilde{F}_{...},\end{aligned}$$

where

$\widetilde{F}_{...}$ is the overall geometric mean of all the frequencies F_{ijk};

$\widetilde{F}_{i..}$ is the geometric mean of all the frequencies F_{ijk} holding i fixed;

$\widetilde{F}_{.j.}$ is the geometric mean of all the frequencies F_{ijk} holding j fixed;

$\widetilde{F}_{..k}$ is the geometric mean of all the frequencies F_{ijk} holding k fixed;

$\widetilde{F}_{ij.}$ is the geometric mean of all the frequencies F_{ijk} holding i, j fixed;

$\widetilde{F}_{.jk}$ is the geometric mean of all the frequencies F_{ijk} holding j, k fixed;

$\widetilde{F}_{i.k}$ is the geometric mean of all the frequencies F_{ijk} holding i, k fixed.

For each of the models introduced above for three-way tables the cell frequencies F_{ijk} have different properties. These properties imply that some of the μ parameters are zero.

Independence Model

In the case of the independence model, $F_{ijk} = \frac{F_{i..}F_{.j.}F_{..k}}{n^2}$ implies that $\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)}$ with all remaining μ parameters zero.

Partial Independence Model

For the partial independence model the two-way interaction between i and j results in $\mu_{12(ij)}$ being non-zero. The other possible interactions are zero. The loglinear model for this particular partial independence model is therefore given by

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)}.$$

If the table is collapsed over k, the resulting two-dimensional table is fitted exactly.

Conditional Independence Model

In the conditional independence model, the relationship between i and k is captured by $\mu_{13(ik)}$, and the relationship between j and k is captured by $\mu_{23(jk)}$. Since i and j are independent at every level of k, $\mu_{12(ij)} = 0$. The loglinear model in this case is

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)} + \mu_{23(jk)}.$$

If the table is collapsed over i or over j, the resulting two-dimensional tables are fitted exactly.

No Three-way Interaction Model

In the no three-way interaction model all pairs are related, but these relationships are independent of the third variable. Only the term $\mu_{123(ijk)}$ is zero. The loglinear model is given by

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}.$$

In this case the three two-dimensional tables obtained by collapsing the fitted table on the third variable have cell frequencies identical to the observed two-dimensional tables.

Saturated Model

The saturated model given at the beginning of this section fits the three-dimensional table perfectly. While this model is not needed to determine expected frequencies, it is often useful for characterizing the interactions in a three-way table.

Multiplicative Form of the Loglinear Model

Taking the anti-logarithm of both sides of the loglinear model yields a multiplicative model for the cell frequency F_{ijk}. The equation becomes

$$F_{ijk} = \beta_0 \beta_{1(i)} \beta_{2(j)} \beta_{3(k)} \beta_{12(ij)} \beta_{13(ik)} \beta_{23(jk)} \beta_{123(ijk)}.$$

The beta parameters are sometimes useful for characterizing the variation in the table. The beta parameters are defined by

$$\beta_0 = e^{\mu}, \quad \beta_{1(i)} = e^{\mu_{1(i)}}, \quad \beta_{2(j)} = e^{\mu_{2(j)}}, \quad \beta_{3(k)} = e^{\mu_{3(k)}},$$
$$\beta_{12(ij)} = e^{\mu_{12(ij)}}, \quad \beta_{13(ik)} = e^{\mu_{13(ik)}}, \quad \beta_{23(jk)} = e^{\mu_{23(jk)}},$$
$$\beta_{123(ijk)} = e^{\mu_{123(ijk)}}.$$

Hierarchical Models

The above collection of models does not include all possible variants using the parameters specified by the saturated model. Such models as

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)} + \mu_{23(jk)} + \mu_{13(ik)}$$

and

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{123(ijk)}$$

have not been considered. In order to maintain the practice of defining higher order terms using deviations of lower order terms, the ***hierarchy principle*** is followed. This principle requires that, if a given term is fitted, all lower order terms involving those variables must also be included. The main difficulty with ***non-hierarchical*** models is the interpretation of the fitted parameters. An additional problem, however, is that the iterative proportional fitting procedure cannot be used to fit the model without some transformation of the model being carried out first.

Notation for Loglinear Models

To simplify the notation for the remainder of this chapter, the various models in the hierarchical system will be denoted by symbols such as [1], [23] and [134]. Only the symbols for the highest order interaction for each variable will be used. All lower order terms containing that variable are automatically included in the hierarchical system. The model [12], [234], for instance, implies that the terms [23], [24] and [34] are also present, while the parameters corresponding to [13] and [14] are not present.

Model Selection

Given a three-dimensional table of observed cell frequencies n_{ijk}, a variety of models in the hierarchical system can be fitted by replacing F_{ijk} by E_{ijk} in the above formulae for the loglinear model parameters. The expression for E_{ijk} depends on the model being fitted. The various formulae for E_{ijk} for the various models have been outlined above. The goodness of fit of a particular model can be judged using the χ^2 goodness of fit statistics G^2 and L^2. A probability level of 0.15 to 0.25 is usually required to confirm that the model adequately represents the interactions in the table. In practice we seek to fit the simplest model while maintaining a reasonable fit.

In addition to the overall measure of goodness of fit, the likelihood ratio statistic L^2 has the advantage that it can be used to compare nested models in the hierarchical system. Let L_1^2 and L_2^2 denote two likelihood chi-square statistics for two alternative models and assume that model 2 is the larger model which contains all the parameters of model 1. The conditional likelihood chi-square statistic $L_{2.1}^2 = (L_1^2 - L_2^2)$ can be used to determine whether model 2 is superior to model 1. Under the null hypothesis that model 1 is equally as good as model 2, the statistic $L_{2.1}^2$ is asymptotically a χ^2 distribution with degrees of freedom equal to the difference (d.f. model 1 - d.f. model 2). An example of such a test might involve a comparison of the model [13] [2] to the model [12] [13] [23]. The null hypothesis would be that the terms [12] and [23] are superfluous.

Summary of Loglinear Model Fitting Procedure

The system of fitting loglinear models for the purpose of explaining interaction in a multidimensional contingency table can be demonstrated by the diagram in Figure 1.

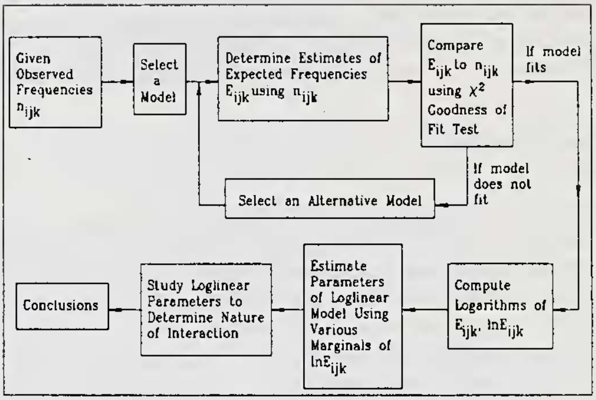

Given Observed Frequencies n_{ijk}
Select a Model
Determine Estimates of Expected Frequencies E_{ijk} using n_{ijk}
Compare E_{ijk} to n_{ijk} using χ^2 Goodness of Fit Test
If model fits
If model does not fit
Select an Alternative Model
Conclusions
Study Loglinear Parameters to Determine Nature of Interaction
Estimate Parameters of Loglinear Model Using Various Marginals of $\ln E_{ijk}$
Compute Logarithms of E_{ijk}, $\ln E_{ijk}$

Figure 1. System for Fitting and Using Loglinear Models

Product Multinomial Sampling

In product multinomial sampling, certain marginals are held fixed. In the three dimensional table we consider the two cases corresponding to the fixing of the marginals for one or two of the three variables. If the marginals are fixed for the first variable, then the loglinear model must contain the term $\mu_{1(i)}$. This will ensure that the fitted marginals $E_{i..}$ are equal to the observed marginals $n_{i..}$. Similarly, if the marginals for both variables 1 and 2 are fixed, then the model must contain the parameters $\mu_{1(i)}$, $\mu_{2(j)}$ and $\mu_{12(ij)}$. In this case the fitted marginals $E_{i..}$, $E_{.j.}$ and $E_{ij.}$ are equivalent to the sample marginals $n_{i..}$, $n_{.j.}$ and $n_{ij.}$.

In product multinomial sampling some of the variables can be viewed as response variables, while the remainder can be viewed as fixed or controlled. The control variables have the fixed marginals, while the marginals for the response variables are viewed as an outcome of the sampling process. The weighted least squares approach assumes product multinomial sampling.

Example

For the example presented in Table 2, the entire set of loglinear models were fitted using the maximum likelihood estimators E_{ijk}. Table 3 summarizes the χ^2 goodness of fit statistics for the various models. The first row is the independence model which permits all three marginals to vary but contains no interaction.

Rows 2, 3 and 4 show the results for the fitting of the three possible partial independence models. In row 2 the model [2], [13] allows variables 1 and 3 to be related, but both are assumed to be independent of variable 2. Similarly, in row 3 variables 2 and 3 are independent of 1, and in row 4 variables 1 and 2 are independent of variable 3.

Table 3. Summary of χ^2 Goodness of Fit Statistics for System of Hierarchical Models

[2] = seatbelt usage, [1] = driver condition, [3] = injury level

	Model	d.f.	Likelihood	Prob	Pearson	Prob
1.	[1], [2], [3]	10	940.02	0.0000	1057.47	0.0000
2.	[2], [13]	7	444.85	0.0000	372.21	0.0000
3.	[1], [23]	7	877.16	0.0000	967.92	0.0000
4.	[3], [12]	9	542.50	0.0000	682.37	0.0000
5.	[12], [23]	6	479.69	0.0000	610.75	0.0000
6.	[13], [23]	4	382.02	0.0000	317.32	0.0000
7.	[13], [12]	6	47.34	0.0000	44.51	0.0000
8.	[12], [13], [23]	3	5.02	0.1705	5.02	0.1705

The three conditional independence models are shown in rows 5, 6 and 7. In row 5 the model [12], [23] requires that 1 and 3 be independent at each level of variable 2. Similarly, in row 6 variables 1 and 2 are independent at each level of 3, and in row 7 variables 2 and 3 are independent at each level of 1. The no three-way interaction model is fitted in the last row. In this model all two-way interactions among the three variables are assumed to explain all the interactions in the table.

An examination of the χ^2 goodness of fit statistics reveals that the no three-way interaction model can be used to explain the interactions among the three variables. The fitted parameters for this model are summarized in Table 4. The ratios of the loglinear model parameter estimates to their standard error are also shown in this table for selected parameters. Plots of the values of the parameter estimates are shown in Figure 2. The loglinear model is given by

$$\log F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}.$$

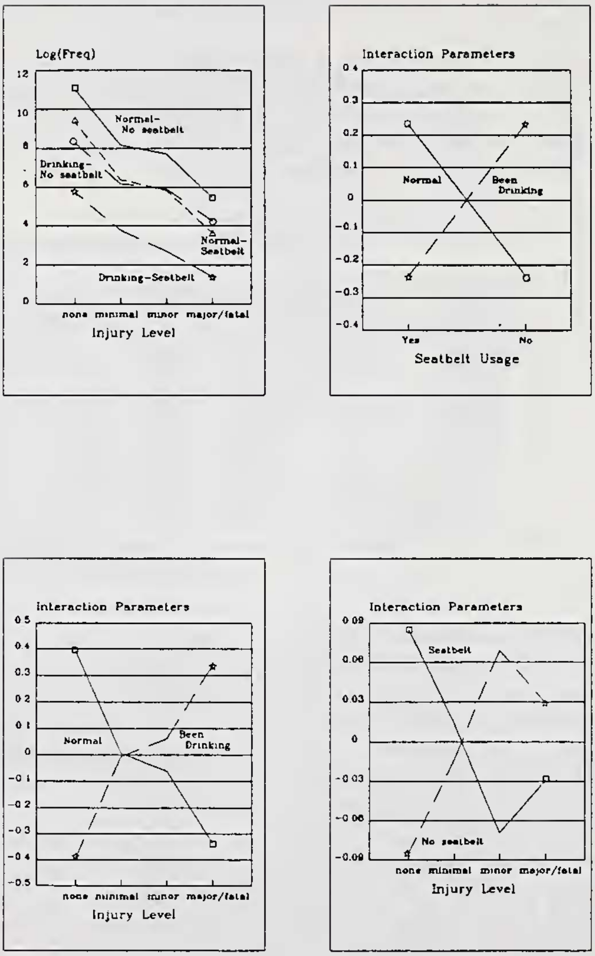


Figure 2. Parameter Estimates for Loglinear Model for Accident Data

From the fitted parameters in Table 4 the logarithm of the geometric mean of the expected frequencies is 6.002. The driver condition effects indicate that the normal condition is much more frequent than the "been drinking" condition. The seatbelt usage effects indicate that many more drivers were not wearing seatbelts than were wearing them. The injury level parameters indicate that the large majority of drivers were not injured and that very few drivers sustained major or fatal injuries.

Table 4. Fitted Parameters for No Three-Way Interaction Loglinear Model

Overall Mean		$\hat{\mu} = 6.002$	
Driver Condition Effects		$\hat{\mu}_{1(1)} = 1.212$	$\hat{\mu}_{1(2)} = -1.212$ (−53.906)
Seatbelt Usage Effects		$\hat{\mu}_{2(1)} = -1.119$	$\hat{\mu}_{2(2)} = 1.119$ (43.938)
Injury Level Effects			
$\hat{\mu}_{3(1)} = 2.626$ (97.270)	$\hat{\mu}_{3(2)} = 0.071$ (2.152)	$\hat{\mu}_{3(3)} = -0.376$ (−10.315)	$\hat{\mu}_{3(4)} = -2.322$ (−32.387)
Driver Condition – Seatbelt Usage Interaction			
$\hat{\mu}_{12(11)} = 0.234$ (17.147)	$\hat{\mu}_{12(12)} = -0.234$	$\hat{\mu}_{12(21)} = -0.234$	$\hat{\mu}_{12(22)} = 0.234$
Driver Condition – Injury Level Interaction			
$\hat{\mu}_{13(11)} = 0.392$ (19.698)	$\hat{\mu}_{13(12)} = 0.006$ (0.219)	$\hat{\mu}_{13(13)} = -0.061$ (−2.249)	$\hat{\mu}_{13(14)} = -0.337$ (−6.578)
$\hat{\mu}_{13(21)} = -0.392$	$\hat{\mu}_{13(22)} = -0.006$	$\hat{\mu}_{13(23)} = 0.061$	$\hat{\mu}_{13(24)} = 0.337$
Seatbelt Usage - Injury Level Interaction			
$\hat{\mu}_{23(11)} = 0.085$ (3.714)	$\hat{\mu}_{23(12)} = 0.013$ (0.490)	$\hat{\mu}_{23(13)} = -0.069$ (−2.286)	$\hat{\mu}_{23(14)} = -0.029$ (−0.465)
$\hat{\mu}_{23(21)} = -0.085$	$\hat{\mu}_{23(22)} = -0.013$	$\hat{\mu}_{23(23)} = 0.069$	$\hat{\mu}_{23(24)} = 0.029$

The interaction effects in Table 4 suggest that normal condition drivers were more likely to be wearing seatbelts than drivers who had been drinking. The driver condition-injury level interactions indicate that, in comparison to drivers who had been drinking, a larger proportion of drivers in the normal category had no injury and a smaller proportion of the normal category drivers were in the major or fatal injury category. For the minimal and minor injury categories, the interaction terms were quite weak. The interaction between driver injury level and driver condition therefore seems to affect only the two extremes of the injury level range. The seatbelt usage-injury level interaction appears to be relatively weak. There is some tendency, however, for seatbelt users to be over-represented in the no-injury category and under-represented in the minor injury category. The minimal injury category and the major/fatal category show only slight interactions with seatbelt usage.

In conclusion, we could say that a large majority of drivers appeared normal, had not been wearing seatbelts, and were not injured. For drivers wearing seatbelts, there were proportionately fewer who sustained an injury and proportionately more were in normal condition than for non-seatbelt users. Among those who had been drinking, proportionately more sustained a minor or major/fatal injury than among those who appeared normal.

A comparison of the observed frequencies to the expected frequencies under the no three-way interaction fitted model is shown in Table 5. The expected frequencies are shown in round brackets under the corresponding observed frequencies. The fit seems to be excellent with only minor differences in the minimal and minor categories for drivers who had been drinking and were wearing seatbelts. The values of the standardized residuals are shown in square brackets for each cell. The largest standardized residuals occurred in the minimal and minor categories for drivers who had been drinking. These residuals, however, were quite small indicating an excellent fit. In these two cells the frequencies are relatively small and hence the prediction errors are proportionately larger.

Table 5. Comparison of Observed and Expected Frequencies

		Driver Injury Level			
Driver Condition	Seatbelt Usage	None	Minimal	Minor	Major/Fatal
Normal	Yes	12500 (12497.0) [0.0]	604 (613.3) [-0.4]	344 (337.8) [0.3]	38 (37.9) [0.0]
	No	61971 (61974.0) [0.0]	3519 (3509.7) [0.2]	2272 (2278.2) [-0.1]	237 (237.1) [0.0]
Been Drinking	Yes	313 (316.0) [-0.2]	43 (33.7) [1.6]	15 (21.2) [-1.3]	4 (4.1) [-0.1]
	No	3992 (3989.0) [0.0]	481 (490.3) [-0.4]	370 (363.8) [0.3]	66 (65.9) [0.0]

The fitted parameters in Table 4 were converted to multiplicative parameters and are summarized in Table 6. The multiplicative form of the fitted model is given by the equation

$$F_{ijk} = \beta_0\beta_{1(i)}\beta_{2(j)}\beta_{3(k)}\beta_{12(ij)}\beta_{13(ik)}\beta_{23(jk)}.$$

Table 6. Multiplicative Parameters for No Three-Way Interaction Model

Geometric Mean		$\hat{\beta} = 404.362$	
Driver Condition Effects		$\hat{\beta}_{1(1)} = 3.361$	$\hat{\beta}_{1(2)} = 0.298$
Seatbelt Usage Effects		$\hat{\beta}_{2(1)} = 0.237$	$\hat{\beta}_{2(2)} = 3.061$
Injury Level Effects			
$\hat{\beta}_{3(1)} = 13.824$	$\hat{\beta}_{3(2)} = 1.074$	$\hat{\beta}_{3(3)} = 0.687$	$\hat{\beta}_{3(4)} = 0.098$
Driver Condition – Seatbelt Usage Interaction			
$\hat{\beta}_{12(11)} = 1.263$	$\hat{\beta}_{12(12)} = 0.792$	$\hat{\beta}_{12(21)} = 0.792$	$\hat{\beta}_{12(22)} = 1.263$
Driver Condition – Injury Level Interaction			
$\hat{\beta}_{13(11)} = 1.481$	$\hat{\beta}_{13(12)} = 1.006$	$\hat{\beta}_{13(13)} = 0.941$	$\hat{\beta}_{13(14)} = 0.714$
$\hat{\beta}_{13(21)} = 0.675$	$\hat{\beta}_{13(22)} = 0.994$	$\hat{\beta}_{13(23)} = 1.063$	$\hat{\beta}_{13(24)} = 1.401$
Seatbelt Usage – Injury Level Interaction			
$\hat{\beta}_{23(11)} = 1.088$	$\hat{\beta}_{23(12)} = 1.013$	$\hat{\beta}_{23(13)} = 0.934$	$\hat{\beta}_{23(14)} = 0.971$
$\hat{\beta}_{23(21)} = 0.919$	$\hat{\beta}_{23(22)} = 0.987$	$\hat{\beta}_{23(23)} = 1.071$	$\hat{\beta}_{23(24)} = 1.030$

Three-way Interaction

When a saturated model is required in order to obtain a good fit for a three-way table, the ***three-way interaction*** $\mu_{123(ijk)}$ is said to be significant. The presence of such an interaction indicates that each of the three two-way interactions cannot be assumed to be constant over the various levels of the third. As an example, consider the two-way interaction $\mu_{12(ij)}$. This parameter measures the interaction between variables 1 and 2 and is estimated using the marginal table obtained after summing over the subscript k. The two-way interaction $\mu_{12(ij)}$ therefore represents an average relationship between variables 1 and 2 summed over the categories of the third variable. The fact that $\mu_{123(ijk)}$ is non zero indicates that the interaction between variables 1 and 2 varies over the levels of variable 3.

Example

To provide an example interpretation for three-way interaction parameters, the estimates $\hat{\mu}_{123(ijk)}$ for the data in Table 2 are shown in Table 7. The largest parameter estimate of 0.086 for the category normal, seatbelt yes, and minor injury allows us to conclude the following:

(1) The proportion of individuals who sustained a minor injury while wearing seatbelts was greater for normal condition drivers than for drivers who had been drinking. In other words, the interaction between seatbelt usage and injury level is not independent of driver condition.

(2) The proportion of individuals who sustained a minor injury while in normal condition was greater for those wearing seatbelts than for those not wearing seatbelts. Thus the interaction between driver condition and injury level depends on seatbelt usage.

(3) The proportion of individuals who wore seatbelts while in normal condition was greater for those who sustained a minor injury than for those who sustained a minimal injury. The interaction between seatbelt usage and driver condition varies with the injury level.

Table 7. Threeway Interaction Terms from Saturated Model

Driver	Seatbelt	Driver Injury Level			
Condition	Usage	None	Minimal	Minor	Major
Normal	YES	−0.007	−0.080	+0.086	0.000
	NO	+0.007	+0.080	−0.086	0.000
Been Drinking	YES	+0.007	+0.080	−0.086	0.000
	NO	−0.007	−0.080	+0.086	0.000

2. Logistic Regression

In the multiple linear regression model the dependent variable Y is always assumed to have an interval scale. The explanatory variables in $\mathbf{x}$, however, can be either interval scaled or categorical. If the dependent variable is categorical, a logistic regression model can be used. The discussion here is restricted to the case of a dichotomous dependent variable.

The Point Binomial

We assume that individuals or objects can be classified into one of two mutually exclusive categories A or B, and that the probabilities associated with these two categories are p and $(1 - p)$ respectively. As an example, the categories A and B might represent the events that a business firm will or will not go bankrupt in the next year.

We define the dummy random variable Y to indicate the two categories by letting $Y = 1$ for category A and $Y = 0$ for category B. The probability density for Y is therefore given by

$$f(Y \mid p) = p^Y (1-p)^{(1-Y)}$$

which is the density of a *point binominal.*

Probability as a Function of Other Variables

To continue the example of business firms and bankruptcy, we assume that the probability of bankruptcy depends on a measure of financial health D, where D is a linear function given by $D = \beta_0 + \beta_1 X$ and where X is a measure of a company's ability to repay its debts, such as debt-equity ratio. In other words the probability of bankruptcy is a function of D and will be denoted by $p(D)$. For an individual firm i with debt-equity ratio x_i, $d_i = \beta_0 + \beta_1 x_i$ and the probability density for y_i is given by

$$f\Big(y_i \mid p(d_i)\Big) = [p(d_i)]^{y_i}[1-p(d_i)]^{(1-y_i)}.$$

For a random sample of n firms we observe $(d_1, d_2, \ldots, d_n)$, and the joint density for $(y_1, y_2, \ldots, y_n)$ is given by

$$\begin{aligned} & f(y_1, y_2, \ldots, y_n \mid p(d_1), p(d_2), \ldots, p(d_n)) \\ & \quad = [p(d_1)]^{y_1}[1-p(d_1)]^{(1-y_1)}[p(d_2)]^{y_2}[1-p(d_2)]^{(1-y_2)} \\ & \qquad \ldots [p(d_n)]^{y_n}[1-p(d_n)]^{(1-y_n)} \\ & \quad = \prod_{i=1}^{n}[p(d_i)]^{y_i}[1-p(d_i)]^{(1-y_i)}. \end{aligned}$$

Note here that the parameters β_0 and β_1 are assumed to be constant across the complete sample.

The Logit Function

To be able to relate the value y of the response variable Y to the value d of the variable D, a more specific assumption about the form of the function $p(d)$ is required. The logistic regression model assumes that $p(d)$ is given

Logistic Regression With c Explanatory Variables

The logistic regression model can be extended to include c explanatory variables. In this case it is assumed that $p = p(D)$ where $D = \beta_0 + \sum_{j=1}^{c} \beta_j X_j$ is a linear function of c explanatory variables. Thus for the bankruptcy example we assume that there are a total of c variables that are related to the probability of bankruptcy.

The logit of p is given by

$$\ln[p/(1-p)] = D = \beta_0 + \sum_{j=1}^{c} \beta_j X_j$$

which has the form of a multiple regression model. The estimation of the parameters $\beta_0, \beta_1, \ldots, \beta_c$ is usually obtained using maximum likelihood, which must be determined using Newton–Raphson procedures.

Inference for the Dichotomous Logistic Regression Model

The dichotomous logistic regression model assumes that the logit function $\ln[p/(1-p)]$ can be modelled as a linear function of a set of explanatory variables $\beta_0 + \sum_{j=1}^{c} X_j \beta_j$. Given a random sample of observations $(y_i, x_{i1}, x_{i2}, \ldots, x_{ic})$, $i = 1, 2, \ldots, n$, the maximum likelihood estimator of the coefficient can be obtained as outlined. In comparison to the multiple linear regression model, the coefficients in this case must be interpreted differently. A marginal one unit increase in X_j brings about an increase in $\ln[p/(1-p)]$ of the amount $\tilde{\beta}_j$. The magnitude of the increase in p, however, depends on the initial value of p.

Comparing Nested Models

Inferences regarding the coefficients in the logistic regression model can be made by comparing models and sub models using a likelihood ratio test. To compare a full model with c explanatory variables plus an intercept to a reduced model with $(c-q)$ explanatory variables plus an intercept, the logarithm of the likelihood ratio yields the statistic $-2[\ln L_0 - \ln L]$, which has a χ^2 distribution with q degrees of freedom if the q deleted variables are superfluous. L is the likelihood function for the full model, while L_0 is the likelihood function for the reduced model. The reader may recall that this approach was also used for loglinear models in the three-dimensional contingency table discussed above.

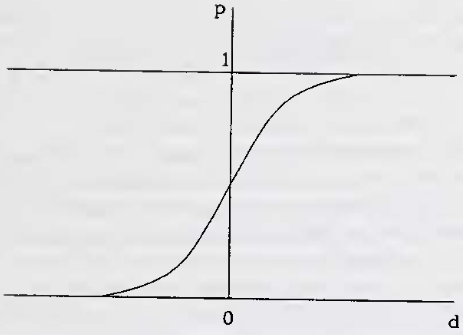


Figure 3. Shape of Logistic Distribution Functions

by the distribution function for a logistic density $G(d)$. Hence $p(d) = G(d)$, $-\infty \leq d \leq \infty$. The shape of $G(d)$ is illustrated in Figure 3.

The equation for $p(d)$ is given by

$$p(d) = e^d/(1 + e^d),$$

which is the *logistic distribution function.* The shape of $p(d)$ for the logistic is quite similar to the shape for a normal distribution function. As we outline next, the logistic transformation lends itself to a useful explicit functional relationship between $p(d)$ and d. The standardized logistic density is given by

$$f(w) = e^{-w}/(1 + e^{-w})^2$$

and the mean and variance of this density are 0 and $\pi^2/3$ respectively. The standard normal and standardized logistic distribution yield very similar shaped densities and distribution functions. Like the standard normal density, the standardized logistic density has a median and mode of zero and a skewness of zero. The kurtosis of the logistic density is 4.2 which indicates fatter tails than the normal which has a kurtosis of 3. The standardized logistic distribution with $w^* = w/\sqrt{\pi^2/3}$ has slightly heavier tails than the standard normal distribution.

An important advantage of the logistic distribution in this context is that the logit transformation $\ln[p/(1-p)]$ has the form

$$\ln\left[\frac{p(d)}{1-p(d)}\right] = \ln\left[\frac{e^d/(1+e^d)}{1/(1+e^d)}\right] = d.$$

Therefore, if d is assumed to be a linear function of x, $d = \alpha + \beta x$, the logit has the familiar linear model form. This logit model is usually referred as the *logistic regression model.*

Goodness of Fit

A pseudo measure of goodness of fit is given by

$$R^2 = 1 - \ln L / \ln L_0^*$$

where L_0^* denotes the likelihood function value when all variables are excluded except the constant term β_0. Thus the sample value of L_0^* is the value of L evaluated using the sample proportion for the maximum likelihood estimator of p. This R^2 measures the proportion of uncertainty in the data that is explained by the model. If the full model is a perfect indicator, then $L = 1$, $\ln L = 0$, and $R^2 = 1$. If the reduced model yields the same likelihood as the full model, then $\ln L = \ln L_0^*$ and $R^2 = 0$. In this case the explanatory variables contribute nothing to the likelihood.

An alternative measure of goodness of fit is given by

$$R_a^2 = (L^{2/n} - L_0^{*2/n})/(1 - L_0^{*2/n}).$$

Hosmer-Lemeshow Goodness of Fit Test

Since a large majority of the observations $(y_i, x_{i1}, x_{i2}, \ldots, x_{ic})$ $i = 1, 2, \ldots, n$ are unique in the sense that in general no two observations yield identical values on all variables, the fitted model cannot be evaluated using the χ^2 goodness of fit tests introduced for the contingency table. A goodness of fit test, known as Hosmer–Lemeshow, divides the range of $\hat{p}$ $[0, 1]$ into s mutually exclusive categories, and then a comparison of the observed and predicted frequencies be carried out using a χ^2 statistic. The categories can be determined by ranking the $n\hat{p}$ values and then dividing them into s equal groups or by dividing the range of p into s equal intervals.

We denote the actual frequency in group j by o_j, the predicted frequency by n_j, and the average value of $\hat{p}$ in group j by $\bar{p}_j$. The statistic $\sum_{j=1}^{s} \frac{(o_j - n_j\bar{p}_j)^2}{n_j\bar{p}_j(1 - \bar{p}_j)}$ is approximately χ^2 with $(s-2)$ degrees of freedom if the fitted logistic regression model is correct.

Example — Bivariate Relationships

To provide examples for the discussion of qualitative response regression models the data summarized in Table 8 will be used. The data represents a sample of 100 observations on married women selected from the Michigan Panel Study of Income Dynamics. The variables THISYR and LASTYR are indicator variables for whether the wife worked (=1) or did not work

(=0) in the current year and the previous year respectively. The variables CHILD1, CHILD2, and BLACK are dummy variables indicating whether the wife has children under 2 (CHILD1), children between age 2 and age 6 (CHILD2) or is BLACK respectively. Finally the three variables AGE, EDUC and HUBINC are measures of the years of age and years of education of the wife and the income of the husband, respectively. The variables THISYR and LASTYR will be used as response variables and the remaining variables will be used as explanatory variables.

To examine the bivariate relationships between THISYR and LASTYR and each of the six explanatory variables, single variable logistic regression models were estimated. The results are summarized in Table 9.

To illustrate the information contained in Table 9, we shall examine in detail the results for the interval scaled variable HUBINC and for the categorical variable CHILD1. For the variable HUBINC the fitted logistic regression model in the case of THISYR has the equation

$$\ln[p/(1-p)] = 1.6001 - 0.0675 \text{ HUBINC}$$

where p is the probability that the wife will choose to work THISYR. The log of the likelihood ratio for the model is given by $\ln L_1 = -56.982$ while the log of the likelihood ratio with HUBINC omitted is $\ln L_0 = -59.295$. The likelihood ratio χ^2 statistic is therefore $-2[\ln L_0 - \ln L_1] = 4.62$ which has a p-value of 0.0315 for a 1 d.f. χ^2. The fitted equation indicates that as HUBINC decreases the value of p increases. From Table 8 the range of HUBINC is 0 to 54.3 and hence the value of the logit varies from +1.6001 to −2.0652. The range of values for p is therefore given by

$$p = \exp[+1.6001]/(1 + \exp[+1.6001]) = 0.83$$

and

$$p = \exp[-2.0652]/(1 + \exp[-2.0652]) = 0.11.$$

For the categorical variables CHILD1, CHILD2 and BLACK, dummy variable coding was used with CHILD1 = 1, CHILD2 = 1 and BLACK = 1 indicating the presence of a child under 2, a child of age 2–6 and an individual of the black race. For the variable CHILD1 the fitted model is given by

$$\ln[p/(1-p)] = 1.1272 - 2.7366 \text{ CHILD1}.$$

The probability that the woman chooses to work therefore varies from $p = \exp[1.1272]/(1+\exp[1.1272]) = 0.76$ for CHILD1 = 0 to $p = \exp[-1.6094)/(1 + \exp[-1.6094]) = 0.17$ for CHILD1 = 1. The significance of the coefficient of CHILD1 is obtained from $-2[\ln L_0 - \ln L_1] = -2[-59.295 - (-55.006)] = 8.58$ which has a p-value of 0.0034 for a 1 d.f. χ^2.

Table 8. Full Time Work Outside the Home for Married Women

OBS	LASTYR	THISYR	CHILD1	CHILD2	BLACK	HUBINC	EDUC	AGE
1	0	0	0	0	0	4.340	12	42
2	0	0	0	1	0	13.648	12	31
3	1	1	0	1	1	4.973	10	38
4	0	1	0	0	0	8.427	12	46
5	0	1	0	0	0	18.320	18	46
6	0	1	0	1	1	7.680	10	29
7	1	1	0	1	0	5.612	12	25
8	0	0	0	1	0	13.554	12	32
9	1	0	0	0	0	5.329	12	26
10	1	1	0	0	0	10.511	12	29
11	1	1	0	0	0	10.486	12	34
12	0	1	0	0	0	14.071	16	38
13	1	1	0	0	0	9.024	12	32
14	1	1	0	1	0	14.329	12	36
15	1	1	1	0	0	5.118	18	28
16	1	1	0	0	1	3.044	12	37
17	1	1	0	0	1	2.640	7	38
18	1	1	0	0	1	2.050	7	43
19	0	0	0	1	1	6.750	12	23
20	1	0	0	0	0	3.383	12	24
21	1	1	0	0	0	6.630	12	40
22	1	1	0	0	0	7.000	12	46
23	0	1	0	0	0	8.815	12	42
24	1	1	0	0	0	3.450	12	46
25	0	0	0	0	0	12.031	12	42
26	1	1	0	0	1	6.144	12	31
27	0	0	0	1	0	11.513	12	39
28	0	1	0	1	0	12.167	12	46
29	0	0	1	0	0	9.968	16	28
30	0	0	1	0	0	5.888	12	23
31	1	1	0	0	0	10.232	12	32
32	1	1	0	0	0	8.017	12	40
33	1	1	0	0	0	11.686	12	45
34	1	0	0	1	0	28.363	12	31
35	1	1	0	0	1	4.343	7	46
36	1	1	0	0	0	10.554	12	38
37	1	1	0	1	0	2.484	10	29
38	0	0	0	0	0	5.672	12	44
39	1	1	0	0	1	13.319	18	31
40	1	1	0	0	1	7.678	18	35
41	1	1	0	0	0	7.162	12	24
42	0	0	0	0	0	7.804	12	34
43	0	1	0	1	0	13.648	16	28
44	0	0	0	1	0	9.311	12	27
45	1	1	0	0	0	27.938	12	46
46	1	1	0	1	0	6.704	12	27
47	1	1	0	0	0	7.711	12	32
48	1	1	0	0	0	8.576	16	38
49	0	1	0	1	0	7.223	16	26
50	0	0	1	0	0	11.259	16	31

Table 8. Full Time Work Outside the Home for Married Women (continued)

OBS	LASTYR	THISYR	CHILD1	CHILD2	BLACK	HUBINC	EDUC	AGE
51	0	0	0	1	0	26.063	12	30
52	1	1	0	0	0	11.776	12	42
53	1	1	0	0	0	12.793	18	46
54	1	1	0	0	0	11.080	12	44
55	1	1	0	0	0	7.074	12	31
56	1	1	0	1	0	6.679	12	36
57	0	1	0	0	0	15.868	12	45
58	1	1	0	0	0	7.972	16	42
59	0	0	1	0	1	0.000	12	29
60	1	1	0	0	0	3.030	10	43
61	1	1	0	0	0	2.970	16	27
62	1	1	0	0	0	9.305	12	40
63	1	1	0	0	0	8.125	12	30
64	1	0	0	1	1	13.033	10	29
65	1	1	0	0	1	0.000	12	39
66	1	1	0	1	1	2.781	12	30
67	1	1	0	0	1	3.010	12	35
68	0	0	0	0	0	26.056	12	40
69	0	0	0	0	0	5.795	12	46
70	1	1	0	1	0	0.000	12	36
71	1	1	0	1	0	2.639	12	28
72	1	1	0	0	0	9.087	12	24
73	1	0	0	0	0	12.312	12	34
74	1	0	0	0	0	7.325	12	33
75	1	1	0	0	0	3.517	10	26
76	1	1	0	0	0	17.140	12	35
77	1	1	0	0	0	24.054	12	40
78	1	1	0	0	1	6.144	12	42
79	0	1	0	0	1	13.211	12	34
80	0	1	0	0	0	9.309	12	45
81	1	1	0	0	0	3.135	10	40
82	1	1	0	0	0	2.935	10	45
83	1	1	0	0	0	9.607	12	41
84	0	1	0	0	0	10.629	12	44
85	1	1	0	0	0	8.207	12	24
86	1	1	0	0	0	9.772	12	42
87	1	1	0	0	0	8.955	12	46
88	1	1	0	0	0	6.204	10	46
89	0	1	0	0	1	9.378	12	32
90	0	0	0	0	0	54.281	12	45
91	1	1	0	1	0	7.525	12	31
92	0	0	1	0	0	11.504	12	32
93	0	1	0	0	0	5.763	12	42
94	0	0	0	1	0	5.683	12	32
95	0	1	0	0	0	10.937	12	40
96	1	1	0	0	0	9.361	12	45
97	0	0	0	1	0	6.342	12	35
98	1	0	0	0	0	7.160	10	31
99	1	0	0	1	0	7.788	12	31
100	1	1	0	0	1	2.402	10	25

Table 9. Bivariate Relationships Between THISYR, LASTYR and Each of the Six Explanatory Variables

			THISYR					LASTYR		
Variable	Intercept	Coefficient	Model Log Likelihood	Coefficient Chi-Square	Chi-Square Hosmer-Lemeshow	Intercept	Coefficient	Model Log Likelihood	Coefficient Chi-Square	Chi-Square Hosmer-Lemeshow
					p-Value					*p*-Value
AGE	-1.8260 (0.1151)	+0.0794 (0.0147)	-56.318	5.95	0.49	+1.1585 (0.2836)	-0.0150 (.6123)	-64.616	0.26	0.21
EDUC	+0.5050 (0.7127)	+0.0360 (0.7432)	-59.242	0.11	0.13	+2.5742 (0.0434)	-0.1585 (0.1209)	-63.542	2.41	0.19
HUBINC	+1.6001 (0.0000)	-0.0675 (0.0315)	-56.982	4.62	0.30	+1.4264 (0.0001)	-0.8545 (0.0096)	-61.394	6.70	0.71
BLACK	+0.8065 (0.0005)	+0.8675 (0.1682)	-58.346	1.90	–	0.5306 (0.0189)	0.4990 (0.3693)	-64.342	0.81	–
CHILD1	1.1272 (0.0000)	-2.7366 (0.0034)	-55.006	8.58	–	0.7577 (0.0004)	-2.3671 (0.0117)	-61.569	6.35	–
CHILD2	1.2272 (0.0000)	-0.9861 (0.0452)	-57.290	4.01	–	0.8158 (0.0007)	-0.7357 (0.1205)	-63.539	2.41	–
Log Likelihood for Constant Only			-59.295					-64.745		

An examination of Table 9 reveals that the most important variables in predicting whether a woman will choose to work THISYR are AGE, HUBINC and CHILD1. The coefficients of these variables indicate that p tends to be larger if AGE is large, HUBINC is small, CHILD1 = 0 and if CHILD2 = 0. For the variable LASTYR the most significant explanatory variables are HUBINC, CHILD1 and CHILD2. The coefficients in these three models indicate that p increases with decreasing HUBINC, and that p is larger if CHILD1 = 0 and if CHILD2 = 0.

Example - Logistic Regression With Multiple Explanatory Variables

To determine how the explanatory variables together predict p, a logistic regression model was fitted using all six explanatory variables. The fitted models for THISYR and LASTYR are given by

<u>THISYR</u>

$$\begin{aligned}\ln[p/(1-p)] = &- \underset{(0.0472)}{6.0624} - \underset{(0.0040)}{0.1079}\ \text{HUBINC} + \underset{(0.0148)}{0.4777}\ \text{EDUC} \\ &+ \underset{(0.0765)}{0.0773}\ \text{AGE} + \underset{(0.0708)}{1.5451}\ \text{BLACK} \\ &- \underset{(0.0003)}{4.5179}\ \text{CHILD1} - \underset{(0.0621)}{1.1238}\ \text{CHILD2},\end{aligned}$$

<u>LASTYR</u>

$$\begin{aligned}\ln[p/(1-p)] = &+ \underset{(0.0076)}{6.3641} - \underset{(0.0257)}{0.0799}\ \text{HUBINC} - \underset{(0.4638)}{0.0911}\ \text{EDUC} \\ &- \underset{(0.0423)}{0.0870}\ \text{AGE} - \underset{(0.8937)}{0.0879}\ \text{BLACK} \\ &- \underset{(0.0008)}{3.6948}\ \text{CHILD1} - \underset{(0.0057)}{1.6928}\ \text{CHILD2}.\end{aligned}$$

The fitted logistic regression model for THISYR indicates that at the margin the probability that a woman will choose to work increases with decreases in HUBINC, but decreases with decreases in AGE and EDUC. For the dummy variables, a woman of the black race is more likely to work, while if children are present the woman is less likely to work. For the variable LASTYR the variables HUBINC, CHILD1 and CHILD2 have the same impact as in the case of THISYR while the remaining variables are insignificant.

The log likelihoods for the two models are -44.044 and -53.314 for THISYR and LASTYR respectively. Excluding all six variables yields log likelihoods of -59.295 and -64.745. The likelihood ratio χ^2 statistics for the significance of all six variables are given by $-2[-59.295-(-44.044)] = 30.502$ and $-2[-64.745-(-53.314)] = 22.862$ which have p-values of 0.000 and 0.001 when compared to a χ^2 distribution with 6 d.f. The pseudo R^2

values are given by $[1 - 44.044/59.295] = 0.26$ and $[1 - 53.314/64.745] = 0.18$, respectively. The Hosmer–Lemeshow χ^2 p values are 0.05 and 0.46 for the variables THISYR and LASTYR respectively.

The Fitted Model and Classification

The fitted logistic regression model can be used to obtain the value of $\hat{p}_i$ for each observation by determining the value of $\ln[\hat{p}_i/(1-\hat{p}_i)] = \tilde{\beta}_0 + \sum_{j=1}^{r} \tilde{\beta}_j x_{ij}$ and then solving for $\hat{p}_i$. The value of $\hat{p}_i$ is given by $\hat{p}_i = e^{\mathbf{x}_i'\tilde{\beta}}/(1 + e^{\mathbf{x}_i'\tilde{\beta}})$. Assume that the observation is placed in the category $Y = 0$ if $\hat{p}_i < 0.50$, and otherwise the observation is placed in the category $Y = 1$. A prediction success matrix or confusion matrix in this case can be constructed as shown below.

	True Category	
	$Y = 0$	$Y = 1$
$\hat{p}_i < 0.50$	n_{00}	n_{01}
$\hat{p} \geq 0.50$	n_{10}	n_{11}

$n = (n_{00} + n_{01} + n_{10} n_{11})$

This table shows the distribution of the predictions for each of the two categories. The proportion of correctly classified observations is given by $(n_{00} + n_{11})/n$. The logistic regression model therefore provides a discriminant function which can be used to classify unknowns.

Example

It is of interest to examine the abilities of the two fitted models to predict the values of THISYR and LASTYR. If no explanatory variables are included in the model, the probabilities based on the observations are p[THISYR = 1] = 0.72 and p[LASTYR = 1] = 0.65. Prediction success tables based on these probabilities are therefore given by

	Predicted THISYR		
	0	1	
THISYR			
0	8	20	28
1	20	52	72
	28	72	

	Predicted LASTYR		
	0	1	
LASTYR			
0	12	23	35
1	23	42	65
	35	65	

We would therefore expect to predict correctly 60% of the values of THISYR and 54% of the values of LASTYR. An equal priors model would only be expected to predict 50% correctly.

To determine the predictions based on the fitted logistic models, values of $\hat{p}$ were determined for both models for all 100 observations. If $\hat{p} < 0.50$ for a particular individual, then that individual was placed in the 'not work' category; otherwise the individual was placed in the 'work' category. The prediction success tables are shown below.

	Predicted THISYR		
	0	1	
THISYR			
0	13	15	28
1	5	67	72
	18	82	

	Predicted LASTYR		
	0	1	
LASTYR			
0	14	21	35
1	8	57	65
	22	78	

For the variable THISYR the use of the fitted logistic regression model results in a correct classification for 80% of the cases, while for the variable LASTYR, use of the fitted model results in a correct classification for 71%. The increases in % correctly classified as a result of the fitted logistic regression model are 20% and 17% respectively. In other words, in the case of THISYR an additional 20 of the 100 cases were correctly classified, while for LASTYR an additional 17 of the 100 cases were correctly classified.

STATISTICAL GRAPHICS

Marion Herbut
Alberta Environmental Centre
Vegreville, Alberta

Introduction

The use of SAS/Graph[1] to produce regression analysis plots and plots with standard deviation/standard error bars is demonstrated. Examples are given for enhancing graphical output, manipulating data, and transferring graphical output to other graphics, word processing and desktop publishing software for further modification.

SAS/Graph regression analysis plots

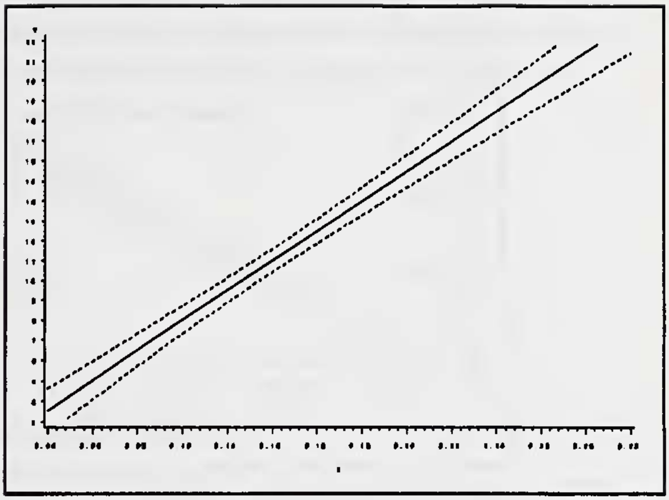
The ability of SAS/Graph to analyze data statistically and manage large data sets simplifies regression analysis. The GPLOT procedure is used to plot Y against X variables. The regression analysis is specified in the SYMBOL statement with the interpolation option (I=option). **I=RL** requests a linear regression and **I=RLØ** would set the intercept to zero.

The regression equations can be linear(RL), quadratic (RQ) or cubic (RC). Confidence limits can be added to the regression line by further specification in the I=option. **I=RLCLI99** requests a linear regression with lines representing 99-percent confidence limits on individual predicted values. **I=RLCLM90** requests the same analysis, but with 90-percent confidence limits on the mean predicted values.

The following example produces a regression analysis plot with system defaults.

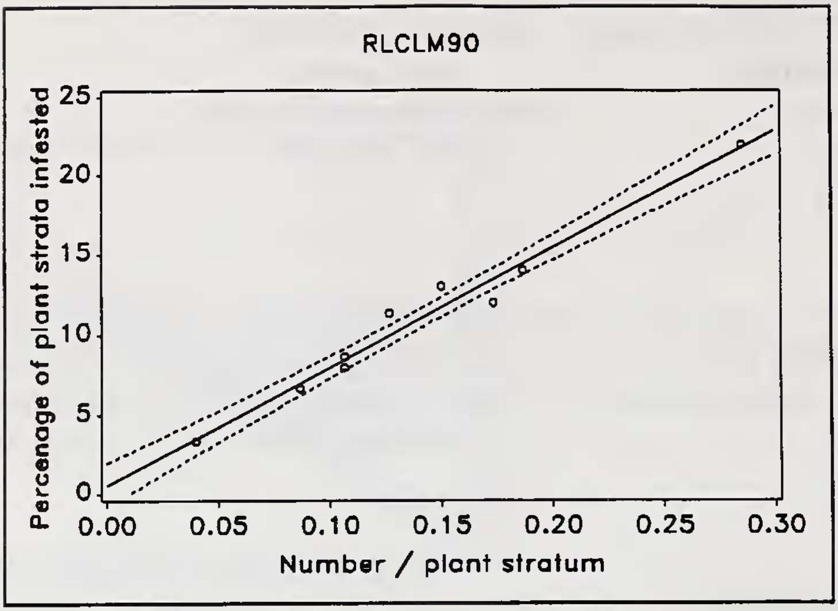
[1] SAS Institute, Inc. SAS/Graph R. Guide for personal computers. Version 6 edition. Cary, N.C.: SAS Institute, Inc., 1987. 500 pp.

```
GOPTIONS DEVICE=VGA16;
DATA LEAVES;
INPUT X Y;
CARDS;
.1267 11.33
.1067  8.00
.1067  8.67
.1867 14.00
.1733 12.00
.1500 13.00
.0867  6.67
.2850 22.00
.0400  3.33
;
SYMBOL I=RLCLM90;
PROC GPLOT;
PLOT  Y*X;
RUN;
```



This plot can be enhanced by supplying specific values for various options. For example, title and axes labels can be added using different text fonts and heights, the X and Y axis increments can be modified, and data point symbols can be added.

```
GOPTIONS DEVICE=VGA16 GUNIT=CM HTEXT=.8 FTEXT=DUPLEX;
TITLE  'RLCLM90';
AXIS1 LABEL=(A=90 'Percentage of plant strata infested')
          MINOR=NONE
          ORDER=0 TO 25 BY 5;
AXIS2 LABEL=('Number / plant stratum')
          MINOR=NONE
          ORDER=0 TO .3 BY .05;
SYMBOL V=CIRCLE I=RLCLM90 H=.5 L=1;
PROC GPLOT  DATA=LEAVES;
PLOT Y*X / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
RUN;
```



Text height and font are specified in the GOPTIONS statement and remain in effect until specified otherwise or until the session is terminated. Output can be further controlled by specifying options within TITLE, FOOTNOTE, NOTE, AXIS, LEGEND, PATTERN and SYMBOL statements.

The SAS/Graph manual documents in detail the many options available to enhance graphics output text and design.

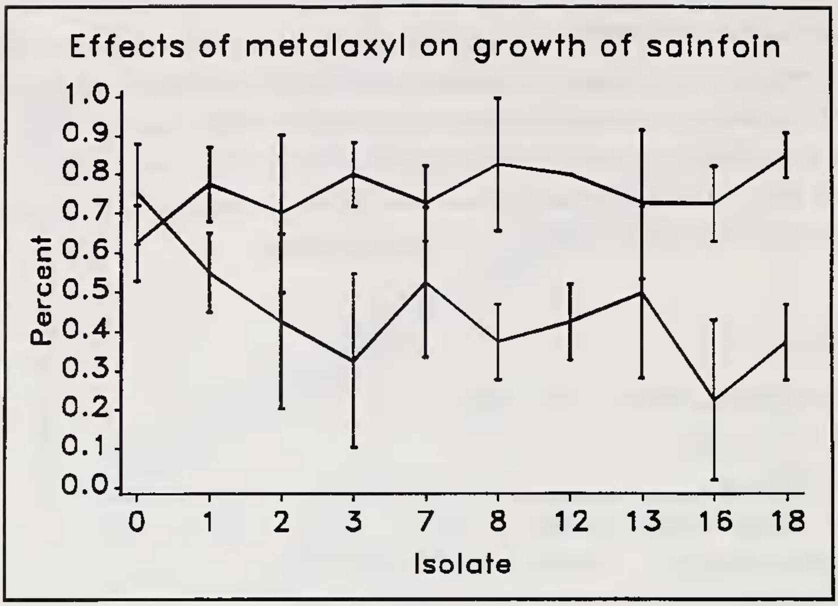
SAS/Graph standard deviation/standard error bar plots

As with regression, the I=option in the SYMBOL statement is used with the GPLOT procedure to add standard deviation/standard error bars to the data plotted. **I=STD** is used when multiple Y values occur for each level of X, and it is desired to join the mean Y value with (±) 1, 2, or 3 standard deviations for each X. **I=STDM** computes the standard error. **I=STD1JT** computes 1 standard deviation, the means are connected from bar to bar (J), and tops and bottoms are added to each bar (T). Further options are available for the interpolation values (I=option).

In the next example **I=STDMJT** requests standard error bars (default value of 2) with means connected from bar to bar and tops and bottoms added to each bar. This program also illustrates some data manipulation to overlay two lines on one plot. The original data set is subset into two data sets, the variable to be plotted renamed uniquely for each isolate, and finally the two data sets are merged again into one. The example also demonstrates how the output can be controlled separately within the TITLE, SYMBOL and AXIS statements.

Example:

```
GOPTIONS DEVICE=VGA16 GUNIT=CM;
DATA M; SET F2.SAIN;
IF CHEMICAL= 'M';
RENAME  PERCENT=PERCENTM;
DATA N; SET F2.SAIN;
IF CHEMICAL='N';
RENAME  PERCENT=PERCENTN;
DATA PLOT; SET M N;
TITLE F=DUPLEX H=1 'Effects of metalaxyl on growth of sainfoin';
SYMBOL1 I=STDMJT V=NONE W=2 L=1 C=B;
SYMBOL2 I=STDMJT V=NONE W=2 L=1 C=B;
PROC GPLOT;
AXIS1 VALUE=(H=.9 F=DUPLEX)
        ORDER=0 TO 1 BY .1
        WIDTH=2 MINOR=NONE
        LABEL=(A=90 H=.9 F=DUPLEX 'Percent');
AXIS2 VALUE=(H=.9 F=DUPLEX)
        ORDER= 0 1 2 3 7 8 12 13 16 18
        WIDTH=2  OFFSET=(.5CM)
        LABEL=(H=.9 F=DUPLEX 'Isolate');
PLOT (PERCENTM PERCENTN)* ISOLATE / OVERLAY
       VAXIS=AXIS1 HAXIS=AXIS2;
RUN;
```



If the data set already has standard deviation/standard error values, some data manipulation is necessary to create a plot with error bars. Multiple values of Y for each value of X will have to be created. Data set STD contains the variables ISOLATE, CHEMICAL, MEAN and STD. The following SAS DATA step can be used to reshape the data:

```
DATA NEWSTD;
SET STD;
PERCENT=MEAN+STD; OUTPUT;
PERCENT=MEAN; OUTPUT;
PERCENT=MEAN-STD; OUTPUT;
DROP MEAN STD;
RUN;
```

In the DATA step, the original data set (STD) is read in. Three values for PERCENT are created for each ISOLATE: the original mean, the mean plus the standard deviation and the mean minus the standard deviation.

Again the I=option in the SYMBOL statement is used with the GPLOT procedure to add error bars to the plot. **I=HILO** is used in this case since the mean and standard deviation are already known.

Similar options are available as are for **I=STD**.

Transferring graphics

Frequently it is necessary to incorporate SAS/Graph output within a document, or the output requires further modification. If so, this can be facilitated by correct device driver selection in SAS/Graph.

The graphics within this document were incorporated using WordPerfect®. The output from SAS/Graph was sent to a file in Hewlett-Packard® Graphics Language (HPGL) format and then placed in WordPerfect®. HPGL is WordPerfect's® recommended export format for SAS/Graph. To create the HPGL file, the GOPTIONS statement was used to select the HP7475 plotter driver in SAS/Graph and to direct the output to a file instead of the default serial port:

```
GOPTIONS DEVICE=HP7475 GACCESS='SASGASTD>hpexport.plt';
```

Using PageMaker® (the desktop publishing software demonstrated in the workshop), the most suitable SAS/Graph export format for placing graphics was found to be Encapsulated Postscript (EPS). With other export formats some detail is lost, the scale is altered, and there is a 64K file size limit. With EPS there is no loss of detail, the scale is retained even when reduced considerably, and there is no file size limit so complex graphics are not restricted. However, EPS graphics appear as a shaded box on the monitor because EPS is a language readable only by a postscript printer, and a postscript printer is required to print the document. Placing, sizing and moving graphics in PageMaker® is very simple and versatile, accomplished with mouse click and drag routines, and annotations can be added easily around and within the shaded graphics box.

SAS/Graph output can also be exported to other graphics software such as Harvard Graphics® and Lotus® Freelance for further modification. The preferred export format in this case would be Computer Graphics Metafile (CGM). SAS/Graph has a number of CGM drivers specific [illegible] the software to which one is exporting. Depending on the graphics software, once imported, the SAS/Graph output can now be edited and annotated. Each graphics software may have different size limits on the CGM file that it can import. In Harvard Graphics®, the metafile can be no larger than 32K, but with Zenographics Mirage™ you are limited only by the memory capacity of your system.

With the various options available for SAS/Graph export, the GREPLAY procedure is useful for storing graphics in a device independent catalog and replaying them later with the appropriate device driver required by the destination software.

Summary

The integration of statistical analysis and graphics in SAS/Graph along with its data management capacity makes it a powerful tool for statistical graphics.

Regression analysis plots and plots with standard deviation/standard error bars are requested in the SYMBOL statement with the interpolation option (I=option) in GPLOT.

SAS/Graph has many options available for enhancing graphical output design and text. If further modification is required, careful selection of the many device drivers available in SAS/Graph can ensure that graphics are transferred correctly to the destination software.

A DISCUSSION OF DISCRIMINANT ANALYSIS USING DYSTOCIA IN BEEF HEIFERS

J. A. Basarab, Beef Management Specialist
Beef Cattle and Sheep Branch, Animal Industry Division
Alberta Agriculture, Edmonton, Alberta

Abstract

Classification of subjects into two or more groups on the basis of one or more numeric measurements has long posed a problem to researchers. The most commonly used approach has been to apply consecutive numbers to the groups, treat them as the dependent variable and then subject the dependent variable and independent variables to multiple regression analysis. This approach assumes that the group variable is continuous, or at least that there is some numeric interval between the groups. How this numbering is applied can dramatically affect conclusions. In addition, the results from the multiple regression analysis lack clarity. The analysis does not classify subjects into groups, but merely identifies variables which significantly affect the dependent variable. Discriminant analysis, on the other hand, does not assume numeric continuity among groups and classifies subjects into discrete groups on the basis of a battery of measurements. Discriminant analysis using the SAS procedures will be discussed with calving difficulty in beef heifers as the example.

APPLICATION OF THE REPEATED MEASURES ANOVA IN AGRICULTURE

L.A. Goonewardene
Research Analyst, Beef Cattle and Sheep Branch
Alberta Agriculture
Edmonton, Alberta

Introduction

In the agricultural sciences it is often of interest/necessary to collect multiple observations on the same sampling unit. If the example is a plot of land and samples are taken from equal sized strips or sub plots which are randomized, the design is called a split plot. The common notation is to call the larger plot the 'main plot' and the randomized units on which measurements are made the 'sub or split' plots. The split plot is a very efficient design in soil/plant experiments due to its ability to use space. Furthermore, compared with a factorial design, where plot sizes for the different combination of main effects are equal, the split plot can sub-divide the main plot into many levels of splits and still maintain all of the interactions among main effects.

Often times, multiple measurements are taken on each main plot at different times and such a design is called a split plot in time. The design can remain a 'true' split plot provided the time at which samples are taken is randomly assigned to plots, but when the sampling unit is an animal or individual, time often becomes fixed. Thus, where levels of the sub or split plot are fixed (i.e. not assigned randomly) it is called a repeated measures design (Pendergast and Littel 1988).

The objective of the Workshop/Paper was to work through the analysis of a split plot and repeated measures design using the SAS^{R} system and compare the univariate and multivariate analyses.

The Split Plot

The split plot analysis is a univariate type where there is only a single dependent variable (Y) and many independents (X^{S}). The example used is hypothetical, where 12 animals have been subject to two treatments F & C and an enzyme labelled as G6P sampled at four periods (time) designated as 1,2,3 & 4.

SAS CODE

```
   DATA  SPLIT;
   INPUT ANIMAL TREAT $ PERIOD G6P;
   CARDS;
          1     F     1     14
          1     F     2     12
          1     F     3     16
          1     F     4     11
          2     F     1     12
          2     F     2     16

         12     C     1     21
         12     C     2     22
```

The SAS code shown above is the usual format for entering data designed for a split plot type of analysis (in columns). However, if the data is entered for a multivariate (repeated) measures analysis of variance in rows, then some modification is needed to get it into a split plot format. For example, if the data were read in as G6P1 to denote G6P at Period 1, G6P2 to denote G6P at period 2 , and so on, a DROP statement and four OUTPUT statements would be needed as shown below to get the data into shape (ie. a row to column conversion) for analysis.

<u>SAS Code</u>

```
DATA  SPLIT1;
INPUT ANIMAL TREAT $ G6P1 G6P2 G6P3 G6P4;
DROP G6P1 - G6P4;
PERIOD = 1; G6P = G6P1; OUTPUT;
PERIOD = 2; G6P - G6P2; OUTPUT;
PERIOD = 3; G6P = G6P3; OUTPUT;
PERIOD = 4; G6P = G6P4; OUTPUT;
CARDS;
```

Once the data are in columns, either PROC ANOVA or PROC GLM can be invoked. If the data are balanced, either procedure may be used, but if the data has missing values, it is recommended that PROC GLM; be used and least squares estimates obtained.

SAS Code

```
PROC GLM;
CLASSES TREAT ANIMAL PERIOD;
MODEL G6P = TREAT ANIMAL(TREAT) PERIOD
          TREAT * PERIOD/SS3;
TEST H = TREAT E = ANIMAL(TREAT)/ETYPE = 3 HTYPE = 3;
MEANS TREAT/E = ANIMAL(TREAT) ETYPE = 3 SNK;
LSMEANS TREAT/E = ANIMAL(TREAT) ETYPE = 3 STDERR;
MEANS PERIOD TREAT * PERIOD/ETYPE = 3 SNK;
LSMEANS PERIOD TREAT * PERIOD/ETYPE = 3 STDERR;
RUN;
```

The above program should generate SAS output which shows the 'split plot type' analysis of variance in two parts, commonly called the analysis above and below the split. The variation associated with TREAT is usually considered as being above the split or main plot , whereas PERIOD and interaction of PERIOD and TREAT are considered below the split or sub plot. The SAS System by default tests all main effects such as TREAT with the residual mean square (error mean square) as the Model is fixed and only the residual or error term is random. However, in a split plot, the correct error term (based on expected mean squares) for testing TREAT is ANIMAL(TREAT) and this should be specified in the program using a TEST Statement. It is also necessary to specify the error term for the calculation of the standard error in the LSMEANS Statement. A type III sums of squares is used in all cases.

The SAS output (modified) will look somewhat like this:

Table 1 SAS Printout From a Typical Split Plot Design

```
DEPENDENT VARIABLE G6P
SOURCE         DF      SS      MS      F VALUE      PR>F      R-SQ      C.V.
MODEL          17     789.5   46.6      13.05      0.0001     0.88      11.3
ERROR          30     106.7    3.5        ROOT MSE        G6P MEAN
CORR TOTAL     47     896.3              1.8                 16.6
SOURCE         DF     TYPE III SS      F VALUE      PR>F      Comment*
TREAT           1     500.5             140.6       0.0001  Above the split*
ANIMAL(TREAT)  10     268.0               7.5       0.001    (incorrect)

PERIOD          3       8.2               0.77      0.51    Below the split*
TREAT*PERIOD    3      12.7               1.19      0.32     (correct)

TEST OF HYPOTHESIS USING TYPE III MS FOR ANIMAL(TREAT)
TREAT           1      500.5            18.6        0.001   Above the split*
                                                             (correct)
* Will not appear in SAS output.
```

As you may notice, the terms below the split i.e. PERIOD TREAT*PERIOD are correctly tested and the F and P values are accurate as the appropriate error term has been used for testing. However, above the split the analyst should only use the F value of 18.6 and P>F of 0.001 as the TEST statement has now directed the program to use the appropriate error term ANIMAL(TREAT) for testing. Thus, it could be seen that TREAT effects are significant P=0.001, whereas PERIOD and TREAT*PERIOD is not significant P>0.05. In addition, the program will generate means with a Student Newman Keuls' comparison, least squares means and standard errors, for TREAT, PERIOD and TREAT*PERIOD, as they have been specified by the MEANS and LSMEANS statements.

The Repeated Measures analysis

The Repeated Measures analysis is a univariate and multivariate ANOVA and assumes that: at least one factor consists of multiple measurements taken on the same subject; repeated measures are independent across subjects; and assumes a normal multivariate distribution with a common covariance matrix. A multivariate normal distribution for data and residuals could be checked out by processing the data through PROC UNIVARIATE and testing for normality. A common or symmetrical covariance matrix is assumed when there is compound symmetry, correlation between all pairs in the matrix are similar and equal. This is called auto correlation. There is, however, a trend in animal data where partial correlations are higher when time intervals are closer together and correlations decrease as time intervals widen.

<u>SAS Code</u>

```
DATA  RPTD;
INPUT ANIMAL TREAT $ G6P1  G6P2  G6P3  G6P4;
CARDS;
  1     F     14    12    16    11
  2     F     12    16    19    20
  3     F     12    15    12    11

 11     C     18    17    16    20
 12     C     23    26    21    22
```

Note: The data appears in rows rather than columns.

```
PROC  PRINT;
TITLE 'REPEATED  MEASURES  ANOVA';
RUN;
PROC GLM DATA = RPTD;
  CLASS  TREAT;
  MODEL G6P1 G6P2 G6P3 G6P4 = TREAT;
  REPEATED  PERIOD  4  CONTRAST(1)/SHORT  PRINTM
              PRINTE  SUMMARY;
  RUN;
```

To perform both the univariate repeated measures and multivariate analyses in a single procedure the MODEL statement is set up with multiple dependent variables (G6P1 - G6P4) to reflect the between animal design and the REPEATED statement links the between and within subject effects ie. effects above and below the split. The resulting SAS print out will provide univariate analyses of variance within each period and since this example has 4 periods there will be 4, one Way ANOVA'S, the Repeated Measures ANOVA with tests of hypothesis for between and within subject factors.

Table 2 SAS Print out (ANOVA only) from a Repeated Measures ANOVA

```
REPEATED MEASURES ANOVA
GLM PROCEDURE
TESTS OF HYPOTHESES FOR BETWEEN SUBJECTS EFFECTS  -  Above the Split*
SOURCE        DF     TYPE III SS    MS    F VALUE   PR>F
TREAT          1       500.5      500.5   18.6      0.001   ADJUSTED PR>F
ERROR         10       268.0       26.8                       G-G      H-F
PERIOD         3         8.2        2.7    0.77     0.51     0.48     0.51 Below*
PERIOD*TREAT  3        12.7         4.2    1.19     0.32     0.32     0.32  the
ERROR(PERIOD) 30     106.7          3.5                                     split
     G-G  EPSILON = 0.73
     H-F  EPSILON = 1.04

* These comments will not appear in the SAS output.
```

A comparison with the values in table 1 shows that the mean squares, F values and P>F are identical. Therefore, the same ANOVA table with effects above and below the split are obtainable either through a split plot or repeated measures analysis.

There is one rather serious limitation using the REPEATED statement. Although both the split plot and repeated measures can derive Means for TREAT and PERIOD it is often of interest to obtain Means for the TREAT*PERIOD interaction. If the differences between the treatments are the same across periods, then the interaction between TREAT and PERIOD will not be significant. However, if this interaction is significant ($P<0.05$) it is necessary to know when the difference(s) occur. Unfortunately, the TREAT*PERIOD interaction means cannot be obtained in a repeated measures analysis (using a REPEATED statement) and thus one may have to revert to a split plot type of analysis to obtain interaction means. The repeated measures analysis provides partial correlation coefficients among the repeated measurements (G6P1-G6P4); table 3. One could use this to see if auto correlation exists in the matrix. There is often a tendency for the correlation to decrease as the time interval widens, which often results in the violation of the H-F/sphericity condition (to be discussed later).

The PRINTM option prints the transformation matrices that define the contrasts, while the PRINTE option instructs the SAS System to print the error sums of squares and cross products matrix associated with the repeated measurements G6P1 to G6P4, as well as the sphericity tests for each transformed variable that are orthogonal. The short option prints the multivariate test criteria and associated F tests in a condensed form. (See SAS-Users Guide version 5 Ed or SAS/STAT 6.03 for details).

Table 3 Partial correlation coefficients from the error SS & CP matrix in a repeated measures analysis (modified)

PARTIAL CORRE. COEFF. FROM THE ERROR SS & CP MATRIX/PROB.lRl

DF=9	G6P1	G6P2	G6P3	G6P4
G6P1	1.0	0.71	0.60	0.45
G6P2	0.71	1.0	0.66	0.55
G6P3	0.60	0.66	1.0	0.79
G6P4	0.45	0.55	0.79	1.0

The univariate repeated measures = standard split plot analysis provided certain conditions are met. The first is of compound symmetry ie. when correlations between all pairs of repeated variables is equal. This condition is sometimes stronger than required and is called the Huynh - Feldt (H-F) condition in SAS. The H-F condition is met only if all possible pairs of repeated measurements have equal variances. The second requirement is of sphericity ie. where any set of orthogonal contrasts have equal variances and zero covariance. The validity of H-F can be determined by a Chi-Square test based on Mauchly's sphericity test for Othogonal contrasts (Mauchley 1940, Pendergast and Littel 1988). If the sphericity test fails, the H-F condition is not valid. Inclusion of the PRINTE option generates a Chi-square value for Mauchley's sphericity criterion. If the X^2 value has a probability <0.05 one would reject the null hypothesis of sphericity and conclude that the H-F and sphericity criteria are violated. If this be the case the univariate analysis of variance is not correct for the within subject effects of PERIOD and TREAT*PERIOD. If on the other hand the sphericity criterion is not violated, (acceptance of ho) then a split plot type of analysis can be performed, and TREAT*PERIOD means compared by the usual procedures.

In our example, Mauchley's sphericity for orthogonal contrasts was P=0.47, thus accepting the null hypothesis of sphericity. As such, all univariate F tests are valid, and there is no need to use the multivariate or adjusted G-G, H-F tests.

Univariate and Multivariate Tests

In addition, the SAS program provides multivariate F tests for PERIOD and TREAT*PERIOD interaction (Table 4). The multivariate tests do not require the H-F condition or sphericity to hold and can be used to test period (time) and interaction effects in the split. The SAS System provides Wilk's Lambda, Pillai's Trace, Holelling-Lawley Trace & Roy's Maximum Root tests (all multivariate tests). The exact F tests and P>F value is usually given. In our example the P values are 0.72 for PERIOD and 0.19 for TREAT*PERIOD interaction respectively, which are not significant. Furthermore, in our example, H-F and sphericity is not violated and the univariate F test probabilities for PERIOD and TREAT*PERIOD are 0.51 and 0.32 respectively, all being non-significant. It could thus be concluded that the univariate F tests are therefore valid for the within subject or below the split effects in our example.

There are times, however, when the multivariate and univariate F tests do not match, especially if sphericity H-F condition is violated. When this occurs, one can either accept the multivariate F test (which usually lacks power) or use the adjusted F tests, Greenhouse-Geisser (G-G) or Huynh-Feldt (H-F) tests. These two tests are slightly more conservative than the exact univariate F tests. If the univariate F test is used under conditions where the H-F condition is violated, the test is too liberal for the within subject effects (PERIOD & TREAT*PERIOD), and therefore one may encounter a Type I error, i.e. rejection of the null hypothesis when in fact it is true.

Table 4 Multivariate tests for period and period X treatment interaction (Modified)

```
REPEATED MEASURES ANOVA
GLM PROCEDURE
H = TYPE III SS & CP MATRIX FOR PERIOD:
PERIOD.N REPRESENTS THE CONTRAST BETWEEN THE NTH LEVEL OF
PERIOD AND THE 1ST
  DF=1            PERIOD.2       PERIOD.3       PERIOD.4
  PERIOD.2         4.08          4.66           8.16
  PERIOD.3         4.66          5.33           9.33
  PERIOD.4         8.16          9.33          16.33
```

```
MANOVA TEST CRITERIA AND EXACT F STATISTICS FOR THE HYPOTHESIS OF
NO PERIOD EFFECT.  H = TYPE III SS & CP MATRIX FOR;  PERIOD
E = ERROR SS & CP MATRIX
S = 1M = 0.5 N = 3
STATISTIC                VALUE      F     NUM DF     DEN DF       PR>F
WILKS' LAMBDA             0.85     0.45     3          8          0.72
PILLAI'S TRACE            0.14     0.45     3          8          0.72
HOTELLING-LAWLEY TRACE    0.16     0.45     3          8          0.72
ROY'S GREATEST ROOT       0.16     0.45     3          8          0.72
```

```
MANOVA TEST CRITERIA FOR EXACT F STATISTICS FOR THE HYPOTHESIS OF
NO PERIOD*TREAT EFFECT H = TYPE III SS & CP MATRIX FOR TREAT*PERIOD
E = ERROR SS & CP MATRIX
STATISTIC                VALUE      F     NUM DF     DEN DF       PR>F
WILKS' LAMBDA             0.56     2.01     3          8          0.19
PILLAI'S TRACE            0.43     2.01     3          8          0.19
HOTELLING-LAWLEY TRACE    0.75     2.01     3          8          0.19
ROY'S GREATEST ROOT       0.75     2.01     3          8          0.19
```

In the SAS System, both the G-G and H-F estimates of Box's E are available through PROC ANOVA and PROC GLM, providing the Probabilities are calculated on the adjusted degrees of freedom. Violation of the H-F condition does not affect the between subject (TREAT) effect.

The repeated measures ANOVA usually handles balanced data and any animal/subject which does not have a complete set of data (repeated measurements) is not included in the analysis. However Milliken and Johnson (1983) and Pareja (1990) provide details of the procedure and SAS code to handle unbalanced data. Often time, the sphericity condition can be violated by having a large number of repeated measurements on the same animal. Some restriction on the number of repeated measurements may be helpful in conforming to sphericity criteria.

In conclusion, experiments where repeated measurements are taken on the same subject, can be analysed either as a univariate (repeated or split plot) or multivariate (MANOVA) type. If H-F/Sphericity is not violated, the univariate approach can be used. Analysis of the repeated measures design as a split plot in time is particularly useful as interaction means for within subject effects can be computed. Furthermore, a split plot in time will handle unbalanced designs and generate least squares estimates, provided PROC GLM is used. If, on the other hand, the H-F condition is in question, either a Multivariate approach (Wilk's Lambda) or adjusted Univariate (G-G & H-F) tests should be used. It is recommended that Univariate tests be used when appropriate, as it is more powerful than the Multivariate tests. All these conditions are, however, only applicable to the within subject (below the split) effects.

Acknowledgements

I wish to acknowledge the use of the paper "Repeated Measures using the SAS System" by Jane F. Pendergast and Ramon C. Littel, appearing in the 13th Annual SUGI Proceedings, as many ideas for this paper were taken from it.

Bibliography

Greenhouse, W.S. and Geisser, S. 1959. On methods in the analysis of profile data. Psychometrica 24: 95-112.

Huynh, H. and Feldt, L.S. 1970. Conditions under which mean square rations in repeated measurements designs have exact F-distributions. J. Am. Stat. Assn. 65: 1582-1589.

Huynh, H. and Feldt, L.S. 1976. Estimation of the box correlation for degrees of freedom from sample data in the randomized block and split plot designs. J. Educ. Stat. 1: 69-82.

Mauchly, J.W. 1940. Significance test for sphericity of a normal n-variate distribution. The Annals of Math. Statistics 11: 204-209.

Milliken, G.A. and Johnson, D.E. 1984. Analysis of messy data, V1: Designed experiments. Belmont CA. Lifetime Learning Publications.

Olson, C.L. 1976. On choosing a test statistic in multivariate analysis of variance. Psych. Bull. 83: 579-586.

Pareja, G.D. 1990. Comparison of different procedures to analyse an unbalanced repeated measures design. 15th Annual SUGI Proceedings, p.1278-1282.

Pendergast, J. and Littel, R. 1988. Repeated measures analysis with the SAS^{R} System. 13th SUGI Proceedings, p. 1205-1211.

SAS Institute Inc. 1985. SAS User's Guide: Statistics Version 5 Ed., Cary, N.C. SAS Institute Inc.

SAS Institute Inc. 1985. SAS/STAT guide for personal computers. V. 6 Ed. Cary, N.C. SAS Institute Inc.

STATISTICAL ISSUES IN INSECT POPULATION RESEARCH

G. Bruce Schaalje
Agriculture Canada Research Station
Lethbridge, Alberta

Introduction

Research involving insect populations as the experimental units presents the statistician with many interesting challenges. In this paper I discuss some of the characteristics of insect populations that must be taken into consideration when designing experiments and analysing data, and review some of the statistical methods that have been devised for work with insect populations.

Characteristics of Insect Populations

Insects are Small

While this statement is obvious, its implications for statistical work are important. One of the biggest challenges in insect population research is in estimating the size of insect populations, but because insects are small they are hard to find, count, and identify completely as to species, sex, and growth stage. Furthermore, their size often renders them fragile and vulnerable to microenvironmental conditions. As a result, the estimation of population size usually involves shortcut methods such as incomplete counts, indirect assessments, lures, mechanical collection devices (Southwood 1978), and data aggregated over growth stages, sexes, or species (Brillinger et al. 1980). Most of these methods require complicated statistical models to obtain estimates and standard errors, and for calibrating collection devices (Mcdonald and Manly 1989). Data from these methods are often subject to large unexplained fluctuations and lots of variability.

Insects are Poikilothermic

Insect development and activity is dependent on temperature, and thus temperatures must be recorded in experiments involving population change. This limits the use of "replication in time" in experiments using natural populations (particularly if they are univoltine), and often necessitates the incorporation of mathematical models for the relationship between temperature and development into statistical analyses (Curry and Feldman 1987).

Insects are Not Stationary

Insects migrate in and out of the study area during the course of any experiment (Kuno 1991), and often change their location in plants at different times of day. This movement has to be taken into account either in the design stage of the experiments, for example by using buffer areas around the populations and sampling at specific times of the day, or in the analysis stage by adjusting mathematically for migration. This may necessitate the collection of covariates to indicate various aspects of migration.

The Spatial Distribution of Insects is Usually Not Random

The spatial distribution of insects is generally aggregated to some degree (Taylor 1984) so that simple models for sampling or population size estimation based on the Poisson distribution usually do not give good results (Taylor 1987). Also, insect distributions on agricultural fields usually display edge effects which must be taken into account in designing experiments.

The Appropriate Spatial Unit for Sampling may not be Obvious

The degree of aggregation of an insect population may be an artifact of the quadrat size used to sample the population and hence may have little meaning unless the quadrat size is "natural" in some sense. A standard area or volume may not be appropriate because of varying numbers of plants, refugia, etc. (Regniere et al. 1989). A lot of thought has to be put into the choice of a sampling unit.

Insects Populations have an Age Structure

This has to be kept in mind in insect population research because the various growth stages of an insect have unique physiologies, behaviour,catchabilities, survivorship, etc. The different growth stages will react to treatments differently and thus the age structure of the population has to be either controlled in the experiment or observed and adjusted for mathematically (Schaalje et al. 1989).

Insects Develop Fast

As a result of this, experiments on insect populations usually have to involve repeated measurements, pretreatment assessments of the populations, and control populations. Often apparent treatment effects have to be adjusted for population changes (mortality, reproduction, development) that would have occurred naturally during the experiment (Abbott 1925).

The Population Structure, Habitat, and Treatments Effects are Dynamic

All three of these components of an insect population experiment have independent dynamics which may interact in complicated ways. The analysis and interpretation of data from these experiments often have to be based on a detailed population model as well as models for the environment and the treatment (Schaalje et al. 1989).

Statistical Methods for Insect Population Research

Sampling Methods

Many techniques of finite population sampling are important in estimating the size of insect populations. Stratified sampling, double sampling, and multistage cluster sampling (Southwood 1978) are all useful in sampling insect populations.

Binomial sampling (Schaalje et al. 1991, Nyrop and Binns 1991) by which population density is estimated using presence-absence data from a random sample of units is very important in sampling small, highly aggregated insects. Measurement error models (Fuller 1987) are necessary in getting unbiased prediction equations for binomial sampling, and for calculating appropriate estimates of the standard error of prediction.

Sequential sampling (Nyrop and Binns 1991, Kuno 1991) has been applied extensively to pest management applications so that as few samples as possible can be taken to determine whether the population has reached an economic threshold.

Mark-recapture methods in which the population is assumed to be open to migration, and which involve multiple releases but a single recapture, have recently been found to be useful in estimating the population size and survival for certain species of insects (Burnham 1989, Lysyk and Axtell 1986).

McDonald and Manly (1989) discuss how biased sampling procedures arise in insect population sampling and suggest methods for their calibration. Finally, ratios often arise when using data from sampling devices to estimate population size, and Buonaccorsi and Liebhold (1989) discuss the unbiased estimation of such ratios.

Spatial Analysis

Various statistical models for the spatial distribution of insects (Southwood 1978) have been suggested and used, most notably the negative binomial distribution. In addition, indices of aggregation have been developed and discussed (Hurlbert 1990). General models for the relationship between the mean and the variance of populations for a given species (Taylor 1984, Iwao 1976, Kuno 1991) are useful in characterizing insect species and predicting their spatial distribution at a given density. Binns (1986) discusses the relationship between Taylor's variance-mean model and the negative binomial distribution. The relatively new field of geostatistics has been applied to insect populations to provide the ability to predict population densities at any particular location in a field (Schotzko and O'Keeffe 1989), and on a larger scale geographical information systems are providing similar capabilities (Johnson 1989).

Stage-Frequency and Related Demographic Analyses

Several methods have been developed for estimating stage-specific development times, mortality rates, and reproductive rates of insects from a sequence of cross-sectional samples of insect populations (Manly 1989). These methods differ as to the type of data collected (non-overlapping cohorts, multiple cohorts, etc.), the type of information desired, and the assumptions of the models upon which the methods are based. In addition, "key-factor analysis" (Varley and Gradwell 1960, Manly 1989) provides a method for analysing stage-frequency data collected over several generations and determining the relative contribution to population dynamics of mortality in the various growth stages. Finally, Carey (1989) discusses how traditional demographic methods can be applied to the study of insect population dynamics.

Stochastic Modelling of Population Dynamics

Much statistical analysis of insect population data involves fitting a stochastic model of population dynamics to the data. Brillinger et al. (1980) developed a stochastic difference equation to model aggregate insect data and investigate density-dependent mortality. Blough (1989) used time series methodology in connection with a difference equation model in his analysis of insect population changes due to a pesticide. Dennis (1989) discussed the use of stochastic differential equation models for insect populations. Schaalje and van der Vaart (1989) reviewed stage-specific population models which allow for variability in developmental rates, and Schaalje et al. (1989) applied such a model to the analysis of field data on a population sprayed with a pesticide. Curry and Feldman (1987) discuss several issues and strategies important in modelling and analysing insect populations.

References

Those references preceded by an asterisk (*) are particularly good general references for statistical methods for insect populations.

Abbott, W. S. 1925. A method of computing the effectiveness of an insecticide. J. Econ. Entomol. 18:265-267.

Binns, M. R. 1986. Behaviourial dynamics and the negative binomial distribution. Oikos 47: 315-318.

Blough, D. K. 1989. Intervention analysis in multivariate time series via the Kalman filter. pp. 389-403 in EAIP (see below).

Brillinger, D., Guckenheimer, J., Guttorp, P., and Oster, G. 1980. Empirical modelling of population time series data: the case of age and density dependent vital rates. pp. 65-90 in Lectures on Mathematics in the Life Sciences, Vol. 13, Oster, G. (ed.), Amer. Math. Soc., Providence.

Burnham, K. P. 1989. Numerical survival rate estimation for capture-recapture models using SAS PROC NLIN. pp. 416-435 in EAIP (see below).

Buonaccorsi, J. P. and Liebhold, A. M. 1989. Estimating the size of gypsy moth populations using ratios. pp. 404-415 in EAIP (see below).

Carey, J. R. 1989. Demographic framework for analysis of insect life histories. pp. 206-218 in EAIP (see below).

*Curry, G. L. and Feldman, R. M. 1987. Mathematical Foundations of Population Dynamics. Texas A&M Press. College Station.

Dennis, B. 1989. Stochastic differential equations as insect population models. pp. 219-238 in EAIP (see below).

Fuller, W. A. 1987. Measurement Error Models. Wiley. New York.

Hurlbert, S. H. 1990. Spatial distribution of the montane unicorn. Oikos 58:257-271.

Iwao, S. 1976. Relation of frequency index to population density and distribution pattern. Physiol. Ecol. Japan 17:457-463.

Johnson, D. L. 1989. Spatial analysis of the relationship of grasshopper outbreaks to soil classification. pp. 347-359 in EAIP (see below).

*Kuno, E. 1991. Sampling and analysis of insect populations. Ann. Rev. Entomology 36:285-304.

Lysyk, T. J. and Axtell, R. C. 1986. Estimating number and survival of house flies with mark/recapture methods. J. Econ. Entomol. 79:1016-1022.

*Manly, B. F. J. 1989. A review of methods for the analysis of stage frequency data. pp. 3-69 in EAIP (see below).

*Manly, B. F. J. 1989. A review of methods for key-factor analysis. pp. 169- 189 in EAIP (see below).

McDonald, L. L. and Manly, B. F. J. 1989. Calibration of biased sampling procedures. pp. 467-483 in EAIP (see below).

*Nyrop, J. P. and Binns, M. 1991. Quantitative methods for designing and analysing sampling programs for use in pest management. in Handbook of Pest Management in Agriculture, Pimentel, D. (ed.), CRC Press, Boca Raton.

Regniere, J., Lysyk, T. J. and Auger, M. 1989. Population density estimation of spruce budworm on Balsam Fir and White Spruce from 45-cm mid-crown branch tips. Can. Ent. 121:267-281.

*Schaalje, G. B. and van der Vaart, H. R. 1989. Relationships among recent models for insect population dynamics with variable rates of development. J. Math. Biol. 26:399-428.

Schaalje, G. B., Stinner, R. L. and Johnson, D. L. 1989. Modelling insect populations affected by pesticides with application to pesticide efficacy trials. Ecol. Model. 47:233-263.

Schaalje, G. B., Butts, R. A. and Lysyk, T. J. 1991. Simulation studies of binomial sampling: a new variance estimator and density predictor, with special reference to the Russian Wheat Aphid. J. Econ. Entomol. 84:(in press).

Schotzko, D. J. and O'Keeffe, L. E. 1989. Geostatistical description of the spatial distribution of *Lygus hesperus* in lentils. J. Econ. Ent. 82:1277-1288.

*Southwood, T. R. E. 1978. Ecological Methods. Chapman and Hall. New York.

*Taylor, L. R. 1984. Assessing and interpreting the spatial distribution of insect populations. Ann. Rev. Entomology 29:321-357.

Taylor, R. A. J. 1987 On the accuracy of insecticide efficacy reports. Env. Entomol. 16:1-8.

Varley, G. C. and Gradwell, G. R. 1960. Key factors in population studies. J. Anim. Ecol. 29:399-401.

*EAIP = Estimation and Analysis of Insect Populations, Lecture Notes in Statistics, Vol. 55, McDonald, L., Manly, B., Lockwood, J., and Logan, J. (eds.), Springer Verlag, New York.

EXPERIMENTAL DESIGN: BASIS FOR SOUND RESEARCH METHODS

L. Zack Florence
Animal Sciences Division, Biometrics Section
Alberta Environmental Centre
Vegreville, Alberta T0B 4L0

Introduction

Well-planned experiments increasingly form the basis for cost effective research. Few people today would argue that budgeting time and money does not largely dictate whether an experiment will be attempted. For those using humans or other animals as experimental subjects, the care, welfare and ethical use of research subjects is the first consideration: choosing sample sizes takes on a new meaning and has drawn renewed attention to power analysis during project planning (Mann et al. 1991).

Most of us encountered our first meaningful experimental design experience at work or in graduate school. This was also our first introduction to planning research in committee. It brings back some questionably fond memories for many of us and will be left at that. The point to be made here is that few people do research alone; most of us must work in an environment where, in order to meet our project objectives, we must help other scientists meet theirs. This paper will discuss a few of the topics that must be considered when planning a co-operative experiment. It will be assumed that even if you are not involved in the committee approach to doing research that you are at least in contact with a statistician or someone in whom you may confide (a person(s) who helps choose the proper model and the appropriate design, and makes inferences based upon the results of analyses). This paper is a short discussion of the process; several good texts such as Box et al. (1978), Cox (1958) and Anderson and McLean (1974) may be consulted for more in-depth discussions and underlying theory. The one mistake all of us hope to avoid when planning experiments is what A.W. Kimball (Kimball 1957, cited by Box et al) called, "error of the third kind", that is, obtaining the right answer to the wrong problem.

Ordered List of Steps in Planning an Experiment

Anderson and McLean (1974) prepared a list of 12 steps that should be followed when planning an experiment. I have added my comments and examples to their list (as follows). When moving through the list, keep in mind that it is implied that a statistician is involved in this process from the beginning.

1. Recognition that a problem exists

- The problem is usually, but not always, apparent and, if a group or committee is involved, it is important to have consensus.

2. Formulation of the problem.

- Uninhibited discussion among participants in a group situation is healthy. If it is not a group project, it is wise to seek the experience of others in the same area of research. In either case, a collaborative effort will make it easier to come to a decision about the most likely problems or subject areas requiring research effort. Do not be surprised if it takes more effort than you expected to identify the objectives of an experiment.

3. Agreeing on factors and levels to be used in the experiment

- In step 2, you identified a probable cause(s), and you are now ready to identify treatments or factors (independent variables) that you can vary (levels) and measure (dependent variables such as response, yield or product). Continuous independent treatment factors will require different assumptions than discrete discontinuous ones.

4. Specifying the variables to be measured.

- Quantitative responses by dependent variables are usually more apparent than are qualitative ones; the latter responses, are not metric and are at best ordinal. The decision to use one or the other is very important.

5. Definition of the inference space for the problem

- This decision determines how far you may legitimately extend the results of the study; for example, if an experiment is bounded by 20 and 60 degrees C, you are limited to this space and the inferences are bound by choosing the proper error terms when fitting the model.

6. Random selection of the experimental units

- The experimental unit is the material, area, time, plot, pot or whatever, which will receive the treatment. In animal experiments, the experimental unit may be a pen of animals, with the same randomly assigned treatment applied to all animals in the pen.

- The experimental unit should be representative of the inference space, that is, growth trials done in a greenhouse may have little relationship to the field environment.
- The number of experimental units determines the standard error or precision which will be associated with the inferences you wish to make from your results; keeping the standard error low may require blocking (e.g., spatially, temporally, by treatment) to avoid systematic bias.

7. Assignment of treatments to the experimental units

- Each experimental unit (plot, cow, beaker, petri plate) should have equal chance of receiving each treatment.
- Randomization (of error) is the basis for the design of the experiment, for example, in a randomized complete block, split-plot experiment, main plots are randomized within blocks, and sub-plots are randomized within main plots, within blocks.

8. Outline of the analysis corresponding to the design before the data are taken

- Quoting form Anderson and McLean (1974) pp. 87-88: "At this point, the statistician must write down the mathematical model that has evolved as a result of the committee activity in the preceding sections. This mathematical model will give rise to the ANOVA table. The ANOVA table will now consist of the degrees of freedom and the expected mean squares for each of the specific factors...". The point made here is that the design and the model to which data are to be fit go hand-in-hand, and the tests of interest are determined before the experiment is done. This point is discussed further in the following section.

9. Collection of the data

- Too many experiments begin here!
- Quality control and quality assurance are paramount at this stage; several people may be collecting data during the course of a project and someone should be in charge of regularly validating data and collection methods.
- Data forms should be simple and straightforward— have a trial run to make sure everyone understands how the forms are to be used.

10. Analysis of the data

- In this step you should plot means and individual data in scatterplots, calculate descriptive statistics and use statistical packages which will test for normality and produce normal probability plots.
- Ask yourself: Are parametric assumptions met, that is, are variances equal; are

variances independent of the means; are transformations needed?

- If you have decided upon the correct model during the design, applying the correct analysis should be fairly straightforward because you have previously decided how the data are to be analyzed and which tests are appropriate. Beware! Unless you specify otherwise, all statistical software packages assume all effects in your model are fixed. This assumption can lead to spurious conclusions regarding tests for significance.

11. Conclusions

- Once again, if this is a group, interdisciplinary effort, you will likely need to get together and discuss the results of the analyses. Experience suggests that this can be a very valuable exercise: many times, leaving interpretation of results to one person may lead to error, and worthwhile information may be overlooked.

12. Implementation

- Time now for a "management" decision: do you alter a manufacturing process based upon the results, publish the results or wait until additional study is completed; can you implement a technology transfer program to agricultural producers based on your study's conclusion? The ease with which these decisions can be made will be contingent upon how closely you have adhered to a process like the one discussed here.
- If an experiment has been based on prototypes or bench techniques, it will be time to think of applying what you have learned to "scale up", which for example, to a crop scientist means doing field trials to determine how well greenhouse or lab experiments fit the real world.

From Design to Model to Analysis

Ideally, the statistical model to which we hope to fit the data should have been determined by the time you have reached step number 8. This is not always the case and too often the very close linkage between the experimental design and the model, and how data will be analyzed and average differences tested, are not realized. A simple illustration can serve to demonstrate the associations.

Let us assume you have identified two experimental factors, α and β, which you wish to vary (Figure 1). Treatments composed of factorial combinations of the levels (i, j, respectively) of α and β are randomly assigned to experimental units, which are arranged in a completely randomized design. You want to measure the response (Y) due to main effects of α and β and their

FIGURE 1. Two factor (α, β), full factorial experiment showing changes among expected mean squares (EMS) conditional upon assumptions about α and β.

The Model

$$Y_{ijk} = u + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{(ij)k}$$

If the Experiment calls for both α and β factors to be fixed:

Source	EMS
α_i	$\sigma^2 + \beta n \theta(\alpha)$
β_j	$\sigma^2 + an\,\theta(\beta)$
$\alpha\beta_{ij}$	$\sigma^2 + n\,\theta(\alpha\beta)$
$\varepsilon_{(ij)k}$	σ^2

If the Experiment calls for both α and β to be random:

Source	EMS
α_i	$\sigma^2 + n\,\sigma^2_{\alpha\beta} + bn\,\sigma^2_{\alpha}$
β_j	$\sigma^2 + n\,\sigma^2_{\alpha\beta} + an\,\sigma^2_{\beta}$
$\alpha\beta_{ij}$	$\sigma^2 + n\,\sigma^2_{\alpha\beta}$
$\varepsilon_{(ij)k}$	σ^2

If the Experiment calls for α to be Random and β fixed:

Source	EMS
α_i	$\sigma^2 + bn\,\sigma^2_{\alpha}$
β_j	$\sigma^2 + n\,\sigma^2_{\alpha\beta} + an\,\theta(\beta)$
$\alpha\beta_{ij}$	$\sigma^2 + n\,\sigma^2_{\alpha\beta}$
$\varepsilon_{(ij)k}$	σ^2

interaction, $\alpha\beta$. How these effects in your model are to be tested for significance depends upon the expected mean squares [EMS; see Anderson and McLean (1974) for a description of the EMS algorithm].

Referring to Figure 1, assume you decide that factors α and β are "fixed",that is , you wish to set each factor's levels at prescribed values and these explicitly define the extent to which you may make inferences within the design space. Recall that number 10 stated that all commercial, statistical software makes the assumption that all effects in a model are fixed. The EMS reflect this assumption in Figure 1: to do an F-test of each term in the fixed model, each mean square estimate should be divided by the mean square error , ε. Note that each effect attributed to α and β can then be uniquely estimated by this ratio of effect to error.

Note, however, if during the planning of our experiment we wished to make inferences not to a very limited number of explicit levels (and fixed contributions of α and β to the overall mean effects), but to a broader, more universal variation in all α and β (called a random effects model). The EMS would take on a different look(note that fixed treatment effects,θ, are now replaced by σ^2, denoting variance effect). Now there are two error terms we must use to test for treatment effects. The $\alpha\beta$ interaction is used as the error term to test main effects (α and β) because it shares the same EMS components, except for that contributed only by the α or β treatment effect. The random error mean square,ε, is now used only to test the significance of the $\alpha\beta$ interaction.

In the third, and final combination, called a "mixed" model, there are two factors, one fixed and one random (Figure 1). After calculating the EMS, we are able to see that the random error, ε, can be used to estimate the effect of α and the $\alpha\beta$ interaction, but the $\alpha\beta$ interaction is appropriate to test the effect due to β.

Obviously, from this simple illustration, decisions need to be made up-front as to how well the design and model conform and whether our research objective(s) can be met by the appropriate tests of terms in the model. In addition to the source already mentioned, a good explanation of these topics can be found in Snedecor and Cochran(1967, ch. 12).

Summary

This paper has provided a simple schematic outline of 12 steps that ought to be considered during the design and analysis of an experiment. If more effort is expended in the early states of a research project, the later stages — defining the model, fitting the response data, testing for treatment effects, and publicly presenting the results — can be done with greater confidence.

References

Anderson,V.L. and R.A. McLean. 1974. Design of experiments: a realistic approach. Marcel Dekker Inc., New York, 418p.

Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. Statistics for experimenters. John Wiley & Sons, New York, 653p.

Cox, D.R. 1958. Planning of experiments. John Wiley & Sons, New York, 308p.

Kimball, A.W. 1957. Errors of the third kind in statistical consulting. J. Am. Stat. Assoc. 57:153.

Mann, M.D., D.A. Crouse and E.D. Prentice. 1991. Appropriate animal numbers in biomedical research in light of animal welfare considerations. Lab. An. Sci. 41:6-14.

Snedecor, G.W. and W.G. Cochran. 1967. Statistical methods. The Iowa State Univ. Press, Ames, 593p.

PARAMETRIC ASSUMPTIONS

Bob Hardin

Department of Animal Science

University of Alberta, Edmonton, Alberta

Abstract

Statistical hypothesis testing of the linear model is based on assumptions that the data are normally distributed. In many situations, both in teaching and the real world, this is simply assumed to be true. A problem is how to incorporate into a graduate biometrics course the routine testing of these assumptions. The approach presently being incorporated into the course is the computation of residuals from the predicted linear model and the use of these residuals in the computation of sample statistics and normal probability plots. The GLM, UNIVARIATE and RANK procedures of SAS will be used to perform the computations. Examples will be presented to illustrate the approach.

STATISTICAL PROBLEMS IN COMPLIANCE ASSESSMENT

Albert J. Liem
Air and Waste Management Branch
Alberta Environmental Centre
Vegreville, Alberta

Introduction

This note is a condensed description of a paper already published elsewhere (Liem and Wilson 1991). Problem definition is emphasized, but details of the solution have been omitted, with the hope of not compromising accuracy or clarity for sake of brevity. The purpose of presenting this note is to introduce the subject the reader, who may find it either useful in other applications or at least academically interesting.

Description of the problem

The problem is quantifying the confidence of proving that regulatory compliance, given by the following equation, is met:

$$y_r < y_c \tag{1}$$

where y_r is the measured value of the regulated parameter and y_c is the compliance level. Many regulatory standards can be expressed as above, and assessing compliance is not a trivial problem when:

*the accuracy and precision of the method for measuring y_r (herein referred to as variability) are not known, and

*some or all the reported values are expressed as less than detection limits or nondetectable.[1]

Such is the case with the results of many incinerator test burns, as shown in Table I.

[1] A further complicating factor is when the detection limit is not unique, but can be varied by adopting different sampling or sample processing strategies, and hence it can be made as close as possible to the compliance level; see original paper.

In all the cases shown, the reported values are lower than the compliance levels. Is it valid to interpret such evidence as a definitive proof that compliance is met? Our contention is NO. The difference between the reported values and the compliance levels ***and*** the variability of y_r should both be taken into account. Intuitively, one expects that the larger the difference and the smaller the variability, the more assured one is that compliance is actually met.

Consider the accompanying results of a surrogate spiking program that was implemented to address the issue of variability. In the process of measuring y_r, the 'analyst' was given certain quantities (unknown to the analyst) of surrogates and requested to report those quantities. The 'recovery', defined as ratio of the reported to the actual values, ranged from 10% to 300%. Thus, on the assumption that the surrogates and the regulated compounds are similar, the actual values of y_r could be as high as ten times the reported values shown in Table I. Surely in Case C one cannot be 'very sure' that compliance is met; but, one can be 'more sure' that compliance is met in Case B. As previously stated, the problem is, therefore, how to quantify the confidence of concluding that compliance is met. It is a statistical problem.

Table I. Sample Incinerator Test Burn Results*

Case	y_r	y_c
A	2.6	46
	4.2	43
	12.8	46
B	<44	351
	<26	380
	<21	419
C	<14	26
	<21	52
	<22	44

* Actual results from the Alberta Special Waste Treatment Centre (ASWTC) near Swan Hills. Both values are expressed in mg/h of emissions of regulated compounds.

Note that it would be conceptually erroneous to use the results in Case A for computing confidence intervals, from which statistical inferences regarding compliance being met are drawn. One simple reason is that the bias or the accuracy of the method is not known from those results alone. Incidentally, that approach could not be used when nondetectable results are obtained, as in Cases

B and C.

Solution

Outline

There are three elements in the solution:

- * Obtaining the variability of the method for measuring y_r: In the case of incinerator test burns, this can be indirectly obtained from a surrogate spiking program. Other cases may require different schemes. The premise is that the surrogates and the regulated parameters are, in terms of variability, identical.
- *Treatment of nondetectable results: From the point of view of the 'regulator', to whom compliance must be proven, nondetectable results can be assigned the detection limit values. Thus, a 'conservative' approach is used.
- * Statistical analysis: The premise is that measurement results are distributional. That is, given a value of the regulated parameter, repeated measurement results can be described by a probability density function. The Bayesian approach for hypothesis testing can then be used.

The implementation of a surrogate spiking program or other programs can be quite involved and hence will not be described here. The only aspect that will be described is the statistical analysis, starting from the point where the variability of the measurements of y_r has been obtained.

Statistical Analysis

Bayes' theorem. Compliance assessment can be formulated as hypothesis testing. The first hypothesis is H_1:$y \geq y_c$ - compliance is violated - and the second is the alternative H_2:$y < y_c$ - compliance is met - where y is the actual value of the regulated parameter.

Given that a value of y_r is obtained (or a series of values of y_{rk}, where subscript k represents the k^{th} measurement), what are the probabilities of H_1 and H_2 being true? Bayes' theorem is:

$$P[H_j/y_r] \propto P[y_r/H_j]P[H_j] \quad j=1,2 \tag{2}$$

where * P[H_j] is the *a priori* probability of H_j being true, the degree of belief in H_j before the evidence is gathered,

* P[y_r/H_j], referred to as a conditional probability, is the probability of obtaining y_r if indeed H_j is true,

* P[H_j/y_r] is the *a posteriori* probability of H_j being true, the revised degree of belief in H_j

after the evidence is gathered.

The degree of belief in a hypothesis is revised by the evidence gathered and quantitatively modified in proportion to the probability of that evidence being obtained if indeed the hypothesis is true. This revision can be continually updated in a series of measurements, where the *a posteriori* probability of one set of measurements is used as the *a priori* value for the next.

The proportionality constant in Eq. (2) can be eliminated by noting that H_1 and H_2 are complementary, or by using the concept of **odds R**, defined by the ratio of the two probabilities. Eq. (2), written in terms of odds, becomes:

$$R[C/y_r] = R[y_r/C]\,R[C] \qquad (3)$$

where C denotes compliance being met, that is, $R[C]=P[H_2]/P[H_1]$ and the conditional and *a posteriori* odds are similarly defined.

Two inputs are thus required. The first is the *a priori* odds of compliance being met. A value of one, expressing no prior knowledge whether compliance is more or less likely to be met, seems a reasonable value. The second is the conditional probability, which requires statistical models or assumptions and as described below.

Assumptions. The following assumptions are needed, but the actual 'models' or equations used can be changed to suit the situation:

(1) Probability density function of obtaining y_r given that the actual value is y. A simple function is log-normal, with constant variance:

$$p(Y_r) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(Y_r - Y)^2}{2\sigma^2} \qquad (4)$$

where ^{2} is the variance (derived from surrogate spiking results) and the upper case Y denotes log-transformed values.

(2) The density function of the *a priori* probability or odds. The simplest function is a constant density.

Computation of conditional probability. Consider the conditional probability of obtaining y_r under H_1, that is compliance is not met, or $y \geq y_c$. This can be readily computed for the 'models' used in the above assumptions. It is simply the integral of Eq. (4) with respect to Y, from $Y=Y_c$ to $Y=\infty$.

$$P[y_r/H_1] = \frac{1}{\sigma\sqrt{2\pi}} \int_{Y_c}^{\infty} \exp - \frac{(Y_r - Y)^2}{2\sigma^2}\, dY \qquad (5)$$

It can verified that

$$P[y_r/H_1] = \Phi(\frac{Y_r - Y_c}{\sigma}) = \Phi(Z_r) \qquad (6)$$

where (Z$_r$) is the 'tail area' of the standard normal distribution (zero mean and unit variance) to the left of Z_r=(Y_r-Y_c)/ .

***A posteriori* odds of compliance being met.** In a series of measurements the degree of belief is continually updated as more evidence is gathered. Thus at the end of M sets of measurements, the odds of compliance being met is:

$$R_M[C/y_{rM}] = (\prod_{j=1}^{M} R_j[y_{rj}/C])\ R_1[C] \qquad (7)$$

where the conditional odds in the jth measurement is:

$$R_j[y_{rj}/C] = \frac{1 - \Phi(Z_{rj})}{\Phi(Z_{rj})} \qquad (8)$$

As discussed previously, R_1[C]=1 is a reasonable value, the values of , Y_c and Y_r are obtained from the surrogate spiking and test burn results. can be found from statistical tables and hence the final *a posteriori* odds of compliance being met can be readily computed.

Results and Discussion

The method was applied to a series of test burns conducted at the ASWTC. A complete presentation would be too lengthy since there were 'complications', such as the presence of two emission sources. The condition for compliance can no longer be expressed as given in Eq. (1), but is in the following form:

$$\sum_{s=1}^{NS} \omega_s y_{r,s} < y_c \qquad (9)$$

where subscript s denotes the sth emission source, NS is the number of emission sources and $_s$ is the weighting factor for emission source s.

For simplicity, therefore, only selected cases will be presented, and the data 'variability' will be qualitatively expressed by the range of surrogate recovery values in each emission source. The results are shown in Table II.

Table II. Summary of Results

Case	Variability,% †		Test Results‡ k_T	Conditional Odds	A posteriori Odds
	MS	CT			
A	83-120	71-140	18	>2 10^5	>8 10^{13}
			10	>2 10^5	
			3.6	2000	
B	34-290	36-275	18	2.9	36
			15*	3.3	
			20*	3.7	
C	77-130	30-330	1.8*	1.3	2.7
			2.5*	1.5	
			2.5*	1.3	
D	54-185	50-200	0.9	0.9	1
			0.4	0.4	
			3.2	2.7	

Notes: † Range of surrogate recovery values, MS and CT are the two emission sources; ‡ Results are expressed as ratio of compliance level to reported value, thus values of <1 correspond to evidence of compliance being violated, * denotes nondetectable results; The conditional odds are for each set of measurements, and the a posteriori odds are for all sets.

The following interesting features can be noted:

*In Case A, the evidence of compliance being met is so overwhelming that a loss of one set of data would not really matter.. There is no need to prejudge and nullify the whole effort, the remaining evidence may still be sufficiently convincing.

*Case B shows that even if the performance of the 'analyst' was less than desirable, the results for the purpose of proving compliance may still be satisfactory (the *a posteriori* probability of compliance being met is 97%). The explanation is the reported values were much lower than the compliance levels.

*Case C shows that (i) even nondetectable results do not provide a convincing proof of compliance (*a posteriori* probability of 70% as compared to the *a priori* value of 50%), and (ii) when the reported values are close to the compliance levels, it is necessary to have small data variability.

*Case D shows that evidence against compliance (k_T<1) can also be included. By coincidence, the *a posteriori* and the *a priori* odds are identical.

Concluding Remarks

The method that has been developed quantifies what is intuitively expected. Data quality should not be judged by itself, but by the intended use. What is acceptable in one case may not be so in another case. The Bayesian approach can deal with nondetectable results in a logical manner and, in the author's opinion, it produces results that can be readily understood.

Incinerator test burns represent only one example of compliance assessment, which in many cases can be formulated as Eq. (1), or more generally, as Eq. (9). Different means of obtaining variability and different statistical models may be needed, but the same approach can be used in other similar cases.

Reference

Liem, A.J. and M.A. Wilson, "A quantitative method for evaluating incinerator test burn results", *J. Air Waste Manage. Assoc.*, Vol 41, No. 1, Jan. 1991: 47-55.

ENVIRONMENTAL CHEMICAL ANALYSIS

I. Johnson and Y. Kumar
Research and Methods Development Branch, Chemistry Division
Alberta Environmental Centre, Alberta Environment,
Vegreville, Alberta

Abstract

Pollutants from many point sources are complex mixtures of chemical compounds. Correlation of pollutants in environmental samples with point sources can be difficult as the relative amounts of chemical compounds in the mixtures may be altered through evaporation, chemical and biological degradation. The multivariate analysis based procedure presented here is specifically directed to the analysis of polychlorinated biphenyls (PCBs) although it may be applied to many other analytes.

A multivariate analysis based procedure for identification and quantification of complex mixtures of (PCBs) as in industrial and environmental samples is presented. There are 209 individual PCB compounds (congeners). Aroclors, industrial preparations of PCBs, are complex mixtures, comprising of as many as 60 separate congeners. Analysis of PCB samples by capillary gas chromatography results in complex chromatograms with many peaks. Identification and quantification of PCBs as Aroclors becomes difficult when more than one Aroclor are present. Identification and quantification can be further complicated by chemical or biological degradation which changes the relative amounts of congeners present.

This method uses a calibration solution of 36 individual PCB congeners. Authentic Aroclors and samples are analyzed using this calibration mixtures. Multivariate analysis of these results is used to identify and quantify PCBs in the samples as Aroclors. The use of operators to describe PCB degradation is also discussed.

RESPONSE SURFACE ANALYSIS

Robert B. Heimann
Materials Section
Alberta Research Council
Edmonton, Alberta

The experimental environment

According to Tukey [1], industrial experiments can be classified by their depth of intellectual investment as (i) confirmation experiments, (ii) exploration experiments, and (iii) fundamental or "stroke-of-genius" experiments. A second way of classification is based on the distance of their objectives from the real world, i.e. from the market [2]. Finally, the continuity of factors provides a third classification. If the factors are continuous variables[1] and controllable at preset levels, then the response surface methodology is the method of choice. If , however, many factors are orderable but not measurable, i.e. at discrete levels , the response surface analysis becomes less useful and should be replaced by nested or split-plot designs [3].

Every experiment attempts to approximate the real world but must avoid by a set of simplifying assumptions the complex interactions occurring in real systems. There are, in principle, two ways to accomplish this: the "classical" experimental strategy that varies one parameter at a time but attempts to keep all others constant, and the statistical experimental strategy that varies parameters simultaneously to obtain a maximum of information with a minimum of experiments. The classical experimental strategy yields accurate results but requires many experiments. It gives, however, misleading conclusions to problems that have synergistic parameter interactions, and also fails to elucidate the "structure" of a system. Table 1 compares these strategies [4].

[1] In this paper, the terms "variable", "parameter" and "factor" will be used simultaneously.

Table 1: Two Viewpoints of the 'Real World"

	CLASSICAL	STATISTICAL
Number of runs	many	few
Response	complex	simple
Synergism	absent	present
Error	small	large(+)
Strategy	one-factor-at-a-time	factorial
Thought pattern	vertical	lateral [5]

(+) Bias errors can be considered by blocking and randomization; random errors can be accounted for by replication of experiments.

The evolution of the experimental environment usually starts with a screening (Plackett-Burman) design [6] with many independent (up to 40) variables. It yields a crude prediction of the ranking of importance of parameters through a first-order polynomial model. The experimenters should list and investigate all possible parameters they can think of but should refrain from skipping some because of "folklore", laboratory gossip, or preferences and hunches. "Be bold but don't be stupid!" [4]. The tremendous reduction in the number of required experiments, however, will be offset by the failure to detect synergistic interactions between parameters. On the other hand, an advantage of the screening designs is that they can accommodate a mix of continuous and discrete parameters.

With the independent parameters (up to eight) identified to influence the response of the dependent parameter(s), "limited response surface" experiments should be run such as full two-level factorials, or even a fractional three-level (Box-Behnken) design [7] that yields higher quality predictions by allowing interpolation within the experimental space by a second-order polynomial model. Such a model determines non-linear behaviour, i.e. the curvature of the response surface, and thus permits the estimation of synergistic parameter interactions.

The polynomial models approximate the "true" response surface only in the necessarily narrow region of the investigated parameter space. Thus, any extrapolation beyond the proven validity of the predictions is dangerous and may lead to useless or even nonsensical results. To avoid this, eventually a theoretical model has to be built [8] that yields the exact mathematical response surface, usually by the application of first-order differential equations.

EXAMPLE 1: FRACTIONAL TWO-LEVEL FACTORIAL DESIGN 2^{8-4}

In this example, the thickness of 88WC12Co alloy coatings should be optimized. These wear-resistant coatings are being applied to carbon steel surfaces by plasma spray technology [9]. The parameters selected for the fractional two-level factorial screening design are shown in Table 2, the randomized design is shown in Table 3.

Table 2: Parameters and Parameter Levels for Example Design

Variable	X_i	"-"	"+"	Type of Variable
Plasma Arc Current	X_1	700 amps	900 amps	Continuous
Argon Gas Pressure	X_2	0.34 MPa	1.36 MPa	Continuous
Helium Gas Pressure	X_3	0.34 MPa	1.36 MPa	Continuous
Powder Gas Pressure	X_4	0.34 MPa	0.68 MPa	Continuous
Powder Feed Rate	X_5	low (0.5)	high (2)	Discrete
Powder Grain Size	X_6	$(-45+5)\mu$	$(-75+45)\mu$	Continuous
Number of Passes	X_7	20	30	Continuous
Spray Distance	X_8	25 cm	45 cm	Continuous

Table 3: Randomized Fractional Two-Level Factorial Design

Run #	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)	X(7)	X(8)	Response Y (μ)	σ (μ)
1	700	0.34	0.34	0.68	2	Coarse	20	45	118	79
2	900	0.34	0.34	0.34	0.5	Coarse	30	45	16	8
3	700	1.36	0.34	0.34	2	Fine	30	45	203	111
4	900	1.36	0.34	0.68	0.5	Fine	20	45	57	25
5	700	0.34	1.36	0.68	0.5	Fine	30	45	82	35
6	900	0.34	1.36	0.34	2	Fine	20	45	138	87
7	700	1.36	1.36	0.34	0.5	Coarse	20	45	30	12
8	900	1.36	1.36	0.68	2	Coarse	30	45	82	44
9	900	1.36	1.36	0.34	0.5	Fine	30	25	7	4
10	700	1.36	1.36	0.68	2	Fine	20	25	108	104
11	900	0.34	1.36	0.68	0.5	Coarse	20	25	65	30
12	700	0.34	1.36	0.34	2	Coarse	30	25	16	10
13	900	1.36	0.34	0.34	2	Coarse	20	25	26	21
14	700	1.36	0.34	0.68	0.5	Coarse	30	25	9	12
15	900	0.34	0.34	0.68	2	Fine	30	25	30	13
16	700	0.34	0.34	0.34	0.5	Fine	20	25	22	19

This 2^{8-4} fractional factorial design is a 1/16 replicate of a full 2^8 factorial. It has a resolution IV [8] and is able to estimate the eight main effects X_i clear of composite two-factor interactions E_i (Table 6). The effects of higher-order interactions can usually be safely neglected. Composite effects of the sum of four two-factor interactions, however, can be estimated from the unassigned factors. If no interactions exist, the effects of unassigned factors can be used to estimate the experimental error, i.e. the minimum significant factor effect. The arrangement of the parameter levels in the design matrix follows Yates's standard order.

Tables 4 and 5 show the numerical evaluation of the results. First, the sum S(+) of all responses on the "+"-level is calculated. Then the sum S(-) of all responses on the "-"-level is calculated. The factor effect is the difference D of the two sums, divided by the number of "+" signs in each column. The coefficient C of the parameter in the response equation is the factor effect divided by two. The factor significance is checked against the minimum factor significance, FE(min) = s(FE) * t(a,df), where s(FE) = $(1/n\ SE^2)^{1/2}$, t(a,df) is the Student t-value for a confidence level a of a double-sided significance test and df degrees of freedom. All absolute factor effects larger than FE(min) are considered to be significant.

Table 4: Computing of Main Factor Effects

Run#	Main	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)	X(7)	X(8)	Y
1	+	-	-	-	+	+	+	-	+	118
2	+	+	-	-	-	-	+	+	+	16
3	+	-	+	-	-	+	-	+	+	203
4	+	+	+	-	+	-	-	-	+	57
5	+	-	-	+	+	-	-	+	+	82
6	+	+	-	+	-	+	-	-	+	138
7	+	-	+	+	-	-	+	-	+	30
8	+	+	+	+	+	+	+	+	+	82
9	+	+	+	+	-	-	-	+	-	7
10	+	-	+	+	+	+	-	-	-	108
11	+	+	-	+	+	-	+	-	-	65
12	+	-	-	+	-	+	+	+	-	16
13	+	+	+	-	-	+	+	-	-	26
14	+	-	+	-	+	-	+	+	-	9
15	+	+	-	-	+	+	-	+	-	30
16	+	-	-	-	-	-	-	-	-	22
Σ(+)	1009	421	522	528	551	721	362	445	726	
Σ(-)	0	588	487	481	458	288	647	564	283	
ΣΣ	1009	1009	1009	1009	1009	1009	1009	1009	1009	
Δ	1009	-167	35	47	93	433	-285	-119	443	
Effect	63	-21	4	6	12	54	-36	-15	55	
C	32	-11	2	3	6	27	-18	-8	28	

Table 5: Computing of Composite Two-factor Interactions

Run#	E(1)	E(2)	E(3)	E(4)	E(5)	E(6)	E(7)	Y
1	+	+	-	-	-	+	-	118
2	-	-	-	-	+	+	+	16
3	-	+	+	-	+	-	-	203
4	+	-	+	-	-	-	+	57
5	+	-	-	+	+	-	-	82
6	-	+	-	+	_	-	+	138
7	-	-	+	+	-	+	-	30
8	+	+	+	+	+	+	+	82
9	+	+	-	-	-	+	-	7
10	-	-	-	-	+	+	+	108
11	-	+	+	-	+	-	-	65
12	+	-	+	-	-	-	+	16
13	+	-	-	+	+	-	-	26
14	-	+	-	+	-	-	+	9
15	-	-	+	+	-	+	-	30
16	+	+	+	+	+	+	+	22
Σ(+)	410	644	505	419	604	413	448	
Σ(-)	599	365	504	590	405	596	561	
ΣΣ	1009	1009	1009	1009	1009	1009	1009	
Δ	-189	279	1	-171	199	-183	-113	
Effect	-24	35	0	-21	25	-23	-14	
C	-12	18	0	-11	13	-12	-7	

The pattern of the "+" and "-" levels again corresponds to Yates's standard order. Note, that the second half of the main effect design matrix is the mirror image of the first half, and that the two halves of the composite two-factor interaction design matrix are identical. The confounding pattern of the composite two-factor interactions E_i is shown in Table 6.

Table 6: Confounding Pattern of Composite Two-factor Interactions

$$\begin{array}{lll} E(1) = E_i = X_1X_2 + X_3X_7 + X_4X_8 + X_5X_6 & = & 12 + 37 + 48 + 56 \\ E(2) & = & 13 + 27 + 58 + 36 \\ E(3) & = & 14 + 28 + 36 + 57 \\ E(4) & = & 15 + 38 + 26 + 47 \\ E(5) & = & 16 + 78 + 34 + 25 \\ E(6) & = & 17 + 23 + 68 + 45 \\ E(7) & = & 18 + 24 + 35 + 67 \end{array}$$

From Table 5 the minimum factor effect, FE(min) can be calculated as follows.

$$\sigma(FE) = (1/n\ \Sigma E^2)^{1/2} = (3592/7)^{1/2} = 22.6 \qquad [1]$$

$$FE(min) = \sigma(FE) * t(\alpha{=}0.90, df{=}7) = 22.6 * 1.895 = 43. \qquad [2]$$

Thus, all factor effects whose absolute values are larger than 43 are significant at a confidence level of 90%. From Table 4 it follows that X_5 (powder feed rate) and X_8 (spray distance) are the only significant main factor effects. This holds true even when the confidence level is increased to 95%. In this case, the minimum factor effect is FE(min) = 22.6 * 2.36 = 53. There are no significant composite two-factor interactions (Table 5). Both main factor effects have positive signs, i.e. the thickness of the coating increases with increasing powder feed rate and increasing spray distance. Short spray distances lead to overheating of the alloy powder thus causing thermal decomposition and reaction of the tungsten carbide with the cobalt matrix. This will eventually result in brittle phases, loss of carbon, and higher coating porosities [10].

The response polynomial of the thickness of plasma sprayed 88WC12Co alloy coatings can be roughly (zero-order approximation) expressed by the equation

$$d\ [\mu] = 32 + 27\ X_5 + 28\ X_8. \qquad [3]$$

EXAMPLE 2: FULL TWO-LEVEL FACTORIAL DESIGN 2^4

The estimation of radioactive source terms for the safety analysis of a nuclear fuel waste repository involves laboratory leaching experiments to determine the durability of used UO_2 fuel and fuel recycle waste glass under conditions relevant to the disposal of these highly radioactive materials deep in a granitic pluton of the Canadian Shield [11]. Table 7 shows the nine parameters selected for leaching in two different groundwaters of used UO_2 fuel and a borosilicate glass containing 90-Sr, 137-Cs and several actinides such as 239-Pu, 241-Am and 244-Cm. In this example, only the responses of the amount of hydrogen formed by radiolysis of the groundwater and of the normalized mass loss of strontium will be examined.

Table 7: Parameters and Parameter Levels for Example Design

Variable	X_i	"-"	"+"	Type
Waste Form	X_1	Glass	UO_2 fuel	Discrete
CEC of Clay	X_2	160 meq/kg	1350 meq/kg	Continuous
Ionic Strength of Groundwater	X_3	10^{-4} mol/L	1.4 mol/L	Continuous
Surface Area/ Volume Ratio	X_4	12 m^{-1}	120 m^{-1}	Continuous
Metal	X_5	Ti-grade 12		Constant
Rock	X_6	Granite		Constant
Pressure	X_7	10 MPa		Constant
Temperature	X_8	200°C		Constant
Time	X_9	6 months		Constant

The geometrical representation of such a 2^4 design is a 4D-hypercube as shown in Figure 1. The four selected variables are the four axes of this hypercube.

With the technique executed in detail in the previous example, the factor effects and the coefficients of the response equations were calculated. The amount of hydrogen formed during g-radiolysis of

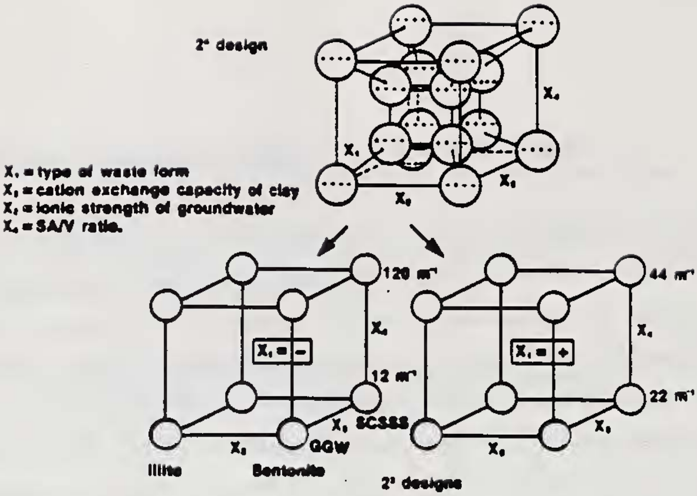
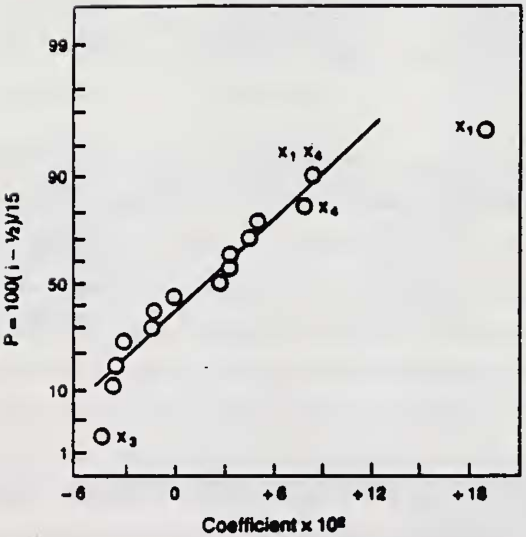


FIGURE 1: 4D-Hypercube as a geometrical representation of a 2^4 design.



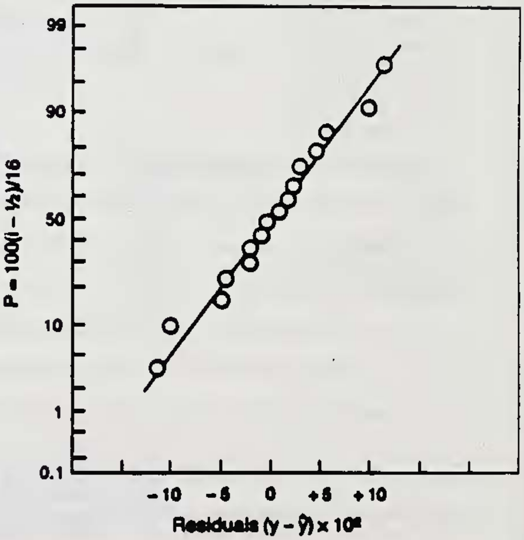


FIGURE 2: Empirical cumulative distribution of the coefficients of a first-order polynomial for the amount of hydrogen in vol% developed by γ-radiolysis of groundwater (left).

FIGURE 3: Empirical cumulative distribution of the residuals (amount of hydrogen) (right).

the groundwater can be described by the polynomial

$$[H_2] * 10^2 = 27 + 19\ X_1 + 8\ X_4 + 8\ X_1X_4 - 4\ X_3 \quad \text{(in Vol\%)}. \qquad [4]$$

The amount of hydrogen formed is a strong positive function of the activity of the waste form, X_1, and the ratio of the surface area of the solid to the volume of the solution, X_4. It is negatively correlated with the ionic strengths of the groundwater, X_3. There is also a rather strong two-factor interaction X_1X_4. Figure 2 shows the empirical cumulative distribution of the 15 calculated coefficients of the first-order polynomial on a probability net. If all the coefficients would randomly fluctuate in a Gaussian fashion around a statistical mean value than the distribution would follow exactly a straight line. Deviations from this line signify non-random factor effects, in this case X_1 ,X_3 and X_4. With the response equation shown above the residuals were calculated and also plotted on a probability net (Figure 3). The figure indicates a reasonable qualitative fit of the assumed model to the true response surface.

The normalized mass loss of 90-Sr is shown in Figure 4 in a 4D-hypercube. The four axes of the hypercube are the selected parameters X_1 to X_4 (see Figure 1). The inner cube contains the data obtained by leaching of used fuel, the outer cube contains the data obtained by leaching of fuel recycle waste glass. The complex response equation is

$$\begin{aligned} NML(Sr)*10^3 = 42 - 32\ X_1 - 17\ X_3 - 12\ X_4 + 16\ X_1X_2 + 21\ X_1X_3 \\ + 15\ X_1X_4 + 11\ X_2X_3 \quad [kg/m^2]. \qquad [5] \end{aligned}$$

The empirical cumulative distributions of the residuals for NML(Sr), and NML(Cs), are shown in Figure 5.

It was suspected that all two-factor interactions but X_2X_3 are merely perturbations of the parameters X_2, X_3 and X_4. The interaction of the cation exchange capacity of the clays, and the ionic strength of the groundwaters, however, determines the pH of the solution. With this, a table of factor assignment can by produced (Table 8).

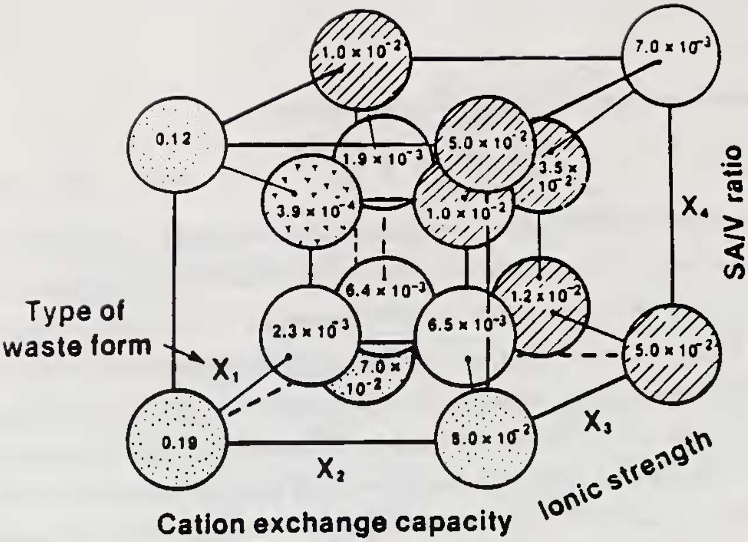
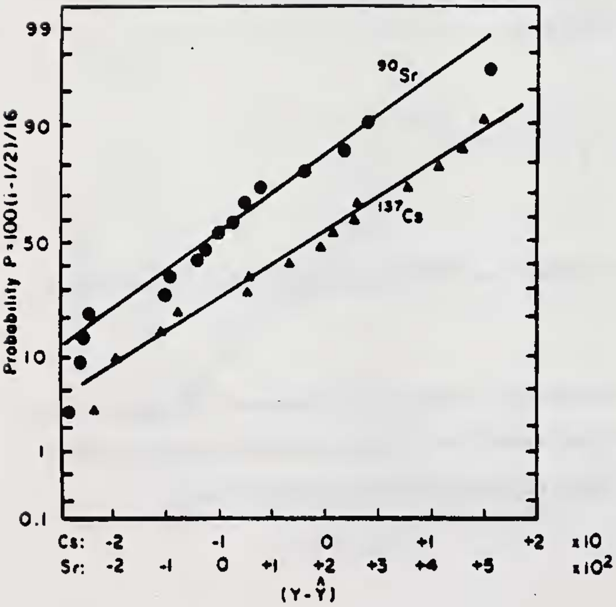


FIGURE 4: Normalized mass loss of 90-Sr in a 4D-hypercube formalism



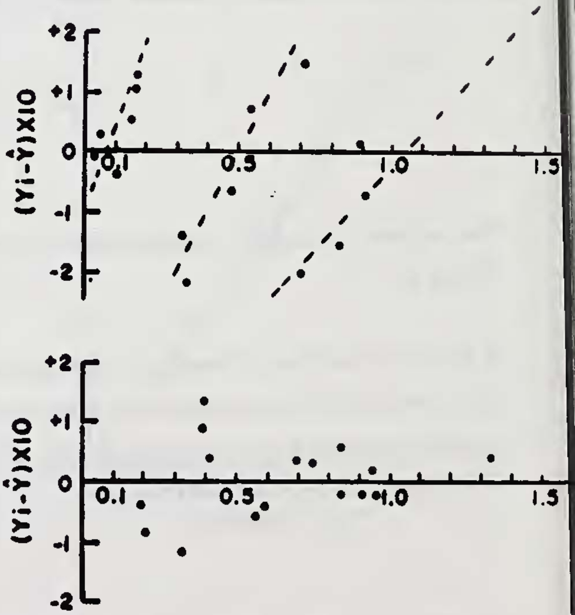


FIGURE 5: Empirical cumulative distribution of the residuals of the normalized mass losses of 90-Sr and 137-Cs.

FIGURE 6: Anscombe-Tukey-plots of residuals vs. linear (top) and parabo (bottom) normalized mass loss of 90-Sr.

Table 8: Order, Signs and Assignment of Factors, and Their Effect on the Normalized Mass Loss of Strontium-90

Factor	Sign	Assignment	Effect
X_1	-	Waste Form	Fuel shows smaller NML(Sr) than glass
X_3	-	Ionic Strength	Increasing I decreases NML(Sr)
X_1X_2	+	Perturbation of X_2	
X_1X_4	+	Perturbation of X_4	
X_1X_3	+	Perturbation of X_3	
X_4	-	SA/V ratio	Increasing SA/V decreases NML(Sr)
X_2X_3	+	pH	Increasing pH increases NML(Sr)

The three-factor interaction $X_1X_2X_3 = X_4X_5$ is suspiciously large thus suggesting the involvement in the leaching process of corrosion products due to interaction of the SA/V ratio and the type of metal.

The plot of the residuals vs. the response Y (Figure 6, top), also called Anscombe-Tukey plot, shows three straight lines instead of the random distribution of data points above and below the zero line as required for a good fit of the data to the model. This suggests a Y-data transformation to achieve better fit to the model.

Indeed, several data transformations increase the regression coefficient (coefficient of determination)[1] from 0.928 for Y (linear) to 0.960 for $Y^{1/3}$ to 0.963 for $Y^{1/2}$, considering the seven parameters identified above. Figure 6 (bottom) shows random scatter of the data points for the $Y^{1/2}$ transformation. This could mean that the true response surface has a parabolic or cubic character. However, the selected experimental design does not allow for non-linear parameters to determine the curvature of the response surface. A different model must be built, and additional experiments must be run.

[1] R^2=SSREG/SSTOT; SSREG = Sum of squares of regression, i.e. fraction of the total scatter of the original observations that is accounted for by the selected effects. SSTOT = Total sum of squares of deviation of the observations about their mean.

EXAMPLE 3: THREE-LEVEL FRACTIONAL FACTORIAL (BOX-BEHNKEN) DESIGN

This example deals with the dissolution of a simulated borosilicate-based nuclear waste glass in the presence of three different clays and three different groundwaters [12]. Three parameters were varied at three levels (Table 9). This is the minimum number of levels for each parameter to estimate non-linear responses. Designs with more than three levels yield higher quality predictions such as the five-level central composite [13] or the Box-Wilson [14] designs but require many more experiments. A Box-Behnken design is a subset of a full three-level factorial, 3^{3-f} that uses 13 of the 27 points of the full factorial plus 2 extra replicates at the centre. There are 5 more points than the minimum 10 points required to estimate the 10 parameters of the second-order polynomial (3 linear, 3 quadratic, 3 two-factor interactions, 1 three-factor interaction). Thus the design provides 5 degrees of freedom for error. Geometrically the Box-Behnken design can be described by an edge-centred cube with three centre points (Figure 7).

Table 9: Parameters and Parameter Levels for Example Box-Behnken Design

Variable	"-"	"0"	"+"
Cation Exchange Capacity of Clay, X_1	160 meq/kg	490 meq/kg	1360 meq/kg or 820 meq/kg($)
Ionic Strength of Groundwater, X_2	10^{-4} mol/L	0.013 mol/L	0.07 mol/L
Ratio Clay to Groundwater, X_3	0.01 kg/L	0.05 kg/L	0.10 kg/L

($) The "+" level was split between Ca-montmorillonite (CEC= 1360 meq/kg), and Na-montmorillonite (CEC=820 meq/kg).

Figure 7 shows the experimental results. Responses measured were the specific mass loss of the glass, the mass of dissolved silicon, the mass of dissolved boron, all in kg/m^2, as well as the final pH of the leach solution measured at room temperature. The right-hand plane of the Box-Behnken design cube shows both the results of sub-design A (using Ca-montmorillonite) and B (using Na-montmorillonite). Figure 8 shows the calculated response surfaces of the specific mass losses of the glass as the function of the coded levels of the cation exchange capacity,X_1 and the ratio

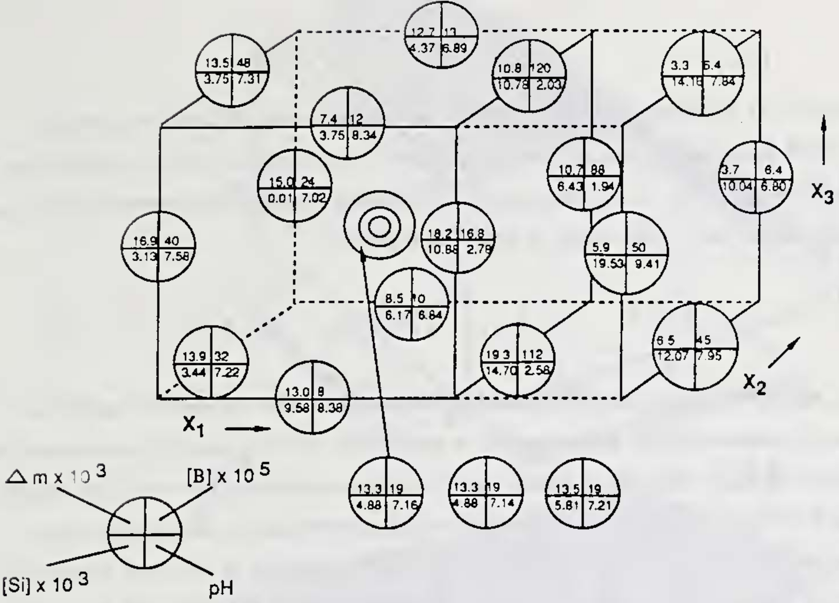


FIGURE 7: Box-Behnken design cube with experimental results of normalized mass loss of boron (1st quadrant), pH (2nd quadrant), normalized mass loss of silicon (3rd quadrant) and specific mass loss of the glass (4th quadrant) for sub-designs A and B (see text).

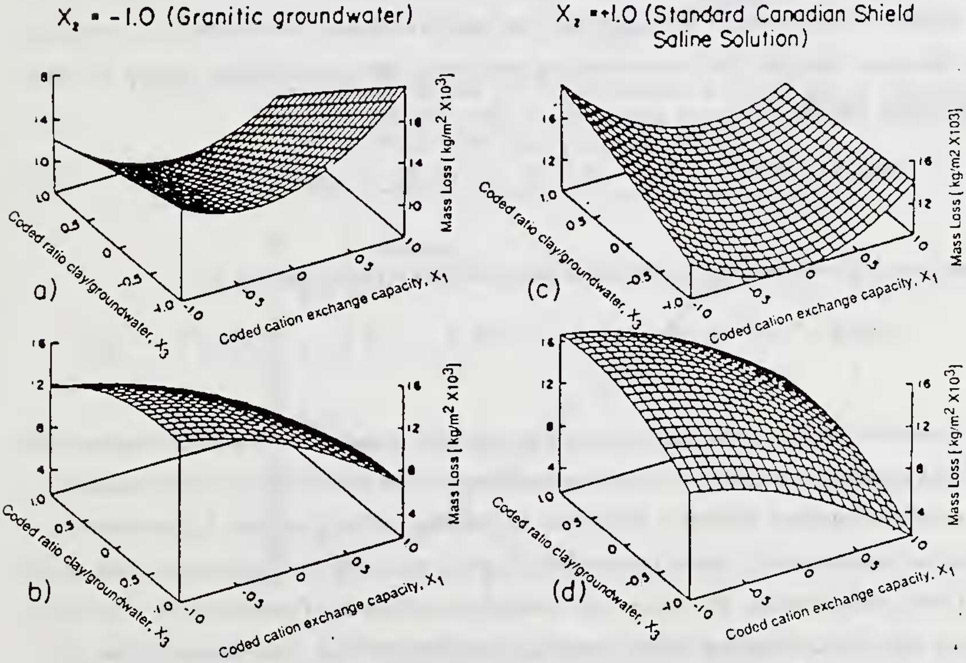


FIGURE 8: Response surfaces of specific mass losses of the glass as a function of the cation exchange capacity of the clay, X_1 and the ratio clay/groundwater, X_3 for constant ionic strengths of the groundwater, X_2.

clay/groundwater,X_3 at constant ionic strengths of the groundwater (left column: granitic groundwater, I=10^{-4} mol/L; right column: saline solution, I=0.07 mol/L; top row: Sub-design A with Ca-montmorillonite at "+" level; bottom row: Sub-design B with Na-montmorillonite at "+" level). The specific mass loss of the glass in kg/m^2 is given by

$$\Delta m * 10^3 = 12.7 + 3.2 X_1^2 - 1.3 X_3 + 2.5 X_2X_3 - 2.1 X_1X_3 - 1.4 X_1X_2. \quad [6]$$

From Figure 8 it appears that the use of Na-montmorillonite as buffer material to isolate the nuclear fuel waste containers from groundwater is preferable over the use of Ca-montmorillonite. The latter interacts strongly with groundwater, and produces exceptionally low pH values that promote the attack of the waste glass as shown by the high specific mass losses of the glass in contact with Ca-montmorillonite (Figures 8a and 8c). The interaction of clay and groundwater can also be seen in Figure 9. Whereas Ca-montmorillonite produces pH-values of 3 in granitic groundwater (X_2 = -1) and even lower values in saline solution (X_2 = +1) (Figure 9a), the Na-montmorillonite produces pH-values of 9 in granitic groundwater and around 7 in saline solution (Figure 9b).

Figure 10 shows the probability plot of the coefficients of the response equation of the normalized mass loss of silicon from the glass. It can be seen that the three linear parameters X_1, X_2 and X_3 do not fit the Gaussian straight line thus indicating significant parameter effects. Indeed, the linear response equation yields

$$NML(Si) * 10^3 = 5.5 + 4.0 X_1 - 1.3 X_2 - 1.4 X_3 \quad [kg/m^2]. \qquad [7]$$

On the other hand, the normalized mass loss of boron follows a parabolic rate law:

$$NML(B) * 10^5 = 13 + 61 X_1^2 + 39 X_1 - 13 X_1X_2 \quad [kg/m^2]. \quad [8]$$

Figure 11 shows the contour lines that indicate a trough with negative slope with increasing ionic strength of the groundwater due to the negative coefficient of the two-factor interaction parameter X_1X_2. The parabolic contours are due to the strong contribution of the quadratic X_1 parameter. Figure 12 shows the probability plot of the coefficients of the second-order response equation of the normalized mass loss of boron. The linear and quadratic coefficients of parameter X_1 due not fit the Gaussian line thus indicating effects significantly different from experimental noise. A simple variance test confirms that the fitted response surface was estimated with sufficient

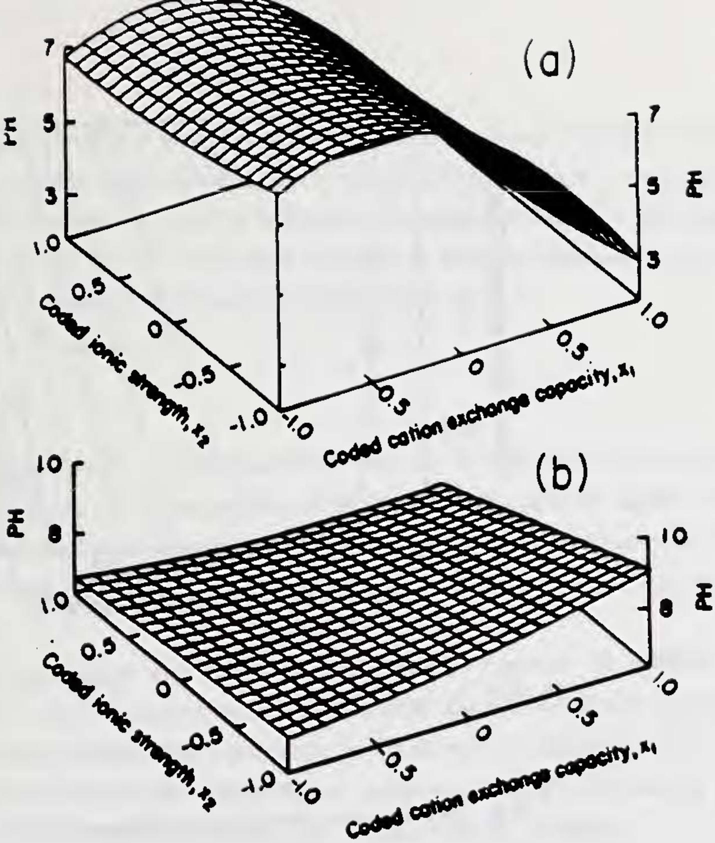


FIGURE 9: Development of pH for sub-designs A (top) and B (bottom) as a function of the cation exchange capacity of the clay, X_1 and the ionic strengths of the groundwater, X_2.

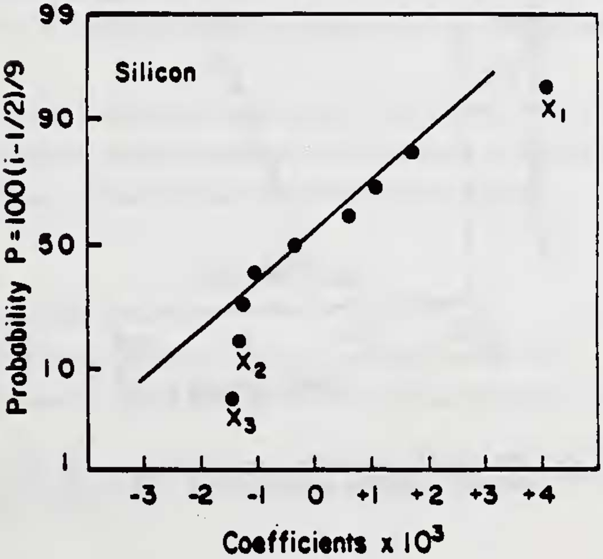


FIGURE 10: Empirical cumulative distribution of the coefficients of the second-order polynomial for the normalized mass loss of silicon.

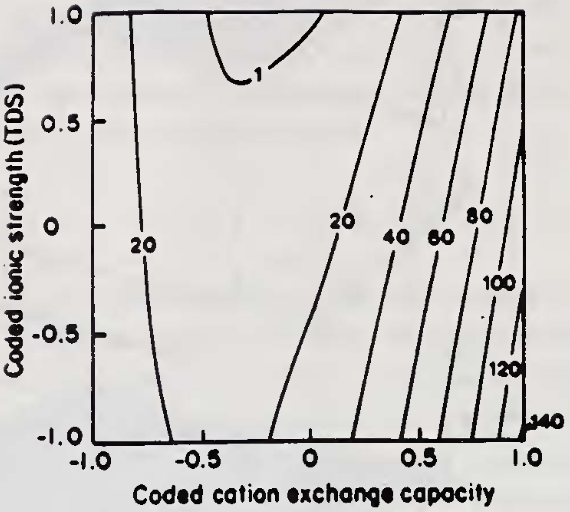


FIGURE 11: Contour lines (isopleths) of the normalized mass loss of boron in the X_1-X_2 plane.

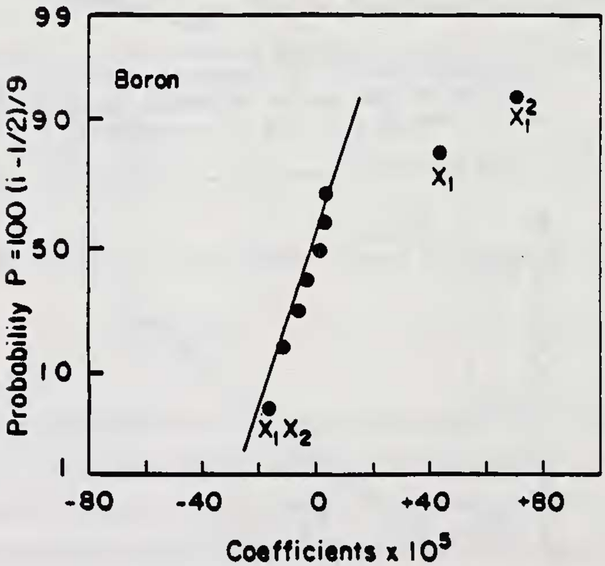


FIGURE 12: Empirical cumulative distribution of the coefficients of the second-order polynomial for the normalized mass loss of boron.

precision using the equation $V(Y) = p*\sigma^2/n$, where p = number of parameters fitted, σ^2 = estimate of error variance of the replicated centre point, and n = number of experiments. Accordingly, $V(Y) = (3)(5.33*10^{-5})/15 = 1.066*10^{-5}$, and $\sigma = 1.032*10^{-5}$. Figure 11 shows that the fitted Y data range approximately from $20*10^{-5}$ to $120*10^{-5}$. Thus we have failed to show any substantial lack of fit. The predicted change of Y is indeed 97 times the standard error of Y.

Conclusion

The response surface methodology is a convenient and easy tool to estimate how a particular response is affected by a given set of independent variables over some specific region of interest. Furthermore, it allows to determine those values of input variables at which a particular response is maximized or minimized , and gives also information on the character of the response surface close to the extremum.

One should always try, if time and resources permit, to develop the full experimental strategy by proceeding from screening (Plackett-Burman design or fractional two-level factorial) to "limited response surface" analysis (full two-level factorial) to "response surface" analysis per se (Box-Behnken design or other three-level factorials) to "exact model" building.

Many aspects of the response surface methodology could not be dealt with in this context such as orthogonal blocking, detailed analysis of variance, and canonical analysis. These subjects are described in more detail in numerous textbooks, for example in Box, Hunter and Hunter [7].

It should be emphasized again that the methodology described in this article is just a convenient tool, not gospel. Common sense, critical judgement, even Ockham's rasor and the "KISS" strategy have to be applied at all stages of the analysis to yield meaningful results.

REFERENCES

[1] Tukey,J.W. (1962): The future of data analysis. Ann.Math.Stat.33:1-67.

[2] Daniel,C. (1976): Applications of Statistics to Industrial Experimentation. John Wiley & Sons. New York.

[3] Cochran,W.G. and Cox,G.M. (1957): Experimental Designs. 2nd ed., John Wiley and Sons. New York.

[4] DuPont de Nemours & Co. (1975): Strategy of Experimentation.

[5] De Bono,E.(1970): Lateral Thinking. Ward Lock Education.

[6] Plackett,R.L. and Burman,J.P. (1946): The design of optimum multifactorial experiments. Biometrika **33**:305-325.

[7] Box,G.E.P. and Behnken,D.W. (1960): Some new three level designs for the study of quantitative variables. Technometrics **2**:455-475.

[8] Box,G.E.P., Hunter,W.G. and Hunter,J.S. (1978): Statistics for Experimenters. John Wiley & Sons. New York.

[9] Heimann,R.B., Lamy,D. and Sopkow,T. (1990): Optimization of vacuum plasm arc spray parameters of 88WC12Co alloy coatings using a statistical multifactorial design matrix. J.Can.Ceramic Soc. **59**(3):49-54.

[10] Chandler,P.E. and Nicoll,A.R. (1987): Plasma Sprayed Tungsten Carbide-Cobalt Coatings. Proc.2nd Intern.Conf.on Surface Engineering. June 16-18, 1987. Stratford-upon-Avon, U.K. pp. 403-411.

[11] Heimann,R.B. (1987): Multicomponent systems tests on used fuel and fuel recycle waste glass at 200°C. Mat.Res.Soc.Symp.Proc. **84**:409-420.

[12] Heimann,R.B. (1987): A Statistical Approach to Evaluating Durability of a Simulated Nuclear Waste Glass. In: Geological Disposal of High Activity Radioactive Wastes (ed.D.G. Brookins). Theophrastus Publications S.A., Athens, Greece, pp. 181-206.

[13] Hill,W.J. and Hunter,W.G.(1966): A review of response surface methodology: A literature survey. Technometrics **8**:571.

[14] Box,G.E.P. and Wilson,K.B.(1951): On the experimental attainment of optimum conditions. J.Roy.Stat.Soc.,Ser.B.**13**:1.

SAS STATISTICAL APPLICATIONS

Serge Dupuis

Software Support

Alberta Public Works, Supplies and Services

Introduction

SAS ™ (Statistical Analysis System) is a comprehensive package designed for data analysis. It includes a programming language for reading and manipulating data of almost any form and a number of procedures designed for analysing and displaying the data.

SAS offers an integrated approach to building systems by replacing several software packages for editing, database functions, programming, reporting and graphics into a single system. Although this paper deals primarily with statistics, SAS should be considered as an complete system. Add-ons to the base product include:

SAS/FSP Interactive facility for data entry,retrieval.

SAS/GRAPH Business graphics(Pie/bar charts), Scientific graphics (contouring, mapping, 3D modelling).

SAS/AF Application facility to create menus and turnkey applications.

SAS/OR Operations research, decision support,project management. Includes Gantt charts, Critical path Analysis

SAS/STAT Wide range of statistical procedure for analysis and modelling such as ANOVA, Regression, Chi-square, survival analysis and others.

SAS/ETS Econometrics,financial planning, time series modelling including ARIMA, forecasting time series, cross spectral analysis.

SAS/IML Interactive matrix programming language similar to APL which can be used for regression.

SAS/QC Quality control software including SHEWHART and several experimental design macros.

SAS is available on mainframes, minis and PCs on several operating systems including VAX, UNIX, MVS, and PC-DOS and Macintosh so that analyses and data can be readily shared among scientists, worldwide. The version for PC's is leased to Alberta Government Departments and Agencies through a master license at Public Works, Supplies and Services (PWSS). The Base product is leased from $90/year with add-ons from $40/year. The cost includes support and upgrades, but not manuals.

Manuals

All SAS/PC users should obtain the LANGUAGE guide (# P5856, $19.95US) and PROCEDURES guide (# P5856, $16.95US). For specific applications, the following listed in the reference section may also be required.

A master index of all documentation is highly recommended (publication# 56000). Many manuals can be ordered from PWSS (charged to mainframe account), others from SAS directly (DPO direct to SAS). A free semi-annual catalog of all publications is available from SAS. The address to order directly is:

SAS (Canada)
225 Duncan Mills Road, Suite 300
North York, Ontario
M3B 3K9
Phone (416) 443-9811 Fax: (416) 443-1269

Hardware Required

SAS/PC will require a PC/AT class computer or better such as IBM PS-2/Model 60-80 and COMPAQ 286/386. A 386 machine is recommended for future upgrades of SAS/PC under Windows. Expanded memory is not normally required but will be very useful. A math co-processor will be useful for complex statistical or graphical applications. With the addition of an IRMA or similar card for communication, SAS/PC programs can be executed on a mainframe where large data files can be processed. Be prepared to allocate 10 to 30 megabytes for storage on the PC. This could be stored on a Local Area Network (LAN) server as SAS/PC operates under many LANs such as NOVELL, IBM, Banyan-Vines and others.

Training

SAS has a steep learning curve, but fortunately, users do not need to understand the whole system to be able to effectively analyse their data. It is highly recommended that users take the tutorial provided with the installation diskettes to get familiar with windows, functions keys and basic syntax of the SAS language.

This basic training can be enhanced with other computer based tutorials available from Software Support, classroom courses and seminars which are held several times a year. The software also includes several help screens and fill the blanks menus which can be activated by function keys (F1=help) or by typing MENU on the command line. For those wanting an automated system, SAS/Assist and SAS/AF can be used to build turnkey systems requiring little knowledge of SAS by its users.

Each product also has its own set of sample programs, complete with data and ready to run. The index is listed in files called *.BLS. For example, the INDEXSTT.BLS file contains names of sample programs in the SAS/STAT guide. To run any of these programs, you need to invoke SAS and INCLUDE the program name, for example, ANOVAEX is included with the command:

```
    command===> INCLUDE 'ANOVAEX.SAS'
```

The program is ready to use by pressing the execute (f10) key.

A partial list of samples for SAS/STAT:

ANOVAEX	Documentation Examples from PROC ANOVA
TTESTEX	Documentation Examples from PROC TTEST
CATKAPPA	KAPPA STATISTIC COMPUTATION
FREQTREN	TREND TEST USING PROC FREQ
CLUSTER	CLUSTER ANALYSIS OF MAMMALS' TEETH DATA
FASTCLUS	USING MACROS TO ANALYZE ARTIFICIAL FIVE-GROUP
DEXCC	Macros for Central Composite Designs
DEXPB	Macros for Plackett-Burman Designs
PLANEX	Documentation Examples from PROC PLAN
CANDISC	CANONICAL DISCRIMINANT ANALYSES OF CARS DATA
CANDPOLY	POLYNOMIAL CANONICAL DISCRIMINANT ANALYSIS
DISCKERN	KERNEL DISCRIMINANT ANALYSES OF IRIS DATA
INTDISC	EXAMPLE FROM INTRO TO SAS DISCRIMINANT PROCS

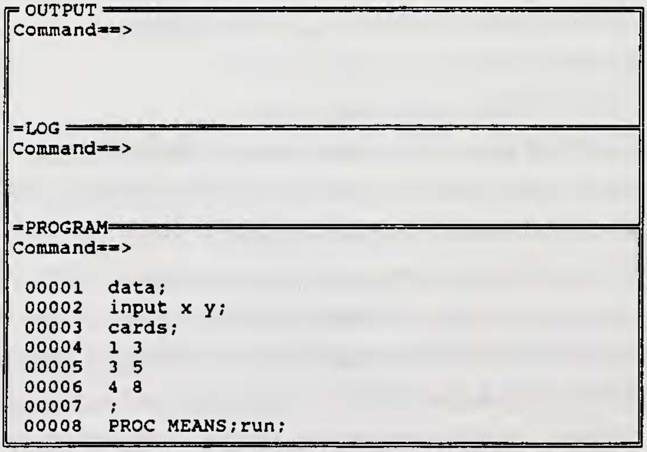
Getting Data into SAS

There are several method of entering data into SAS:

- Internal to the programs, using the INPUT and CARDS statements. This is acceptable when analysing small amounts of data.
- External to SAS, in DOS 'flat files'. These can be of any format and size, read with FILENAME and INPUT statement.
- In Dbase II, III ,III+ or Dbase IV files which can be converted very simply by using the DBF procedure.
- In LOTUS 1-2-3 files pre-converted to DIF or to printed to a PRN file and then read with an INPUT statement.
- Using SAS/FSP for entering the data. This will also do quality control on data entry and allow a screen to be any input form. This method is preferred for large numbers of variables or where data entry is critical. A specific tutorial is available for learning SAS/FSP.

A Sample Session with SAS/PC

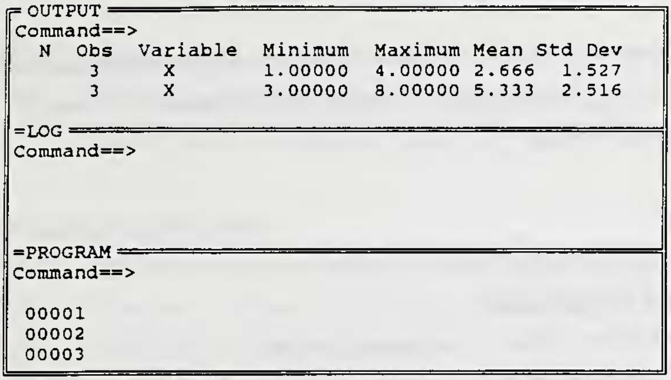
This is a sample program which reads data from a CARDS list and does a simple means calculation.



```
OUTPUT
Command==>

=LOG
Command==>

=PROGRAM
Command==>

data;
input x y;
cards;
1 3
3 5
4 8
;
PROC MEANS;run;
```

(Press F10 to execute the program)

The output is displayed on the output screen. To access it, press f5 and view the output:



```
= OUTPUT =
Command==>
  N  Obs  Variable   Minimum   Maximum Mean Std Dev
       3     X       1.00000   4.00000 2.666  1.527
       3     X       3.00000   8.00000 5.333  2.516

=LOG =
Command==>

=PROGRAM =
Command==>

 00001
 00002
 00003
```

Function Keys

Pressing F2 will display current keys assignment. The most useful function keys are:

F1 - help

F2 - keys

F5 - Jump to next window

F7 - Zoom window to full screen

F9 - Recall program

F10 - END / Execute program

Many other function keys are assigned and all are user definable.

Command Line

There is a command line on every window. The most common commands are:

Command==> X	Shell to DOS (exit to return)
Command==> Include '\dosfile'	Read a DOS file to window
Command==> File '\dosfile'	Write a DOS file to disk eg to printer: FILE '\lpt1'
Command==> BYE	Finish SAS Session

```
Command==> MENU                    Call procedure menu
```

Statistical Products

The base package includes very limited statistical procedures for descriptive statistics (mean,median, modes etc..). More complex statistical procedures can be found in SAS/STAT, SAS/ETS, SAS/IML and SAS/QC.

SAS/STAT Product

REGRESSION PROCEDURES:

CATMOD	Analyses data in continency tables
GLM	Least square fit for simple, multiple, polynomial and weighted regression
LIFEREG	Fits parametric models to failure data (survival analysis)
NLIN	Non linear regression, such as Gauss-Newton
ORTHOREG	For ill condition data using the Gentleman-Givens method Use colinearity diagnostics of REG to determine if ORTHOREG is needed.
REG	Linear regression with method selection from nine options such as backward, forward, stepwise, r-square
RSREG	Builds response surface models
TRANSREG	Obtains linear and nonlinear transformations of variables using alternating least squares to fit data to linear regression canonical regression and anova models

ANALYSIS OF VARIANCE:

ANOVA	Includes multivariate anova and repeated measures anova with several comparison tests. DO NOT use for unbalanced data, use GLM instead
CATMOD	Fits linear models for categorical data
GLM	Regression, analysis of covariance, repeated measures analysis, multivariate anova, hypothesis tests, test of means
NESTED	Anova and analysis of covariance on nested random models
NPAR1WAY	Non-parametric one-way of rank scores
PLAN	Constructs designs and randomizes plans for nested and crossed

	experiments
TTEST	Compare means of two groups
VARCOMP	Estimate of variance components for random or mixed models

CATEGORICAL ANALYSIS:

CATMOD	Fits linear models to functions of categorical data
FREQ	Builds tables or continency tables with chi-squares, Fishers' test

MULTIVARIATE ANALYSIS:

PRINCOMP	Principal component analysis, output component scores
FACTOR	Principal component and common factor analysis with rotations
CANCORR	Canonical correlation analysis

DISCRIMINANT ANALYSIS:

DISCRIM	Compute discriminant functions, including non-parametric methods
CANDISC	Canonical analysis to find linear combinations of quantitative variables
STEPDISC	Forward, backward or stepwise selection

CLUSTERING PROCEDURES:

CLUSTER	Hierarchical clustering using 11 methods applied to coordinate or distance data
FASTCLUS	Finds disjoint clusters using k-means (up to 100,000 observations)
VARCLUS	Hierarchical and disjoint clustering by oblique multiple group component analysis
TREE	Draws tree diagrams (dendrograms or phenograms)

The following may be used prior to clustering:

ACECLUS	Estimate of pooled within cluster covariance matrix
PRINCOMP	Principal component analysis
STANDARD	Standardizes variables to specified mean and variance

CORING PROCEDURES:

STANDARD	Standardizes variables to specified mean and variance
RANK	Ranks numeric variables from high to low
SCORE	Constructs new variables that are linear combination of old variables according to a scoring dataset (used with PROC FACTOR)

SURVIVAL ANALYSIS:

This is used for data that measure length of time to occurrence of an event, for example mean time before failure, or length of time a person stays on the job.

LIFEREG	Fits parametric accelerated failure time or regression models
LIFETEST	Computes nonparametric estimate of survival distribution

OTHER SAS/STAT PROCEDURES

CORRESP	Simple and multiple correspondence analyses. Reads a continency or Burt table or creates these tables from raw data. Also named Appropriate scaling, reciprocal averaging.
PRINQUAL	Obtains linear and nonlinear transformations of variables using alternate least squares.
PROBIT	Maximum likelihood estimates of regression parameters and threshold response rate for biological assay quantal response data
CALIS	Covariance analysis of linear structural equations
LOGISTIC	Fits linear logistic regression models for binary or ordinal data

SAS/ETS (Econometric and time series) PRODUCT

SAS/ETS has a number of statistical procedures for analying time series. This includes:

- Econometric models for market analysis or macro economics
- Corporate financial modelling including planning equations
- Physical models for mechanics, hydraulic and hydrologic models
- Biological models to simulate living systems
- Ecological models to represent systems in nature.

ARIMA	Autoregressive integrated moving average process(Box-Jenkins)
AUTOREG	Regression allowing serially correlated error

FORECAST Forecast series using trend-adjusted autoregressive or exponentially smoothing models

PDLREG Multiple regression with polynomial distributed lag

SPECTRA Computes periodograms, smoothed spectral density estimates, white noise tests

STATSPACE Autocorrelation of stationary vector time series by state space models

X11 Seasonally adjusts quarterly or monthly time series

A Sample Time Series Application

The following is a sample program to extrapolate a time series using the FORECAST procedure.

```
data a;
   input month year date :monyy.
        crude coal;
   format date monyy.;
   label crude='CRUDE PETROLEUM';
   cards;
 1 1965 JAN65 24.09  40.015
 2 1965 FEB65 21.86  37.862
 3 1965 MAR65 24.38  42.816
 4 1965 APR65 23.68  41.862
   ...more data lines...
11 1972 NOV72 28.28  56.297
12 1972 DEC72 28.94  44.904
;
proc forecast data=a out=b outest=c
   trend=2 outactual out1step outlimit interval=month lead=15;
   id date;
   var crude coal;
proc print data=c;
   title 'The Estimates from PROC FORECAST';
```

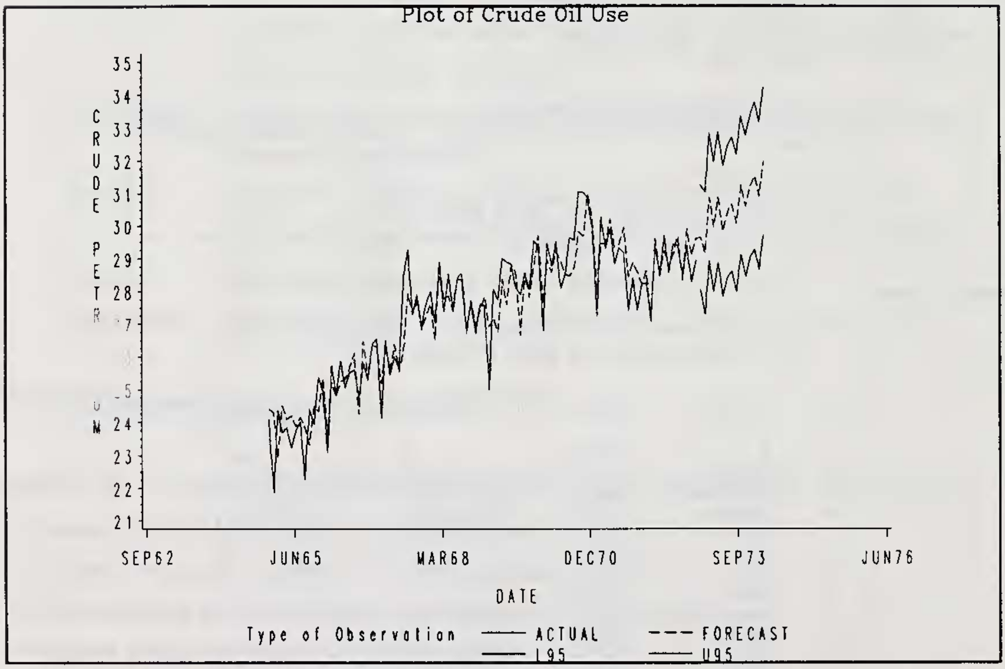
The output dataset:

```
     The Estimates from PROC FORECAST


OBS    _TYPE_       DATE        CRUDE          COAL

  1    N            DEC72            96            96
  2    SIGMA        DEC72     0.7967232     6.3485238
  3    CONSTANT     DEC72     24.349114     42.934934
  4    LINEAR       DEC72     0.0612253     0.0728315
  5    AR01         DEC72     0.6101027             .
  6    AR02         DEC72             .             .
  7    AR03         DEC72             .             .
  8    AR04         DEC72             .             .
  9    AR05         DEC72             .             .
 10    AR06         DEC72             .             .
 11    AR07         DEC72     0.2069177             .
 12    AR08         DEC72             .             .
 13    AR09         DEC72             .             .
 14    AR10         DEC72             .             .
 15    AR11         DEC72     -0.197647             .
 16    AR12         DEC72     0.5571647             .
 17    AR13         DEC72     -0.498052             .
```

The CRUDE variable has several autoregressive terms, there is an indication of seasonality as shown by the significance of the terms around AR12.

The output dataset can also be plotted with actual data, forecast and confidence limits:

```
proc gplot data=b;
   symbol1 i=join c=red    L=1 r=1;
   symbol2 i=join c=green  L=2 r=1;
   symbol3 i=join c=cyan   L=3 r=1;
   symbol4 i=join c=yellow L=4 r=1;
   symbol5 i=join c=Blue   L=5 r=1;
   plot crude*date=_type_;
   title 'Plot of Crude Oil Use';
run;
```



SAS/QC (Quality Control)

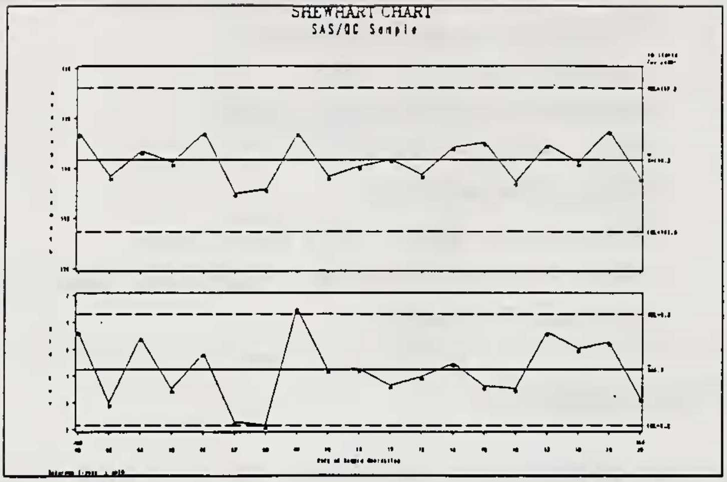
SAS/QC is an add-on package to SAS/PC designed for statistical quality control and experimental design. There are several applications for the product in manufacturing and laboratory. It features several graphics and analytical procedures such as:

CAPABILITY	Process capability analysis (including histograms with Gamma, Weibull and lognormal distribution)
CUSUM	Cumulative sum control charts
FACTEX	Orthogonal fractional factorial analysis
MACONTROL	Moving average control charts
OPTEX	Finding of optimal design
SHEWHART	Shewhart charts (MEAN X AND Range charts)
ADX	Macros for design and analysis of experiments (menu driven)
ISHIKAWA	Cause and effect diagrams

A sample use for PROC SHEWHART

The following was taken from the sample library. It tests and plots means and standard deviation.

```
data lengstat;
  input day mean std n;
  informat day date7. ;
  format day date5. ;
  label day ='Date of Sample Collection'
        mean='Average Length'
        std ='Standard Deviation of Length'
        n   ='Subgroup Sample Size';
  cards;
02JAN86  115.39  5.67  20
03JAN86  113.68  2.96  20
04JAN86  114.69  5.45  20
...more data lines.......
19JAN86  115.51  5.25  20
20JAN86  113.63  3.17  20
data lengstat;
  set lengstat;
  rename mean=lengthx  /* subgroup mean    */
         std =lengths  /* subgroup std. */
         n   =lengthn; /* subgroup sample size  */
proc shewhart history=lengstat graphics;
  xschart length*day='*';
run;
```



The charts show that this process is not in statistical control since the standard deviation (bottom) of the measurement exceeds the upper control limit.

SAS/IML (matrix programming)

The interactive matrix language allows more direct programming and array processing. It can be used for statistical applications such as:

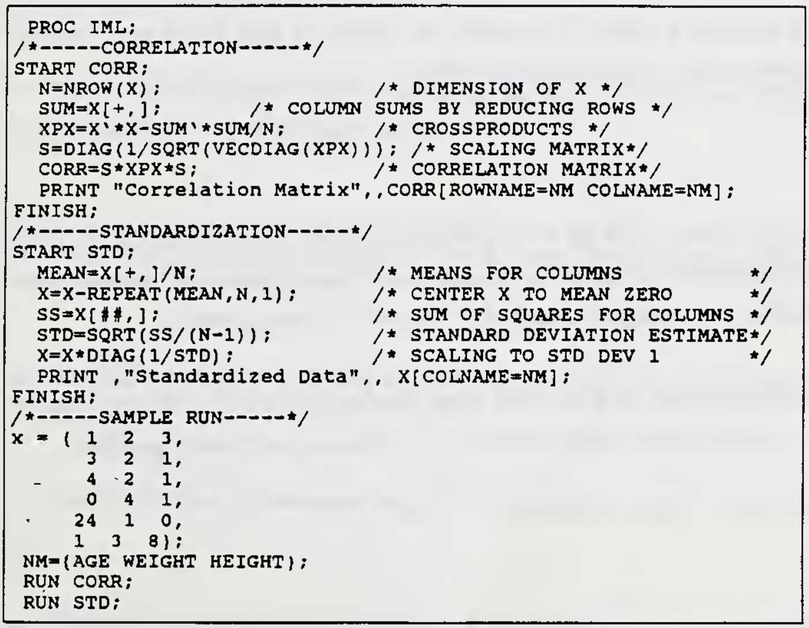
- Correlation
- Solving non-linear equations

- Regression
- Alpha factor analysis
- Categorical linear models
- Response surface analysis
- Logistic and Probit regression
- Linear programming
- And many other applications

SAS/IML can be used to replace the APL language without use of a special keyboard. It has the advantage of being interactive, unlike the DATA step. Data read in matrices can be converted to SAS datasets and SAS datasets can be converted to matrices.

Correlation example with IML

The following program, taken from the sample library supplied with SAS/IML shows the use of matrix language for a simple correlation.



```
 PROC IML;
/*-----CORRELATION-----*/
START CORR;
  N=NROW(X);                   /* DIMENSION OF X */
  SUM=X[+,];         /* COLUMN SUMS BY REDUCING ROWS */
  XPX=X`*X-SUM`*SUM/N;         /* CROSSPRODUCTS */
  S=DIAG(1/SQRT(VECDIAG(XPX))); /* SCALING MATRIX*/
  CORR=S*XPX*S;                /* CORRELATION MATRIX*/
  PRINT "Correlation Matrix",,CORR[ROWNAME=NM COLNAME=NM];
FINISH;
/*-----STANDARDIZATION-----*/
START STD;
  MEAN=X[+,]/N;                /* MEANS FOR COLUMNS          */
  X=X-REPEAT(MEAN,N,1);        /* CENTER X TO MEAN ZERO      */
  SS=X[##,];                   /* SUM OF SQUARES FOR COLUMNS */
  STD=SQRT(SS/(N-1));          /* STANDARD DEVIATION ESTIMATE*/
  X=X*DIAG(1/STD);             /* SCALING TO STD DEV 1       */
  PRINT ,"Standardized Data",, X[COLNAME=NM];
FINISH;
/*-----SAMPLE RUN-----*/
x = { 1  2  3,
      3  2  1,
      4  2  1,
      0  4  1,
     24  1  0,
      1  3  8};
 NM={AGE WEIGHT HEIGHT};
 RUN CORR;
 RUN STD;
```

Two matrices are produced: a Correlation matrix and a Standardized data matrix:

```
                    Correlation Matrix

         CORR            AGE     WEIGHT     HEIGHT
         AGE               1  -0.717102  -0.436558
         WEIGHT  -0.717102            1  0.3508232
         HEIGHT  -0.436558  0.3508232            1

                     Standardized Data

             X         AGE     WEIGHT     HEIGHT
               -0.490116  -0.322749  0.2264554
               -0.272287  -0.322749  -0.452911
               -0.163372  -0.322749  -0.452911
                -0.59903  1.6137431  -0.452911
               2.0149206  -1.290994  -0.792594
               -0.490116  0.6454972   1.924871
```

SAS/IML also contains a number of routines for displaying data which give a greater amount of control over graphs than with SAS/Graph alone.

Summary

SAS has a complete set of tools for the statistical analysis of any source of data. The challenge is knowing which procedure to use under specific conditions. Since it will almost always produce an output, a good understanding of statistics is required for interpretation.

The system has a steep learning curve for those wanting to use SAS in its raw form but it can be 'packaged' as an automated system with the use of menus and sample programs.

References

SAS, SAS/FSP, SAS/OR, SAS/GRAPH, SAS/STAT, SAS/ETS, SAS/IML and SAS/QC are registered of SAS Institute Inc., Cary, N.C., USA.

SAS Institute Inc. SAS Language Guide for Personal Computers, Release 6.03, Edition, Cary, NC: SAS Institute Inc., 1988, 558 pp.

SAS Institute Inc. SAS Procedures Guide, Release 6.03 Edition, Cary, NC: SAS Institute Inc., 1988, 441 pp.

SAS Institute Inc. SAS/GRAPH Users Guide, Release 6.03 Edition, Cary, NC: SAS Institute Inc., 1988, 549 pp.

SAS Institute Inc. SAS/FSP Users Guide, Release 6.03 Cary, NC: SAS Institute Inc., 1988, 331 pp.

SAS Institute Inc. SAS/STAT Guide for Personal Computers, Version 6 edition, Cary, NC: SAS Institute Inc., 1988, 1028 pp.

SAS Institute Inc. SAS Technical Report P-179, Additional SAS/STAT Procedures, Release 6.03, Cary, NC: SAS Institute Inc., 1988.

SAS Institute Inc. SAS Technical Report P-200, SAS/STAT Software: CALIS and LOGIST Procedures, Release 6.04, Cary, NC: SAS Institute Inc., 1990.
236 pp.

SAS Institute Inc. SAS/ETS Users Guide, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1988, 560 pp.

SAS Institute Inc. SAS/OR Users Guide, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1989, 479 pp.

SAS Institute Inc. SAS/QC Software: Reference Version 6, First Edition, Cary, NC: SAS Institute Inc., 1988, 660 pp.

SAS Institute Inc. SAS Technical Report P-188, SAS/QC Software Examples: Version 6, Cary, NC: SAS Institute Inc., 1988.

Appendix A

Program October 21,1990
and
Program October 22, 1990

Appendix A

Program
Thursday, October 21, 1990

TIME LOCATION	TOPIC
0900-0915 AUD	Opening Remarks
0915-1000 AUD	Applying Statistics to Practical Problems Milton Weiss
1000-1030	BREAK
1030-1115 AUD	Time Series Analysis Victor Adamowicz
1115-1200 AUD	Multivariate Methods Dave Jobson
1200-1300	LUNCH-ARC (Catered)

TIME LOCATION	GROUP	TOPIC
1300-1400 LMR	A	Statistical Graphics
CR	B	Discriminant Analysis
BDR	C	Sampling Insect Populations
BDR	D	Sampling Insect Populations
1400-1500 CR	A	Repeated Measures ANOVA
LMR	B	Statistical Graphics
BDR	C	Discriminant Analysis
BDR	D	Discriminant Analysis
1500-1515		BREAK
1515-1615 BDR	A	Sampling Insect Populations

Appendix A-continued

Program
Friday, October 22, 1990

TIME LOCATION	TOPIC
0900-0950 AUD	Experimental Design Zack Florence
0950-1015	BREAK
1015-1100 AUD	Parametric Assumptions Robert "Bob" Hardin
1100-1130 AUD	Statistical Problems in Compliance Assessment Albert Liem
1130-1200 AUD	Environmental Chemical Analysis Ian Johnson and Yogesh Kumar
1200-1300	LUNCH-ARC (Catered)

TIME LOCATION	GROUP	TOPIC
1300-1400 CR	A	Response Surface Analysis
CR	B	Response Surface Analysis
AUD	C	SAS Applications
AUD	D	SAS Applications
1400-1500 AUD	A	SAS Applications
AUD	B	SAS Applications
CR	C	Response Surface Analysis
CR	D	Response Surface Analysis
1500-1515 Auditorium		WRAP-UP

LEGEND

AUD=	Auditorium
LMR=	Library Meeting Room, Main Floor
BDR=	Board Room
CR=	Conference Room

Appendix A-continued

TIME LOCATION	GROUP	TOPIC
BDR	B	Sampling Insect Populations
LMR	C	Statistical Graphics
CR	D	Repeated Measures ANOVA
1615-1715 CR	A	Discriminant Analysis
BDR	B	Repeated Measures ANOVA
BDR	C	Repeated Measures ANOVA
LMR	D	Statistical Graphics

1900 (7:00 PM)
BUFFET (Optional)
Terrace Inn, 4440-Calgary Trail Northbound

LEGEND
AUD= Auditorium
LMR= Library Meeting Room, Main Floor
BDR= Board Room
CR= Conference Room

Appendix B

List of Participants

Appendix B

List of Participants

Adam, Michael	Statistics Branch, Alberta Agriculture
Adamowicz, Vic	University of Alberta,Dept. of Rual Economy
Arthur, Paul	University of Alberta, Dept. of Animal Science
Au, Peter	Alberta Forest Service
Bakowsky, Olenka	Alberta Forest Service
Bamsey, Colin	Alberta Forest Service
Basarab, John	Alberta Agriculture, Animal Industry Division
Bessette, Dale	Alberta Research Council
Brierly, Tony	Agriculture Canada
Christian, Ralph	Alberta Agricultural Research Institute
Darroch, Barbara	Alberta Environmental Centre
Day, Phyllis	Alberta Agriculture, Animal Industry Division
Dosdall, Lloyd	Alberta Environmental Centre
Dupuis, Serge	PWSS Software Support Branch
Florence, Marilyn	Consultant
Foster, Sandra	Hardy BBT
Gabruch, Barbara	Statistics Branch, Alberta Agriculture
George, Lee	Alberta Environmental Centre
Gietz, Michelle L.	Statistics Branch, Alberta Agriculture
Godby, Gavin	University of Alberta, Dept. of Animal Science
Heimann, Robert B.	Alberta Research Council
Henry, Philip J.	Alberta Environmental Centre
Herbut, Marion	Alberta Environmental Centre
Hiley, Jim	Agriculture Canada
Hwang, Sheau-Fang	Alberta Environmental Centre
Iacchelli, Angelo	Alberta Research Council, Devon Coal Research Centre
Jobson, J.D.	University of Alberta, Dept. of Accounting
Johnson, C. Ian	Alberta Environmental Centre
Jonasson, Ralph	Alberta Research Council
Khan, A. Aziz	Alberta Environmental Centre
Kryzanowski, Len	Alberta Agriculture, Soils Branch
Kumar, Yogesh	Alberta Environmental Centre
Lakusta, Tom	Alberta Forest Service, Timber Management Branch
Lamy, Denise	University of Alberta, Dept. of Animal Science
Ledene, Les	Canadian Charolais Association
Liem, Albert	Alberta Environmental Centre
Liu, M.F.	University of Alberta, Dept. of Animal Science
McClay, Alec	Alberta Environmental Centre
McKinnon, Blair	Workers Compensation Board, Research and Evaluation Department
Morgan, Dave	Alberta Forest Service, Timber Management Branch

Appendix B-continued

Mostrum, Michelle	Alberta Environmental Centre
Naazie, Augustine	University of Alberta, Dept. of Animal Science
Pawlowicz, John	Alberta Research Council
Phillips, Paul	Alberta Forest Service
Pittman, Vanessa	Lakeside Research
Rau, Ron	Alberta Research Council
Roy, Larry	Alberta Environmental Centre
Schipper, C	Alberta Agriculture,Health Management Branch
Sen, Amode Ranjan	University of Calgary, Dept. of Math and Statistics
Sherstabetoff, Rick	Alberta Agriculture, Soils Branch
Stilborn, Bob	Lakeside Research
Tickner, Glen	Workers Compensation Board, Research and Evaluation Department
Tong, Hang Mao	Hardy BBT
Wong, Ray	Alberta Research Council