A Multiple Logit Analysis of a Family Planning System

Robert W. Makuch and Daniel H. Freeman, Jr.
Yale University
Nancy S. Henley and Sagar C. Jain
U. of North Carolina

1. Introduction

This study comprises the second phase of an investigation by Henley et al. (1976) on data from a family planning program in the state of Haryana, India. Briefly, the study was designed to evaluate the extent to which various inputs, within the boundary of fixed constraints, affect program output. The output was the delivery of contraceptive devices; the input was both financial and technical; the constraints were socio-economic and demographic variables. Over 150 data variables were obtained for each of the 97 geographical blocks composing Haryana; however, many of these variables were eliminated on the basis of known unreliability, level of detail they represented, or large numbers of missing values. Thirty-seven independent variables and four measures of program output (IUD acceptors, tubectomies, vasectomies, and number of condoms distributed) remained. The first phase focused on summary statistics such as four-year averages across the period studied, 1968-1971. Five rural blocks and two urban blocks were dropped from the analysis due to 1 or more missing values; thus, 90 blocks remained for analysis. A maximum R-squared improvement technique was applied to ascertain the relative importance of the variables on the program output, and the objective was simply to obtain equations which had as many statistically significant variables as possible.

The second phase involved a multiple logit analysis which focused on the IUD acceptance rates to describe yearly patterns within the data. In particular, a linear model was constructed to examine both the relationship among the input variables in the context of fixed constraint variables, and the interaction between the constraint and year (trend) effect. The linear model was of the form

$$K \ln \pi \doteq X\beta \qquad (1)$$

where K, X are known matrices subject to the condition that the matrix of first partials $K \ln \pi$ with respect to the elements of π has full row rank and X has full column rank, π is a vector of multinomial probability parameters, β is a vector of parameters to be estimated, $\ln \pi$ denotes the logarithm of the elements of the vector π, and '$\doteq$' means 'is approximated by'. The regression coefficients β were estimated using a regression approach developed by Grizzle, et al. (1969). This methodology entails a weighted least squares approach to determine a BAN estimator $\hat{\beta}$ of β. Moreover, minimum modified chi-square statistics described by Neyman (1949) were employed to assess the adequacy of the model in fitting the data and to test hypotheses of interest.

The choice of independent variables (program inputs) was based on a preliminary optimal regression analysis (see Hocking, 1976), and the algorithm used for selecting input variables was described by Furnival and Wilson (1974). The procedure allowed the selection of the 'best' and several 'nearly best' subsets according to a specified criterion. Therfore, a comparison of the possible regression equations in terms of a summary statistic measuring the aspects of the adequacy of each equation was needed to aid in deciding which subsets to choose. The estimate C_p of the standardized total squared error Γ_p was chosen as the selection criterion. This estimate is defined as

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n-2p)$$

where p = number of variables in the reduced regression equation, n = number of data points, RSS_p is the residual sums of squares for the particular p-variate regression equation under consideration, and $\hat{\sigma}^2$ is the residual mean square from the full term p'-variate equation ($p' \geq p$). Employment of C_p statistics rather than the R-squared criterion was favored since C_p plots are more amenable to graphical analysis. Furthermore, this statistic served to confirm the observations of Henley et al. who utilized the P^2 statistic.

In summary, our analysis was two-fold. On the one hand, the optimal regression procedure was implemented so that the most important sets of input variables could be determined. The second part of the analysis embodied these results in constructing a linear model to represent a parsimonious, yet adequate, description of the data.

2. Overall (Optimal) Regression

The data consisted of responses gathered at four 1-year intervals for 85 urban and rural blocks. Based on the earlier investigation, 27 independent variables (defined in Table 1) were used for the overall regression. Observations on one block were assumed to be independent of those on every other block. Furthermore, a no-interaction linear model was postulated and a

logit transformation of the IUD acceptance rate Y (see Table 1) served as the dependent variable. Residual plots were examined for each of the 27 independent variables and, as a result, log transformations of variables X4, X5, and X26 were performed. Finally, an inspection of the normal probability plots indicated that variables from this set of data could be regarded as samples from a normal distribution.

TABLE 1
Variables Considered for Use in Overall Regression

Variable	Variable Name
X1	Male illiteracy-percent-1971
X2	Female illiteracy-percent-1971
X3	Newspaper subscribers/1000 population
X4	Radios/1000 population (pop.)
X5	Agricultural composite
X6	% of total workers in agriculture-1971
X7	% vaids who motivate for family planning
X8	Dais/100,000 pop. (4 year average)
X9	Auxiliary nurse midwives (ANMs) - 1971
X10	Male physicians/100,000 pop.
X11	IUD referrals from 'other' sources/ total referrals (4 year average)
X12	Health expenditures/1000 pop. (4 year avg.)
X13	Extension educators/100,000 pop.
X14	Family planning field workers/100,000 pop. (4 year average)
X15	Auxiliary nurse midwives/100,000 pop. (4 year average)
X16	Hospital and PHC beds/100,000 pop. (4 year average)
X17	Family planning media events/100,000 pop. (4 year average)
X18	Family planning educational groups/100,000 pop. (4 year average)
X19	Km paved road - 1971
X20	Extension educators trained/in place-1971
X21	Female physicians/100,000 pop.
X22	Family planning expenditures/1000 pop. (4 year average)
X23	Dais trained/in place - 1971
X24	Sterilization referrals from sources 'other' than health and family planning sources/total sterilization referrals (4 year average)
X25	Field workers - 1971
X26	Population density - 1971
X27	% scheduled caste - 1971
Y	# IUD acceptors for 4 years/# of eligible couples in year 4

After this preliminary inspection of the data, simple correlation coefficients for all pairs of variables were computed to implement the all possible regressions procedure. It is noteworthy that over 96% of the coefficients were between -.3 and +.3, thereby raising little concern that problems due to multicollinearity would arise in this analysis. From the optimal regression procedure, the 'best' (in terms of lowest C_p value for all p) and several 'nearly best' subsets of independent variables for each size p, p=1,...,27, were determined. The elements of the most desirable subsets of each size p, p=1,...,11, are given in Figure 1. We note that most of the same variables reappear as the subset size increases. This seems to indicate that these variables were good regressors for the dependent variable, and this interpretation was useful when we constructed our model to analyze time trends.

FIGURE 1
Best and Nearly Best Subsets of Size p, p=1,...,11, and their elements

Set	Elements in Set (denoted by subscripts of variables in Table 1)
A	22
B	12 22
C	3 22
D	3 22 24
E	3 14 22
F	3 14 15 22
G	3 10 14 22
H	3 10 12 14 22
I	2 3 10 14 22
J	2 3 10 15 22
K	2 3 10 14 15 22
L	3 10 11 12 14 15 24
M	3 8 10 11 12 14 15
N	3 10 11 12 14 15 21 24
O	3 8 10 11 12 14 15 24
P	3 8 10 11 12 14 15 21 24
Q	2 3 10 11 12 14 15 21 24
R	2 3 8 10 11 12 14 15 21 24
S	2 3 10 11 12 14 15 21 24 27
T	2 3 4* 10 11 12 14 15 21 24
U	2 3 4* 6 10 11 12 14 15 21 24
V	2 3 4* 10 11 12 14 15 21 24 27
W	1 2 3 4* 10 11 12 14 15 21 24

* denotes a transformed variable

3. Modeling For Time Trends

Ten rural blocks were randomly chosen to illustrate a modeling procedure for time trends. Let n_{ij} be the number of IUD acceptors for year i (i=1,2,3,4) in block j (j=1,...,10), and N_j be the number of eligible couples in year 4 for block j. Then define $p_{ij} = n_{ij}/N_j$ to be the proportion of IUD acceptors for year i in block j, where

$$\underset{1x50}{p'} = (p_1', \ldots, p_{10}')$$

and

$$p_j' = (p_{1j} \cdots p_{4j}\ 1 - \sum_{i=1}^{4} p_{ij}) .$$

Choose

$$\underset{40x50}{K} = \underset{}{K^*} \otimes \underset{10x10}{I} , \quad \text{where}$$

$$K^* = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

and ⊗ denotes the Kronecker product, to form the vector

$$\underset{40\times 1}{\underset{\sim}{\ell}} = \underset{\sim}{K}\ln\underset{\sim}{p} = \begin{bmatrix} \ell_{1,1} \\ \ell_{2,1} \\ \ell_{3,1} \\ \ell_{4,1} \\ \ell_{1,2} \\ \vdots \\ \ell_{4,10} \end{bmatrix} .$$

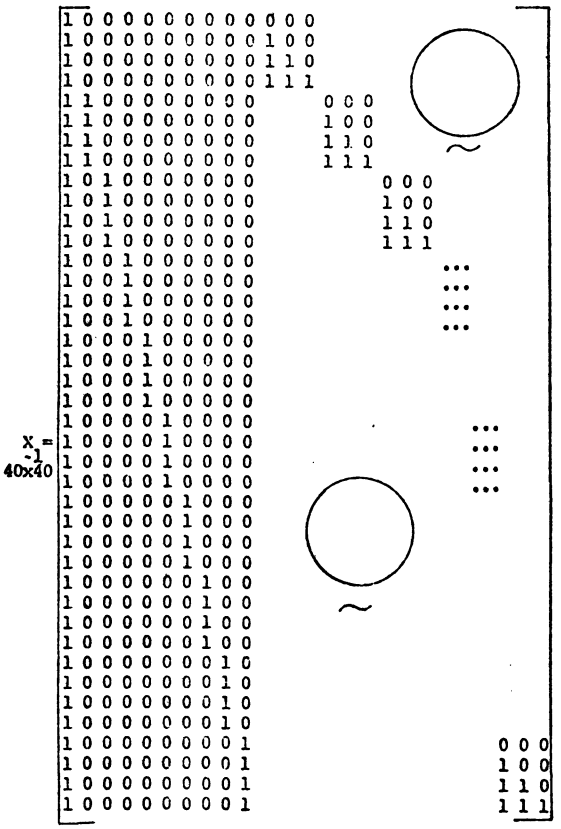
Each $\ell_{i,j} = \ln \dfrac{p_{ij}}{1-\sum_{i=1}^{4} p_{ij}}$ corresponds to an element of the multiple logit response vector $\underset{\sim}{\ell}$ for year i in block j, and variation among these elements can be investigated by fitting linear regression models defined in (1). Briefly stated, the modeling procedure can be characterized by writing

$$\underset{\sim}{E}_{A} \{K\ln p\} = K\ln\pi \doteq X\beta$$

where '$\underset{\sim}{E}_{A}$' means 'asymptotic expectation', so that BAN estimates $\hat{\beta}$ of β and minimum modified χ^2 statistics for testing hypotheses of interest by the method of weighted least squares can be obtained.

Patterns in the data suggested that 1) the proportion of IUD acceptors decreased as we moved from year 1 to year 4 and 2) these yearly trends were different in some blocks. Taking these patterns into account, the vector $\underset{\sim}{\ell}$ was fit to the right hand side of (1) by constructing the design matrix $\underset{\sim}{X}_1$. The first ten columns of $\underset{\sim}{X}_1$ spanned the space of all constraint variables, and the next thirty columns of ten identical matrices pertained to the constraint-by-year, or trend, space. Hence, our model initially lacked any input variables since the rank of $\underset{\sim}{X}_1$ was equal to the number of responses (i.e. the model was saturated).

$\underset{40\times 40}{\underset{\sim}{X}_1} =$



Since one aim of our analysis was to explain away as much variation in the data as possible using the program inputs, some of the ten rural blocks were grouped together to allow for the addition of these input variables to the model. The chi-square test statistics in Table 2 served as a guide in this grouping procedure.

TABLE 2
Test Statistics for Preliminary $\underset{\sim}{X}_1$ Model

Source of Variation		Degrees of Freedom	χ^2	p
Block	1 vs 2	1	36.74	0
	2 vs 3	1	25.88	0
	3 vs 4	1	9.75	0
	4 vs 5	1	.13	.72
	5 vs 6	1	11.68	0
	6 vs 7	1	40.00	0
	7 vs 8	1	.97	.32
	8 vs 9	1	14.07	0
	9 vs 10	1	20.27	0

One possible grouping is given in Table 3. Another possible grouping arrangement was identical to this grouping except that the last two blocks were not combined into one group. However, in order to utilize as many program inputs as possible, the combination of ten blocks into five groups as indicated in Table 3 was chosen for further analysis.

TABLE 3

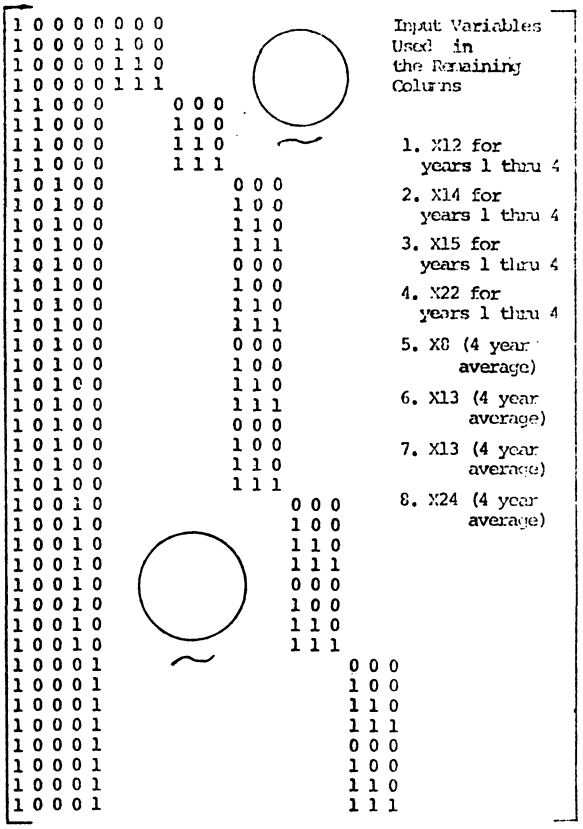
Number of Rural Blocks in Each Group	Range of Overall Non-acceptance Rate Within Each Group
1	85%
1	90%
4	90.4%-91.4%
2	93.7%-94.6%
2	94.7%-95.6%

Based on the overall regression analysis, 9 input variables were considered. They were:

1. Family planning expenditures
2. Family planning field workers
3. Auxiliary nurse midwives
4. Female medical doctors
5. Sterility referrals
6. Health expenditures
7. Dais
8. Family planning media events
9. Extension educators

Since data on these variables were collected for each of the four years studied, 36 possible regressors were available. However, the overall regression suggested that family planning expenditures (X22), family planning field workers (X14), auxiliary nurse midwives (X15), and health expenditures (X12) were good candidates for year-by-variable interaction. Adopting this suggestion, we introduced parameters into the model for each of the four years for each of these four variables.

X_2 = 40x40



To adequately describe the grouping of the blocks in Table 3, 5 columns to distinguish the 5 groups and five identical three-column matrices to span the constraint-by-year interaction space were also included. Finally, four of the next most important variables were represented by one column each in X_2 to again form a saturated model. This resulted in the design matrix X_2 as shown.

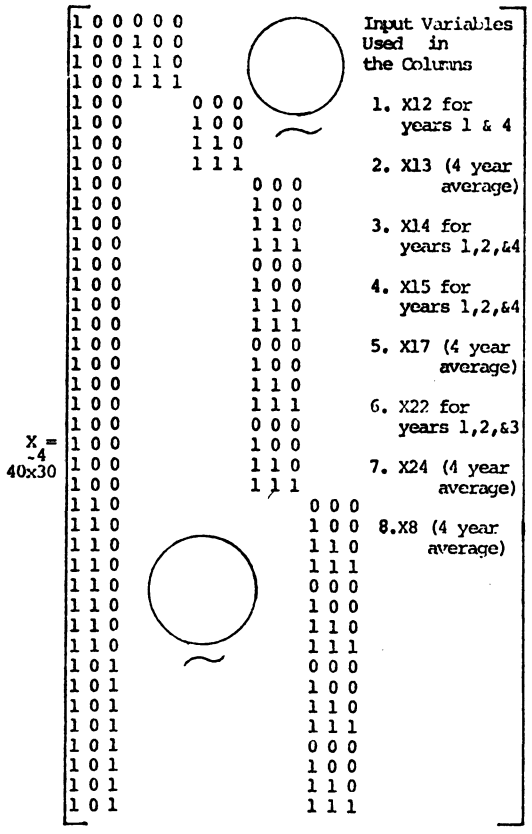
Based on significance tests, only non-significant constraint-by-year (trend) effects and group means were collapsed since it was of interest to determine the importance of the program inputs on the family planning system. As a result, the trend parameters of the last two groups and the constraint parameters of the first three groups were collapsed, as both χ^2 tests were not significant ($p>.25$). An intermediate model was formed after taking these results into account and including the ninth variable, extension educators (X13), as another parameter. Statistical tests for this intermediate model appear in Table 4.

TABLE 4

Test Statistics for Intermediate Model Analysis of Variation

Source	Degrees of Freedom	χ^2	p
Model	35	6020.58	0
Constraints	2	34.60	0
Trend	12	308.57	0
Across groups for year:			
2	4	110.93	0
3	4	106.68	0
4	4	36.33	0
Significance of trends within each group			
Group			
1	3	18.17	0
2	3	59.63	0
3	3	34.74	0
4	3	30.28	0
Input variables	21	373.27	0
Family planning expenditures	4	56.83	0
Health expenditures	4	51.18	0
Family planning field workers	4	457.50	0
ANMs	4	47.15	0
Family planning media events	1	17.00	0
Dais	1	25.97	0
Sterility referrals	1	10.63	0
Extension educators	1	13.09	0
Female doctors	1	.01	.91
Error	4	1.88	.76
Total	39	6022.46	

Moreover, one degree of freedom tests on parameters for the number of family planning workers in year three, the amount of health expenditures in years two and three, and the amount of family planning expenditures in year four were non-significant ($p>.18$) . Another model was fit to the data excluding these parameters as well as the non-significant parameter for female doctors. Statistical significance of the parameters for this new model were examined and only one parameter (the number of ANMs in year three) was not statistically important. Our next model was then constructed by removing the parameter for ANMs in year three and constructing the design matrix X_4.



This model fit quite well (χ^2 = 11.55 on 10 degrees of freedom) and so valid test statistics, shown in Table 5, were also obtained. All the parameters were highly significant ($p<.05$) and so no further removal of any of the parameters could take place without the fit of the model becoming unsatisfactory. As a result, this model was deemed the final model.

TABLE 5

Test Statistics for Final Model X_4

Analysis of Variation

Source	Degrees of Freedom	χ^2	p
Model	29	6010.91	0
Constraints	2	35.76	0
Trend	12	279.47	0
Across groups for year:			
2	4	134.73	0
3	4	140.44	0
4	4	45.69	0
Significance of trends within each group			
Group 1	3	18.21	0
2	3	81.69	0
3	3	40.26	0
4	3	44.06	0
Input variables	15	366.25	0
Family planning expenditures	3	98.65	0
Health expenditures	2	59.91	0
Family planning field workers	3	164.65	0
ANMs	3	54.81	0
Family planning media events	1	8.83	0
Dais	1	31.39	0
Sterility referrals	1	51.18	0
Extension educators	1	15.40	0
Error	10	11.55	.32
Total	39	6022.46	

4. Conclusion

One thing that is immediately apparent from this table is that the input variables have a relatively large effect on the family planning system in comparison to the trend or constraint variables considered individually. Secondly, within the set of input variables, the magnitude of the χ^2 statistics reflects the importance of the number of family planning field workers, the amount of family planning expenditures and overall health expenditures, and the number of auxiliary nurse midwives. Of course, we must keep in mind that these results are preliminary as only ten rural blocks were examined. However, it is worthwhile to point out that these conclusions were similar to those in the investigation of Henley et al.. Furthermore, the methodology described here can be extended to analyze all the rural and urban blocks to see if similar results would be obtained.

REFERENCES

Furnival, G.M. and Wilson, R.W., Jr. (1974). Regression by leaps and bounds. Technometrics 13, 499-512.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Henley, N.S., Jain, S.C., and Wells, H.B. (1976). Relative importance of program input and environmental constraints to family planning programs in Haryana, India. Presented at the 1976 Annual Meeting of the Population Association of America (in Montreal).

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

Neyman, J. (1949). Contributions to the theory of the χ^2 test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles: University of California Press. 239-72.

Mallows, C.L. (1967). Choosing a subset regression. Bell Telephone Laboratories, unpublished report.

Mallows, C.L. (1973a). Some comments on C_p. Technometrics 15, 661-75.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426-82.