A MONTE CARLO STUDY OF MODEL SELECTION PROCEDURES FOR THE ANALYSIS OF CATEGORICAL DATA

Elizabeth Martin Fischer, University of North Carolina

Introduction

Researchers and social scientists frequently confront data analysis situations for which existing theory provides little or no guidance concerning either the determinants of the dependent variable of interest or the nature of the relationships among variables. In such situations, researchers must rely upon a combination of intuition, previous empirical studies, theory, and exploratory data analysis in order to select an appropriate subset of explanatory variables and a model which adequately describes relationships among them. The present paper is concerned with techniques which may be used in the exploratory, model-building stage of research to analyze multi dimensional contingency tables.

We begin with a brief overview of model selection procedures for contingency table data. Four procedures are examined in detail; they are:

1. Stepwise backward elimination of parameters from a saturated model;
2. Stepwise backward elimination of parameters from a homogeneous baseline model;
3. Stepwise forward selection from a homogeneous baseline model;
4. Direct estimation, in which terms are eliminated from a saturated model based upon tests of significance of standardized parameter estimates.

The procedures are evaluated using Monte Carlo simulation techniques. We first specify a true model characterizing a hypothetical population and then analyze repeated samples generated from the hypothetical population. Because the true model is known, the results permit comparison of selection procedures. The following questions are considered:

1. How adequate are the various model selection techniques? That is, how likely is each to lead to the selection of the "true" model, when the true model is known?
2. What are appropriate criteria for an acceptable model? In particular, how well should a model fit in order to be considered acceptable?
3. How much confidence should be placed in the results of applying selection procedures when samples are small? Does the adequacy of the techniques depend upon the nature of the underlying population model?

Overview of Model Selection Procedures

In developing a model to describe a set of data, the researcher first identifies a set of variables for inclusion in the analysis and then specifies the model or equation relating the variables. We assume here that an appropriate set of variables has been identified, and consider techniques to identify a good fitting, parsimonious model to account for the data. A variety of statistical procedures to select log linear models have been developed to assist the researcher in the decision. (See e.g. Goodman 1971, 1973, Birch 1964, Bishop, Fienberg, and Holland 1975, and Brown 1976.) The search procedures vary in several key ways.

Stepwise versus simultaneous tests of parameters may be employed to eliminate or add terms to a model. Stepwise selection requires a test for each parameter to be included in or deleted from a model, while simultaneous procedures test multiple effect parameters simultaneously, and thus require fewer tests to select a final model. Goodman (1973) suggests that simultaneous tests may be employed as an initial screening procedure to eliminate some models from consideration before applying stepwise procedures. As with linear regression, forward or backward stepwise procedures may be employed. Forward selection involves stepwise addition of effect parameters to a model according to some criterion of statistical importance, while backward elimination "prunes" a saturated model by sequential deletion of parameters whose estimated values are statistically insignificant. Goodman (1971, p. 45) cites Draper and Smith (1966) to suggest that backward elimination is superior to forward selection, but provides no evidence concerning their relative performance.

Different methods rely upon different statistical criteria for adding or deleting effect parameters. Goodman (1971) advocates the use of the difference chi-square test statistic, which is the difference in chi-square values for two models, one including the parameter and one excluding it. A statistically significant difference between the two models implies that the effect is significant and must be included in the final model. Higgins and Koch (1977) rely upon chi-square divided by its degrees of freedom to assess the magnitude of an effect parameter. Goodman (1971) also advocates significance tests of the standardized parameter estimates as a criterion for inclusion in a model. Benedetti and Brown (1976) suggest that with large samples selection should not be based upon statistical significance, and advocate the selection of a model which explains a certain fraction of the lack of fit of a baseline model.

In all of these cases, the researcher must also determine the α-level to be used as the criterion of acceptance or rejection of parameters and models. Most applications rely upon conventional α levels of .05 or .10, but there is no evidence that these levels produce optimal results for any or all of the procedures. (In the context of linear regression, α-levels of .10 or .05 do not produce optimal results for all selection procedures.) Perhaps for this reason Goodman (1971) cautions against strict interpretations of significance levels associated with models, suggesting that they should be used as a simple way of taking account of degrees of freedom in assessing the relative goodness of fit of different models.

There have been few studies which assess the adequacy of different techniques and decision rules for selecting log linear models. One exception is Benedetti and Brown (1976), who use real world contingency table data to evaluate forward selection, backward elimination, direct estimation, and other procedures. Using as a criterion of success the selection of a model which cannot be significantly improved by adding parameters, and from which parameters cannot be dropped, they find that forward selection, backward selection, or a combination of the two performs most adequately. They recommend against the use of simultaneous tests to screen models from consideration, because they find that prior screening led to the exclusion of relevant parameters from the selected model. In addition, they recommend against the use of the difference chi-square when samples are large, arguing that such tests will always be statistically significant with large samples, and that a more appropriate criterion is selection of a model which explains a certain fraction of the lack of fit of a baseline model.

Although the Benedetti and Brown study is useful, the conclusions which can be drawn from two sets of analyses are limited. In addition, because the study is based upon analysis of real data, it is not known which (if any) of the selection procedures arrived at the true population model. Analyses based upon artificial data with known properties provide a more systematic basis for comparing different procedures. There have been a number of such simulation studies of procedures for selecting linear models. Although the findings may not be generalized directly to the log linear case, they are relevant to the issues raised here. The findings suggest that the performance of different linear regression model selection procedures depends in a complex fashion upon the data analysis situation. Dempster, Schatzoff, and Wermuth (1977) find that the performance of different selection procedures is affected by collinearity and multicollinearity among independent variables, centrality in the original model, and the pattern of true regression coefficients. In addition, the choice of significance level has an inconsistent effect upon the accuracy of stepwise selection procedures which rely upon significance tests as decision criteria. Based upon a simulation analysis, Kennedy and Bancroft (1971) recommend sequential deletion (using an α-level of .25) over forward selection. They further find that no single α-level is universally superior for all combinations of parameter values. Finally, Bendel and Afifi (1977) find that the relative performance of different stopping rules in forward stepwise regression depends upon sample size and the number of effect parameters. They recommend that the α-level used with backward elimination be half that used with forward selection.

The complexity of model selection in the linear regression context suggests the importance of evaluating procedures for selecting log linear models. The present paper offers preliminary findings of a Monte Carlo investigation of selecting models to describe multidimensional contingency tables. Two factors which affect the performance of algorithms for selecting linear regression models (the characteristics of the true population model and the size of the sample) are systematically varied. The results suggest that the selection procedures generally perform well, although the adequacy of different procedures depends upon the data analysis situation.

Method

The present analysis attempts to replicate typical analysis problems by simulating a variety of data analysis situations. The characteristics of the true model are varied, and the size of the sample is varied from 50 to 8000. It should be noted at the outset, however, that all of the simulations are based upon four-way cross-classifications of dichotomous variables, and that all models are relatively simple. We further assume that:

1. One true population model gives rise to the observed data.
2. All relevant variables and no irrelevant variables are included in the model.
3. Although the form of the equation is unknown, the correct model is hierarchical. That is, inclusion of a higher order term necessarily results in inclusion of lower order terms involving the same variables. (E.g. if the three-way interaction pertaining to variables A, B, and C is included, then all two-way and one-way effects involving A, B, and C are also included.)

Model selection procedures are compared according to how accurately they identify the correct form of the population model. Thus, a "true" model is one which includes all relevant effect parameters, and excludes all irrelevant effect parameters. The rationale for this broad definition is that in the exploratory stages of research the presence or absence of an effect is very often of primary interest; it is this aspect of specification error that is the focus of the present study.

The true models used to generate the simulated data are reported in Table 1. The only difference between the five hypothetical models is the magnitude of the three-way interaction pertaining to variables A, B, and C, which varies from .00 in Model 1 to .55 in Model 5. Based upon the true population parameters, the multinomial distributions underlying each of the five models are calculated and used to generate random samples via computer algorithm. Fifty replications are generated for each combination of model type and sample size. The selection algorithms are then applied to each random sample to identify models which describe the data.

The selection procedures compared here are discussed by Goodman (1971, 1973), Bishop, Fienberg, and Holland (1975), Benedetti and Brown (1976) and others.[1] In the first set of analyses, a significance level of α = .05 is used for all steps in the selection methods. The methods are:

1. Direct estimation. Goodman (1971, 1973) advocates the use of direct estimation as a guide to further stepwise selection of models; here it is applied as a procedure to select a final model. Under the null hypothesis of no effect, standardized parameter estimates are distributed normally and may be tested directly for statistical significance. Standardized parameter estimates in the saturated model are tested (using a critical value of 1.96) and non-significant parameters are deleted, unless deletion would result in a non-hierarchical model. Zero cells in the multiway table are replaced by 1/2 prior to estimating the

parameters of the saturated model. (This practice conforms to Goodman's recommendation in 1964 (see p. 633) but not his later recommendation to add 1/2 to all cells.)

Table 1. Effect parameters for simulated samples

λ effect pertaining to:	Non-standardized λ effects for Model: 1	2	3	4	5
A	.00	.00	.00	.00	.00
B	.00	.00	.00	.00	.00
C	.00	.00	.00	.00	.00
D	.00	.00	.00	.00	.00
AB	.25	.25	.25	.25	.25
AC	.25	.25	.25	.25	.25
AD	.25	.25	.25	.25	.25
BC	.25	.25	.25	.25	.25
BD	.25	.25	.25	.25	.25
CD	.25	.25	.25	.25	.25
ABC	.00	.10	.25	.40	.55
ABD	.00	.00	.00	.00	.00
BCD	.00	.00	.00	.00	.00
ABCD	.00	.00	.00	.00	.00

2. Backward elimination. Models are selected by deleting parameters in a stepwise fashion from a model. The decision to delete a parameter is based upon the statistical significance of the difference chi-square values[2] comparing two models which differ by the presence or absence of the parameter in question. The backward elimination algorithm first tests parameters of the highest order, and proceeds in systematic fashion to tests of lower order terms; at each stage, the parameter associated with the minimum p-value for the difference chi-square is deleted. That is, the term which contributes least to the overall goodness-of-fit of the model is eliminated. Deletion of parameters stops when further deletion would result in a statistically significant loss of fit, or when the goodness of fit of the overall model falls below the specified rejection level.

The model from which parameters are deleted may be either:

a) The saturated model including all main and interaction effects, or
b) A homogeneous baseline model which is selected as an initial best-fitting model.

Homogeneous models which include terms of uniform order (i.e. for an n-way table, models which include all terms of order 1, 2, . . ., k, k + 1, . . ., n) are sequentially compared using the difference chi-square to assess differences in goodness of fit. If the highest order (n-way) interaction is statistically significant, selection is terminated and the saturated model is selected (without further stepwise testing) as the final model. Otherwise, the baseline model is chosen as the model of order k, where k is the lowest order model (1) which fits acceptably ($p > .05$), (2) which fits significantly better than the model of order k - 1, and (3) which is not improved by the addition of all terms of order k + 1. Terms are then deleted in stepwise fashion from the k^{th} order model.[3]

3. Stepwise forward selection. Models are built by adding parameters in a sequential fashion to a baseline model, beginning with lower order terms and proceeding in systematic fashion to higher order terms. The algorithm is analogous to backward elimination, except that at each stage the parameter which most improves the goodness of fit is added. (That is, the term associated with the highest p-value for the difference chi-square is added.) The baseline model is selected using the procedure described for (2b), except that the baseline model is of order k - 1. Addition of parameters stops when no further addition results in a statistically significant improvement in goodness of fit.

Results

Results are found in Table 2; each entry represents the proportion of 50 replications for which the true model is selected using different selection strategies when sample size and the hypothetical population model are varied.

When averaged over samples of varying size and different population models, the data suggest relatively small overall differences in the success of the four model selection strategies. The proportion of correct selections in 2000 replications varies from .64 for backward elimination from the saturated model to .53 for direct estimation. However, the relative and absolute performance of different selection procedures varies according to the data analysis situation.

Not surprisingly, the probability of selecting the correct model is considerably greater when samples are large than when they are small, regardless of which selection procedure is used. However, sample size has a greater effect upon the accuracy of some procedures than others. Direct estimation performs very poorly, and worse than any of the stepwise procedures, when samples are small. For $n \leq 250$, the proportion of correct selections using direct estimation is .14, while the stepwise procedures select the true model for an average of .40 of the replications in samples of the same size. When samples are large ($n \geq 500$), average differences in accuracy among selection procedures are small; the proportion of correct selections is about .77 for all techniques. Sample size has a nonmonotonic effect upon the accuracy of selection procedures. All four techniques are most likely to select true models for sample sizes of 2000 or 4000, with accuracy declining somewhat in larger samples.

The nature of the true population model affects the likelihood that a correct selection will be made using any of the search procedures. When the true model contains a small three-way interaction effect (ABC = .10 in Model 2), no search procedure reliably selects the correct model unless the sample size is 2000 or larger. This finding suggests that if a small effect is theoretically or practically important, a relatively large sample is required to detect it reliably using the procedures examined here. In contrast, when the true model includes a large interaction term (ABC = .40 and .55 for Models 4 and 5, respectively) the stepwise procedures select the true model for over half of the replications even in samples of size 50.

The relative advantage of different model selection strategies also depends upon the true population model. Comparison of results for models 2-5 indicates that the larger the ABC interaction term, the more likely the three stepwise

Table 2. Selection of "true" model using four selection strategies and varying sample size and the characteristics of the true model.

	Sample Size								
	50	100	250	500	1000	2000	4000	8000	Total
Model 1 (ABC = .00)									
Backward deletion from saturated model	.00	.00	.36	.56	.72	.82	.78	.72	.50
Backward deletion from baseline model	.00	.02	.46	.82	.90	.90	.98	.82	.61
Forward selection from baseline model	.02	.04	.40	.56	.72	.84	.78	.72	.51
Direct estimation	.00	.04	.32	.64	.80	.78	.80	.68	.51
Model 2 (ABC = .10)									
Backward deletion from saturated model	.04	.08	.24	.38	.40	.82	.86	.78	.45
Backward deletion from baseline model	.04	.02	.04	.18	.32	.72	.86	.78	.37
Forward selection from baseline model	.04	.02	.16	.38	.38	.72	.86	.78	.42
Direct estimation	.00	.00	.22	.35	.40	.74	.86	.80	.42
Model 3 (ABC = .25)									
Backward deletion from saturated model	.16	.36	.82	.76	.84	.96	.86	.86	.70
Backward deletion from baseline model	.10	.20	.72	.76	.84	.96	.86	.86	.66
Forward selection from baseline model	.10	.20	.80	.78	.84	.96	.88	.86	.68
Direct estimation	.00	.03	.55	.80	.84	.94	.82	.84	.60
Model 4 (ABC = .40)									
Backward deletion from saturated model	.62	.72	.76	.84	.76	.80	.84	.74	.76
Backward deletion from baseline model	.52	.50	.76	.84	.76	.80	.84	.74	.72
Forward selection from baseline model	.56	.50	.78	.84	.76	.80	.84	.76	.73
Direct estimation	.00	.07	.40	.84	.82	.82	.88	.78	.58
Model 5 (ABC = .55)									
Backward deletion from saturated model	.60	.84	.86	.80	.82	.86	.84	.72	.79
Backward deletion from baseline model	.52	.84	.86	.80	.82	.86	.84	.72	.78
Forward selection from baseline model	.54	.86	.86	.80	.82	.86	.84	.74	.79
Direct estimation	.00	.04	.40	.63	.88	.86	.90	.76	.56
Total									
Backward deletion from saturated model	.28	.40	.61	.67	.71	.85	.84	.76	.64
Backward deletion from baseline model	.24	.32	.57	.68	.73	.85	.88	.78	.63
Forward selection from baseline model	.25	.32	.60	.67	.70	.84	.84	.77	.63
Direct estimation	.00	.04	.38	.65	.75	.83	.85	.77	.53

procedures are to select the true model. In contrast, the accuracy of the direct estimation procedure is relatively unaffected by the nature of the population model. Consequently, the relative superiority of the stepwise selection procedures over direct estimation increases as the size of the ABC interaction term is increased.

Differences among the three stepwise selection procedures are relatively small. Backward elimination from a homogeneous baseline model performs better than other procedures when the underlying population model is homogeneous, as is Model 1, which includes all two-way interaction terms. Backward elimination from a baseline model appears to be less sensitive to the presence of small higher order interaction terms than the other stepwise procedures; this is a disadvantage when the true model includes small higher order terms (as is the case for Model 2) but is an advantage when such terms merely represent "noise" (as is the case for Model 1).

All four procedures for selecting a log linear model to describe a multiway contingency table rely upon criteria of statistical significance to accept or reject individual parameters and models. There has been little investigation of an appropriate α-level to select log linear models, although as noted the optimal significance level for selecting linear models generally differs according to procedure, and is usually higher than a conventional α of .05. Some attention was therefore given to the questions of an appropriate level of significance to be used as the criterion of rejection, and at what stage in the selection process the stopping rule should be invoked.

Table 3 presents the results of varying the α level used to select models which describe Model 3 samples, employing backward elimination from a homogeneous baseline model. The results are somewhat surprising. It was thought that use of a less stringent α level (e.g. .10) might improve the accuracy of model selection procedures when samples are small, but this proved not to be the case. Instead, a more stringent α of .01 yielded improved accuracy when averaged over all sample sizes; the small decrease in accuracy for samples of size 100 or less is more than balanced by improvements in accuracy for larger samples. An α-level of .01 implies that a parameter is included in a model only if it is statistically significant at the .01 level. More controversially, an α level of .01 implies that a model is rejected only if it is associated with a probability of .01 or less. It is counterintuitive that a strategy which accepts models which fit so poorly by conventional standards is nevertheless most likely to lead to selection of the true model, especially when samples are large. Similar results are found for forward selection, and hold as well when tested using Model 1 samples. Although further investigation of appropriate α levels is warranted, these results tentatively suggest that an α-level of .01 may produce better results than α-levels of .05 or .10 when samples are 250 cases or larger.

Table 3. Proportion of correct selections for different levels of α.

Sample size	α = .01	.05	.10	.25	.50
50	.06	.10	.10	.00	.00
100	.10	.20	.22	.08	.02
250	.76	.72	.72	.24	.02
500	.90	.76	.66	.38	.02
1000	.92	.84	.66	.26	.04
2000	.98	.96	.41	.42	.14
4000	.98	.86	.68	.28	.06
8000	.94	.86	.72	.38	.06
Total	.71	.66	.52	.26	.04

A second issue is the question of when in the search process stopping rules should be invoked. The stepwise selection procedures used here terminate search when overall goodness of fit of the selected model falls below the specified rejection level. The criterion that the selected model must be associated with a probability of .05 or greater is applied not only to the final selection, but to all intermediate models in the stepwise selection process. This is particularly problematic for backward elimination if, for example, lower order models fit well although higher order models fit poorly. This may occur when deletion of higher order terms adds degrees of freedom but does not much reduce goodness of fit. If the stopping rule is invoked at an intermediate stage, search will terminate before good-fitting, lower order models are tested. It is possible that the search process would be improved if the criterion for overall goodness of fit of a model is applied only to the final model which results from the search.

A possible example of premature termination of the search process occurs when the highest order (four-way) interaction term is statistically significant. In this case, all four selection procedures terminate search and select the saturated model as the final model. For the 2000 samples analyzed in Table 2, the four-way interaction is significant at the .05 level in 118 samples (or for .06 of the replications, which is slightly greater than the chance expectation under the null hypothesis). Of course, in all cases the four-way term represents random variation, since it is included in none of the true models. In addition, in many cases lower order models fit the data well. If the search for a good-fitting model is continued ignoring the significant four-way interaction term, the true model is selected (and fits acceptably) in 48 of the 118 samples. Thus, for the models considered here, the likelihood of selecting a true model is marginally improved if stopping rules are not employed to terminate search.

Discussion

When averaged over sample sizes and model types, the results indicate relatively small overall differences in the performance of difference selection strategies for the data analysis situations simulated here. The principal finding is that direct estimation performs worse than any of the stepwise procedures, due mainly to its relatively poor performance when samples are small and the true model includes a large interaction term. Thus, the results of the simulation suggest that if the researcher has identified the appropriate set of variables to analyze, if the population model is hierarchical and relatively simple, and if one of the stepwise procedures is used, the true model may be selected with probability between about .25 and .90, depending upon the size of the sample. The results suggest that these model selection procedures (particularly direct estimation) should generally not be applied to very small samples ($n < 250$). However, even direct estimation performs quite well for large samples, suggesting that Goodman (1971) may be too cautious in his recommendation against the use of standardized λ's as a simple guide to the selection of models.

When samples are large, the simple strategies analyzed here perform relatively well and about equally. This suggests that there may be little need for the complex, multidirectional selection strategies such as those proposed by Goodman (1971, 1973) and Benedetti and Brown (1976). Of course, it must be emphasized that four-way tables characterized by relatively simple hierarchical models have been simulated; the results reported here may not generalize to larger tables or more complex situations. Nevertheless, an emphasis upon the development of many alternative methods for selecting models may be misplaced. Instead, it may be more appropriate to focus attention upon other important issues concerning the selection of descriptive models for categorical data. One such issue is the neglected problem of how variables should be selected for inclusion in an analysis. Koch and his students have recently developed criteria for the selection of variables (see e.g. Higgins and Koch, 1977) although the adequacy of such procedures has not been evaluated.

Finally, the results reported here are germaine to two points made by Benedetti and Brown (1976). The recommend against the selection of a homogeneous baseline model prior to application of stepwise procedures, because it may lead to the exclusion of relevant parameters. Comparison of the results obtained by backward elimination from a saturated versus homogeneous baseline model indicates that when models are not screened, backward deletion is somewhat more sensitive to the presence of interaction effects (i.e. in Models 2-5) but is also more likely to detect interaction where there is none (i.e. in Model 1). Thus, neither method is superior in all data analysis situations. Benedetti and Brown (1976) also argue that for large samples the difference chi-square should not be used as the criterion for inclusion of terms, and that a more appropriate test would be based upon explained variance. Although the two decision rules are not compared here, the results do not indicate that the difference chi-square is an inappropriate statistical criterion for model selection. The performance of the search procedure may be improved, especially for large samples, by using an α level of .01 as the criterion for acceptance or rejection of models and parameters.

Footnotes

[1]Stepwise model selection is carried out by a computer program (MAT) developed at the University of Chicago and modified at the University of Michigan, the University of North Carolina and Duke University.

[2]The difference chi-square for a model M_1 and a model M_2 containing additional effect parameter(s) is calculated as $\chi_1^2 - \chi_2^2$, with degrees of freedom $df_1 - df_2$. A difference chi-square value associated with $p \leq .05$ indicates that M_2 fits significantly better than M_1, and the term(s) in M_2 should therefore be retained. If $p > .05$ the term(s) in M_2 do not make a significant contribution to goodness of fit and may be deleted.

[3]This procedure differs somewhat from that described by Bishop, Fienberg, and Holland (1975, p. 157-8). Backward elimination is not confined to intervening models which include all terms of order k - 1 and some or all terms of order k, but may delete terms of order k - 2, etc. However, if terms of order k are statistically significant and must be included in a model, then terms of order k - 2 will not be tested or deleted using the stepwise algorithm employed here.

Acknowledgments

This research was supported by NSF Grant #SOC76-02484. I would like to thank Miltiades Damanakis for assistance in computing.

References

Bendell, R.B. and A.A. Afifi. 1977. "Comparison of stopping rules in forward 'stepwise' regression." Journal of the American Statistical Association 72 (357):46-53.

Benedetti, J.K. and M.B. Brown. 1976. "Selecting log-linear models: A comparison of strategies." Paper presented at the Annual Meeting of the American Statistical Association, Boston, Massachusetts, August 1976.

Birch, M.W. 1964. The detection of partial association, II: the general case. Journal of the Royal Statistical Society Series B27: 11-24.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. Discrete Multivariate Analysis. Cambridge, Mass: The MIT Press.

Brown, M.B. 1976. Screening effects in multidimensional contingency tables. Applied Statistics 25:37-46.

Dempster, A.P., M. Schatzoff, and N. Wermuth. 1977. "A simulation study of alternatives to ordinary least squares." Journal of the American Statistical Association. 72 (357): 77-91.

Draper, N. and H. Smith. 1966. Applied Regression Analysis. New York: John Wiley.

Goodman, L.A. 1964. "Interactions in multidimensional contingency tables." Annals of Mathematical Statistics 35:632-646.

________. 1971. "The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications." Technometrics 13 (1):33-61.

________. 1973. "Guided and unguided methods for the selection of models for a set of T multidimensional contingency tables." Journal of the American Statistical Association 68 (341):165-175.

Higgins, J.E. and G.G. Koch. 1977. "Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey." International Statistical Review 45:51-62.

Kennedy, W.F. and T.A. Bancroft. 1971. "Model building for prediction based upon repeated significance tests." The Annals of Mathematical Statistics 42 (4):1273-1284.