The Philippine Statistician, 1998
Vol. 47, Nos.1-4, pp. 77-88

Statistical Models For The Changes Of Preferential Variables: I. Dichotomous Classification

Emeterio S. Solivas[1]

ABSTRACT

This paper illustrates the reparametrization of the standard log-linear models to formulate special models that can be used to analyze changes in preferential variables observed in repeated measurements. It describes the essential features and potentials of the techniques, and illustrates the techniques with the analysis of a dataset on voting preferences among presidential aspirants obtained from a random sample of University of the Philippines Los Banos employees using a widely available statistics software to show their practical applicability.

Keywords: model matrix, logit model, reparametrization, persistence coefficient, symmetry coefficient.

1. Introduction

Usually, when a researcher has data on categorical variables to analyze, he first summarizes the data in two-way contingency tables and then uses chi-square-based techniques to determine the existence of association among the variables. If ever he further describes the results, he uses the classical log-linear model techniques. However, after obtaining the results from running the data set using a statistical software, the researcher just simply states whether or not there is association existing among the variables, or he may indicate on which categories of the variables they are associated. This kind of interpretation of results may be faulty if he does not consider the default constraint used in the software for parameter estimation. Different constraints necessitate different interpretation of results. Thus, these methods are fraught with limitations especially when the objective is to study changes like preferential shifts, and when there is temporal ordering of the variables as in repeated measurements.

This study considers some specifications of the general log-linear models to analyze changes in attitudinal variables observed in repeated measurement data. The objective is to find how existing techniques for the analysis of log-linear models can be specialized to study processes of change; to apply special models to study shifts in the distributions and associations among attitudinal variables; to describe the essential features and potentialities of the techniques in a less technical fashion for practicality; and to illustrate how the analyses are implemented with widely available statistics software to show their practical applicability.

2. Theoretical Framework

This study deals with the analysis of two-occasion repeated measurements data on

[1] Associate Professor of Statistics, Institute of Statistics, U.P. Los Banos, College, Laguna

dichotomous variables. The extension of this to t-occasion ($t \geq 3$) repeated measurements data on polychotomous variables will be done in a subsequent paper.

2.1. The sample 2 x 2 contingency table

Consider a categorical variable Y with $c = 2$ categories or classification levels (e.g. voting preference on a presidential candidate with yes and no vote categories) observed at two occasions (say, in January and then in March) from the same random sample of n individuals. Let Y_t be the observation on Y at time t, ($t = 1, 2$). When the data on Y_1 and Y_2 for the n individuals are cross-tabulated we have a square 2 x 2 contingency table. Let n_{ij} be the observed frequency in the cell (i,j) where $i = 1, 2$ indexes categories of Y_1 and $j = 1, 2$ indexes categories of Y_2. The contingency table of observed frequencies is given in the table below.

Table 2.1. A *2* x *2* contingency table of observed frequencies.

Y_1	Y_2		*Total*
	1	*2*	
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

We assume that we fixed n when we obtain our random sample so that the cell frequencies follow the *product binomial distribution.* We analyze this sample data to make inference about the population from where this sample is drawn.

2.2. The log-linear model for a 2 x 2 table.

Let F_{ij} be the expected (theoretical) frequency in the cell (i,j) under an assumed model. The table of expected frequencies is given below.

Table 2.2. A *2* x *2* contingency table of expected frequencies.

Y_1	Y_2		*Total*
	1	*2*	
1	F_{11}	F_{12}	$F_{1.}$
2	F_{21}	F_{22}	$F_{2.}$
Total	$F_{.1}$	$F_{.2}$	n

The standard (saturated) log-linear model is defined as

$$\ln(F_{ij}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(i,j)} \quad i = 1, 2 \text{ and } j = 1, 2 \qquad (2.1)$$

where: μ is the general mean of all $\ln(F_{ij})$'s
$\mu_{1(i)}$ is the effect of being in the ith category of Y_1
$\mu_{2(j)}$ is the effect of being in the jth category of Y_2
$\mu_{12(i,j)}$ is the interaction of the ith category of Y_1 with the jth category of Y_2.

The saturated log-linear model (2.1) has a design or model matrix constructed as follows:

Cell(i,j)	Parameters /Design matrix								
	μ	$\mu_{1(1)}$	$\mu_{1(2)}$	$\mu_{2(1)}$	$\mu_{2(2)}$	$\mu_{12(11)}$	$\mu_{12(12)}$	$\mu_{12(21)}$	$\mu_{12(22)}$
(1,1)	1	1	0	1	0	1	0	0	0
(1,2)	1	1	0	0	1	0	1	0	0
(2,1)	1	0	1	1	0	0	0	1	0
(2,2)	1	0	1	0	1	0	0	0	1

It is noted that this design matrix is 4 x 9 of rank 4; that is, there are 9 unknown parameters in only four cell frequency equations. This implies that the parameters are not estimable. In order to have estimates of the parameters there is a need to impose a restriction. Several restrictions, also called constraints or assumptions, may be used. One of the usual restrictions is the set-to-zero constraints.

For the last-category-set-to-zero constraint, the cell model and the model matrix become:

Cell(i,j)	Cell model $\ln(F_{ij}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}$	Model matrix			
		μ	$\mu_{1(1)}$	$\mu_{2(1)}$	$\mu_{12(11)}$
(1,1)	$\ln(F_{11}) = \mu + \mu_{1(1)} + \mu_{2(1)} + \mu_{12(11)}$	1	1	1	1
(1,2)	$\ln(F_{12}) = \mu + \mu_{1(1)}$	1	1	0	0
(2,1)	$\ln(F_{21}) = \mu + \mu_{2(1)}$	1	0	1	0
(2,2)	$\ln(F_{22}) = \mu$	1	0	0	0

This new model matrix is already of full rank, so the parameters can be uniquely estimated. It is noted that

$$\mu_{1(2)} = \mu_{2(2)} = 0; \quad \mu_{12(12)} = \mu_{12(21)} = \mu_{12(22)} = 0 \tag{2.2}$$

so there is only a need to estimate the non-redundant parameters, μ, $\mu_{1(1)}$, $\mu_{2(1)}$ and $\mu_{12(11)}$.

Note: The estimates of these parameters are not equal for the different restrictions and so they have different interpretations. Under the last-category-set-to zero constraint,

1. Column 2 of the design matrix indicates how, at Y_1, the log frequency of category 1 differs from that of the reference category 2. Thus, at Y_1, the $\mu_{1(1)}$ is the increase (or decrease) of the log frequency from category 2 to category 1,

2. Column 3 of the design matrix indicates how, at Y_2, the log frequency of category 1 differs from that of the reference category 2. Thus, at Y_2, the $\mu_{2(1)}$ is the increase (or

decrease) of the log frequency from category 2 to category 1,

3. Column 4 of the design matrix indicates how the log frequency of category 1 differs from that of reference category 2 at Y_2 given the category of Y_1. Thus, the $\mu_{12(11)}$ measures the association between Y_1 and Y_2.

2.3. The odds-ratio as a measure of association

Consider again Table 2.2. Defining the following odds,

F_{11}/F_{12} - the odds of an individual in category 1 of Y_1 to be in category 1 at Y_2,

F_{21}/F_{22} - the odds of an individual in category 2 of Y_1 to be in category 2 at Y_2,

then the ratio of these odds, denoted as OR, called the *odds-ratio* of category 1, is

$$\text{OR} = F_{11}F_{22}/F_{12}F_{21} \tag{2.3}$$

which is the likelihood that an individual be in category 1 of Y_2 when he is in category 1 of Y_1 compared to that when he is in category 2 of Y_1. It is noted that OR = 1 if Y_1 and Y_2 are independent, since the two odds are equal. In which case the ln (OR) = 0.

2.4. The logit model

Let F_{i1}/F_{i2} be the odds of an individual in category i of Y_1 to be in category i at Y_2. The logit of category i is defined as $\ln(F_{i1}/F_{i2})$ or $\ln(F_{i1}) - \ln/F_{i2})$. In the saturated model (2.1), this logit can be expressed as

$$\ln(F_{i1}/F_{i2}) = (\mu_{2(1)} - \mu_{2(2)}) + (\mu_{12(i1)} - \mu_{12(i2)}) \tag{2.4}$$

Equation (2.4) is called the *logit model* or *log-odds model* of category i, which is a linear combination of the standard log-linear parameters. Under the last-category-set-to zero constraint, this logit model is

$$\ln(F_{i1}/F_{i2}) = \mu_{2(1)} + \mu_{12(i1)} \tag{2.5}$$

Other logit models may be similarly defined.

2.5. A reparametrization of the log-linear model to measure changes

The association between Y_1 and Y_2 is expected in repeated measures data, so the interaction parameter $\mu_{12(ij)}$ does not indicate directly the changes in Y. Likewise, the main effect parameters $\mu_{1(i)}$ and $\mu_{2(j)}$ do not describe directly the changing marginals especially when there is higher-order interaction. Thus, there is a need to reparametrize the model (2.1) so that the new parameters will be used to measure the changes that take place in Y.

Looking at the cells in Table 2.2, it is seen that cells (1,1) and (2,2) denote *persistence* – (1,1) represents persistence of category 1 and (2,2) represents persistence of category 2. It is also seen that cells (1,2) and (2,1) denote *changes* – (1,2) represents the change from category 1 to category 2, and (2,1) represents the change from category 2 to category 1. The magnitude of F_{12} and F_{21} tells us the changes that occurred. The situation $F_{12} = F_{21}$ denotes ***symmetry*** or no net change. It is also desired to see if there are changes in the marginal distribution of Y_1 and the marginal distribution of Y_2 . The particular situation of $F_{1.} = F_{.1}$ and $F_{2.} = F_{.2}$ denotes *marginal homogeneity.*

The new specialized model of change is now formulated as

$$\ln(F_{ij}) = \mu + \mu_P + \mu_S + \mu_A \qquad (2.6)$$

where μ_P , μ_S and μ_A are parameters or coefficients which denote ***persistence, symmetry and association***, respectively. The special design matrix of model (2.6) is as follows:

Table 2.3. Design matrix of the specialized model (2.6).

Cell(i,j)	Parameters/Design matrix			
	μ	μ_P	μ_S	μ_A
(1,1)	1	1	0	1
(1,2)	1	0	1	-1
(2,1)	1	0	-1	-1
(2,2)	1	-1	0	1

From the model for each cell log frequency and using simple algebraic operations, the following are obtained:

$$\mu_P = [\ln(F_{11}/F_{22})]/2 \; ; \quad \mu_S = [\ln(F_{12}/F_{21})]/2 \; ; \quad \mu_A = [\ln(F_{11}F_{22}/F_{11}/F_{22})]/2 \qquad (2.7)$$

Summarizing these logits in terms of the usual log-linear parameters gives the following:

Parameter	Under last-category-set-to zero constraint
μ_P	$(\mu_{1(1)} + \mu_{2(1)} + \mu_{12(11)})/2$
μ_S	$(\mu_{1(1)} - \mu_{2(1)})/2$
μ_A	$\mu_{12(11)}$

For more straightforward interpretations, these parameters of changes, μ_P, μ_S and μ_A, may be modified. These modified parameters, μ^*_P, μ^*_S and μ^*_A, are given in the table below:

Table 2.3. Modified coefficients of changes, μ^*_P, μ^*_S and μ^*_A

Parameter	Using logits	Using log-linear parameters
μ^*_P	$\ln(F_{11}/F_{22})$	$\mu_{1(1)} + \mu_{2(1)} + \mu_{12(11)}$
μ^*_S	$\ln(F_{12}/F_{21})$	$\mu_{1(1)} - \mu_{2(1)})$
μ^*_A	$\ln(F_{11}F_{22}/F_{11}/F_{22})$	$2\mu_{12(11)}$

Note: That $\mu^*_S = 0$ if there is symmetry, $\mu^*_P = 0$ if the two level Y have equal persistence, and $\mu^*_A = 0$ if Y_1 and Y_2 are independent.

Thus, the parameters of changes are logit formulations or linear combinations of the usual log-linear parameters. One only have to fit the usual log-linear model.

2.6. Estimation of the parameters

Once a log-linear model is assumed, the parameters of the model are estimated and the expected frequencies under the assumed model are estimated. Many widely available references provide the formulas and algorithms for estimating the standard log-linear parameters and the expected frequencies. Among these are Haberman (1979), Goodman, 1978), Knoke and Burke (1980). The two most commonly used algorithms are the Newton-Raphson method and the iterative proportional fitting method (Neri, 1992).

To find the adequacy of the assumed model, a test of goodness-of-fit is performed, by the Pearson chi-square test statistic (χ^2) or the likelihood-ratio test statistic (L^2) given as

$$\chi^2 = \Sigma (n_{ij} - F_{ij})^2/F_{ij} \qquad (2.8)$$

$$L^2 = 2\Sigma\, n_{ij} \ln(n_{ij}/Fij) \qquad (2.9)$$

Both follow the chi-square distribution with appropriate degrees of freedom.

2.7. Inference about the parameters of change

Given the vector of standard log-linear model parameters, $\underline{\beta}$

$$\underline{\beta} = (\mu \quad \mu_{1(1)} \quad \mu_{2(1)} \quad \mu_{12(11)}),$$

let $\underline{\mathbf{B}}$ be the usual maximum likelihood estimate of $\underline{\beta}$,

$$\underline{\mathbf{B}} = (m \quad m_{1(1)} \quad m_{2(1)} \quad m_{12(11)})$$

$\underline{\mathbf{X}}$ be the design matrix of the model, and **F** be the vector of the estimates of F_{ij}'s.

$$\underline{\mathbf{F}} = (f_{11} \quad f_{12} \quad f_{121} \quad f_{22})$$

The variance-covariance matrix of $\underline{\mathbf{B}}$, denoted as $V(\underline{\mathbf{B}})$, is shown in many references (for example, by Agresti, 1990). Under the multinomial sampling, the asymptotic $V(\underline{\mathbf{B}})$ is

$$V(\underline{\mathbf{B}}) = [\underline{\mathbf{X}}'\{\underline{\mathbf{D}} - 1/n(\mathbf{FF}')\}\underline{\mathbf{X}}]^{-1} \text{ where } \underline{\mathbf{D}} = \text{diag}(F_{ij}) \qquad (2.10)$$

The standard error of the ith element of $\underline{\mathbf{B}}$, $se(m_i)$, can be obtained from the corresponding diagonal element of $V(\underline{\mathbf{B}})$. For sufficiently large sample size n, this estimate b_i is approximately normal, thus we can perform an approximate z-test for the significance of this estimate.

To test the significance of the parameters of change, from Table 2.3, let $\mathbf{W}$ be defined as

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Then $\underline{\mathbf{W}}\,\underline{\mathbf{B}} = (m^*_P \;\; m^*_S \;\; m^*_A)'$ is the vector of the estimates of the μ^*_P, μ^*_S and μ^*_A, and

$$V(\underline{\mathbf{W}}\,\underline{\mathbf{B}}) = \underline{\mathbf{W}}V(\underline{\mathbf{B}})\underline{\mathbf{W}}' \qquad (2.11)$$

The element of $\underline{\mathbf{W}}\,\underline{\mathbf{B}}$ corresponding to the parameter of change estimate can be tested for its significance by the approximate z-test.

For the 2 x 2 contingency table, the results of (2.11) can be easily expressed as:

$$V(m^*_P) = V(m_{1(1)}) + V(m_{2(1)}) + V(m_{12(11)}) + 2Cov(m_{1(1)}, m_{2(1)})$$

$$+ 2Cov(m_{1(1)}, m_{12(11)}) + 2Cov(m_{2(1)}, m_{12(11)}) \qquad (2.12)$$

$$V(m^*_S) = V(m_{1(1)}) + V(m_{2(1)}) + 2Cov(m_{1(1)}, m_{2(1)}) \qquad (2.13)$$

$$V(m^*_A) = 4V(m_{12(11)}) \qquad (2.14)$$

With these variances, we compute the approximate z-test statistics, z_c. Say for m^*_S,

$$z_c = m^*_S / se(m^*_S) \qquad (2.15)$$

and at 5% level of significance, the hypothesis that the coefficient is zero is rejected if $|z_c| \geq 1.96$; at 1% level, the hypothesis is rejected if $|z_c| \geq 2.575$.

3. Data and Methods

The data used to illustrate the techniques described earlier are part of the data set gathered and used by Pacia and Prosuelo (1998). The data set was obtained from a random sample of 200 employees of the University of the Philippines at Los Baños. A questionnaire was administered to the sample employees on the second week of January 1998 and then administered again to these sample employees on the first week of March.

The data are about the preferences of the sample employees on the leading presidential aspirants during the second week of January and then the first week of March, 1998. The preferences were De Villa, Lim, Roco and Others (consisting of DeVenecia, Estrada and Santiago). The observed contingency table of preferences is summarized in the next table.

Table 3.1. Preferences of presidentiables. January - March, 1998.

January	March				Total
	1. De Villa	2. Lim	3. Roco	4. Others	
1. De Villa	19	7	8	3	37
2. Lim	6	51	12	2	71
3. Roco	2	6	47	1	56
4. Others	1	2	10	23	36
Total	28	66	77	29	200

To illustrate the analysis of 2 x 2 contingency table, the categories (presidentiables) were dichotomized. The resulting four sub-tables are given in Tables 3.2 – 3.5.

Table 3.2. Preference for De Villa. Jan-March, 1998

January	March		Total
	Yes	No	
Yes	19	18	37
No	9	154	163
Total	28	172	200

Table 3.3. Preference rate for Lim. Jan-March, 1998

January	March		Total
	Yes	No	
Yes	51	20	71
No	15	114	129
Total	66	134	200

Table 3.4. Preference rate for Roco. Jan-March, 1998

January	March		Total
	Yes	No	
Yes	47	9	56
No	30	114	144
Total	77	123	200

Table 3.5. Preference for "Others". Jan-March, 1998

January	March		Total
	Yes	No	
Yes	23	13	36
No	6	158	164
Total	29	171	200

The log-linear model fittings are performed using the SPSS to obtain estimates of the standard models and other statistics. Then the coefficients of change are computed from the usual output.

4. Results and Discussion

The test of independence on each dichotomized sub-table is summarized below.

Table 4.1. Summary of the test of independence between January and March preferences.

	Likelihood-ratio test		Cramer's V	
2x2 sub-table	L^2 statistic	p-value	Coefficient	p-value
1. De Villa	41.09	0.000	0.51	0.000
2. Lim	76.51	0.000	0.61	0.000
3. Roco	69.83	0.000	0.58	0.000
4. Others	67.01	0.000	0.66	0.000

In all sub-tables, the March preferences were substantially related to January preferences. The dependence model was fitted on each sub-table to obtain parameter estimates, using the last-category-set-to-zero constraints. The results are summarized in the next table.

Table 4.2. Parameters of change estimates for January to March preferences.

	Estimates			
2x2 sub-table	μ	$\mu_{1(1)}$	$\mu_{2(1)}$	$\mu_{12(1,1)}$
1. De Villa	5.037	-2.147	-2.840	2.894
2. Lim	4.736	-1.740	-2.028	2.964
3. Roco	4.736	-2.539	-1.335	2.988
4. Others	5.063	-2.498	-3.271	3.841

All estimates are significantly different from zero at 1% level of significance

The expected logits and odds-ratios were estimated and summarized in the next table.

Table 4.3. Estimates of some logits and the odds-ratios.

2x2 sub-table	$\ln(F_{11}/F_{12})$	$\ln(F_{21}/F_{22})$	OR
1. De Villa	0.054	-2.838	18.03
2. Lim	0.936	-2.028	19.38
3. Roco	1.652	-1.336	19.85
4. Others	0.570	-3.270	46.52

The logit $\ln(F_{11}/F_{12})$ indicates the log-odds of an individual who preferred a candidate to remain with that candidate instead of changing to another candidate. A positive logit indicates that there are more of those remaining with the candidate than those changing to

another, with larger positive values indicating that more remained. All sub-tables have positive values. For example, in the Roco sub-table, the logit 1.652 indicates that there are $e^{1.652} = 5.217$ times more remaining than changing. For De Villa, 0.054 indicates an almost equal number of those who changed and of those who remained.

The logit $\ln(F_{21}/F_{22})$ indicates the log-odds of an individual who did not prefer a candidate but changed preference to that candidate instead of remaining with the other candidates. A negative logit indicates that there are fewer of those changing to the candidate than those remaining with the other candidates; larger absolute values indicating that more remained. All sub-tables have negative values. In Roco, for example, the logit –1.336 indicates that those who changed are only $e^{-1.336} = 0.26$ times as many as those who remained.

The odds-ratios indicate the likelihood of an individual to stay with a candidate compared to that from other candidates to change favor for this candidate. This also measures the strength of relationship between March and January preferences. A log-odds-ratio equal to zero indicates no relationship.

For more direct measures of the changes that took place, the modified coefficients of persistence, symmetry and association given in Table 2.3 were estimated from the usual log-linear parameter estimates in Table 4.2 and are summarized in the next table.

Table 4.4. Modified estimates of the persistence (μ^*_P), asymmetry (μ^*_S) and association (μ^*_A) parameters of the January- March preferences.

2x2 sub-table	μ^*_P	μ^*_S	μ^*_A
1. De Villa	-2.093 ** (.243)	-0.693 ns (.439)	5.788 ** (0.950)
2. Lim	-0.804 ** (.168)	-0.288 ns (.340)	4.592 ** (0.762)
3. Roco	-0.886 ** (.173)	1.204 ** (.424)	5.976 ** (0.355)
4. Others	-1.928 ** (.223)	-0.773 ns (.518)	7.682 ** (1.083)

Figures in parenthseses are the standard errors

** - significantly different from zero at 1% level; ns not significant at 5% level.

Notes:

1. The $\mu_P = 0$ indicates equal persistence between levels of preference.
   The $\mu_P > 0$ indicates more individuals remaining in level 1 than in level 2.
   The $\mu_P < 0$ indicates more individuals remaining in level 2 than in level 1.

2. The $\mu_S = 0$ indicates symmetry; that is, changes from one level to another cancel out each other.
   The $\mu_S > 0$ indicates more individuals changing preference from level 2 of Y_1 to level 1 of Y_2.
   The $\mu_S < 0$ indicates more individuals changing preference from level 1 of Y_1 to level 2 of Y_2.

3. The $\mu_A = 0$ indicates that Y_1 and Y_2 are not associated.
The $\mu_A > 0$ indicates that Y_1 and Y_2 are associated.

The results showed that voting preferences are characterized by persistence. The symmetry coefficients are mainly not significantly different from zero, except for the Roco sub-table. However, the association coefficients are significant indicating the existence of the relationship between March and January preferences.

5. Summary and Conclusion

This study showed how reparametrization of the standard log-linear models can be used in the case of repeated measurements data. The basic concepts on linear model estimation procedures and tests of hypothesis are applied on the log-linear models.

The reparametrization technique is quite easy to apply and the results are readily interpretable. There is only a need for a software that has the facility of the standard log-linear modeling.

This technique of measuring changes in the attitudinal variables is applicable in other areas of research. For market research, the technique may be used for measuring brand loyalty or TV program loyalty. For development research, it may be used to measure growth and progress. For social science research, it may be used to measure shifts of preference, attitudes and many more.

6. Acknowledgment

I wish to thank Ms. Analiza S. Pacia and Mr. Bryan B. Prosuelo, my former Stat 190 (Special Problem) advisees, for lending me their survey data and for the motivation that I got for this study when I was advising them.

7. References

AGRESTI, A. 1990. *Categorical Data Analysis*. Wiley and Sons, Inc.

GOODMAN, L.A. 1978. *Analyzing qualitative/categorical data*. Cambrodge, Mass.: Abt Books

HABERMAN, S. J. 1979. *Analysis of Qualitative Data. Vol. 2*. New Developments. New York: Academic Press

KNOKE, D., and BURKE, P. J. 1980, *Log-linear Models,* Beverly Hills, Calif.: Sage.

NERI, Liza. 1992, *Log-linear Analysis of Categorical Data*. Special Problem . UPLB

NORUSIS, M.J. 1993. *SPSS Advanced Statistics Guide*. SPSS Inc. Chicago. Illinois.

PACIA, A. S. and B.B. PORSUELO. 1998. *Log-linear Modeling: An Approach to the Analysis of Changes in nominal variables*. Special Problem. UPLB.