INTERPOLATING, EXTRAPOLATING, AND FORECASTING QUALITATIVE ATTRIBUTES BY LOG-LINEAR MODELS

Clifford C. Clogg, The Pennsylvania State University

ABSTRACT

Log-linear models provide methods for the time-trending of qualitative attributes on the basis of information contained in crosstabulations obtained at successive points in time. Simple time-trend models are presented whereby the logits ϕ_{it} of some polytomous variable with $i = 1,\ldots,I$ classes are linked to time scores $t = 1,\ldots,T$ by means of a polymonial equation $\phi_{it} = A_i + B_{i1}t + \ldots + B_{i(T-1)}\, t^{T-1}$ of order T-1 or less. Once a suitable model is found, the predicted logit for any time t* is obtained by substituting t* for t in the final model. Formulae for the standard error of interpolated or extrapolated logits (or proportions) are developed such that the variance of a prediction depends upon the distance of that prediction from the mean of the original time scores. These models, except in some special cases, require use of a Newton-Raphson type algorithm for maximum likelihood estimation. Examples of varying complexity show the utility of these methods.

Keywords: Log-linear models; Newton-Raphson algorithm.

1. Introduction

In this paper log-linear models are used for the time-series analysis of qualitative attributes. Attribute data observed over time are common in social research, perhaps the most common example being the repeated cross-sectional survey. Time-series of qualitative attributes form much of the empirical base for the study of "social indicators" [13, 15], and so we can expect this type of data to become even more common in the future. The reasons for analyzing such data are evidently two-fold. The first objective is usually to parameterize the time-trend actually exhibited over the interval spanned by observed data. Statistical models are necessary for this task, since our understanding of the past is usually conditional on sample (rather than population) characteristics. Hopefully, an economical interpretation of the past will emerge from the analysis of only a few parameters of well chosen models.

A second objective of time series analysis, closely related to the first, is the forecasting or "prediction" of the future. Our expectations for the future are often conditioned upon information about the past. Social and economic policy, by its very nature oriented to the future, is most wisely formulated when explicit forecasts (based on past experience) are readily at hand. Time series methods which satisfy these needs are not to our knowledge now available. For attribute or frequency data the more usual time-series methods [e.g., 4, 14] are not strictly appropriate. These other methods can, however, motivate the corresponding methods suited for attribute data. The methods which we propose are applied to labor force data from the Current Population Survey of the United States [5].

2. Time-Trend Models Using Polynomial Equations for the Logits

Let us begin by describing the data which will be analyzed here. Table 1 classifies the civilian population of the United States aged 14 and over into four mutually exclusive and exhaustive categories based upon sample data from the March Current Population Survey for years 1969-1973. These data are easily seen to comprise a 4 x 5 crosstable. The comumn variable will be denoted by a "T," referring explicitly to the Time variable, and has categories $t = 1,\ldots,5$. The scores attaching to the categories of the T variable could just as well be rearranged to -2, -1, 0, +1, +2, but in either alternative the ordered and interval nature of the T variable is to be taken into account.

For simplicity we shall regard each of the five time-period observations as simple random samples (i.e., multinomial samples), but actually these data derive from a very complicated sampling scheme. The methods presented here can be extended to deal with sampling arrangements different from simple random samples. The reader is referred to Haberman [10, 11] for the proper extensions to some situations of possible interest. Note should be taken in Table 1 of the marginals of the T variable, $f_{.t}$, since these are fixed by sampling design. Acceptable models for these data (i.e., models which will generate the frequencies in Table 1) will need to fit this marginal in order not to violate the sampling design.

The attribute under investigation here pertains to the labor force status of persons. These statuses (i.e., categories) will be unfamiliar to some, so we make brief comment about them here. We let U refer to the labor force status variable in the row of Table 1, and we denote its classes by $i = 1,\ldots,4$. Category 1 refers to "adequate employment," and was actually defined as a residual category left over after the measurement of categories 2, 3, and 4. Category 4 refers to "economic inactivity," and comprises all persons who are not seeking out work at the time of the survey. In most respects, this status is similar to the "not-in-labor-force" category widely used in federal statistics. Category 3 denotes a status which we shall refer to as "economic underemployment," and comprises all persons who are unemployed, part-time unemployed, or working full time but receiving sub-standard wages. Category 2 refers to persons whose work wages are satisfactory, but whose skill level (measured by years completed education) is considerably greater than the mean skill level (educational level) of other workers in similar occupations. The labor force statuses contained in Table 1 are not now part of the federal government's system of labor force statistics; even the nomenclature chosen to

describe the categories is different from that of customary labor force reports. A detailed justification for this scheme of measurement is presented in [5]. It will suffice here to note that Table 1 is an example of time series attribute data, and data of this kind appear often in social research. Even though the column variable T (perhaps specified as an "independent" variable) is quantitative, the row variable U (a "dependent" variable) is qualitative.

To begin a time series analysis of Table 1 we first consider a simplified table derived by combining the 1st and 4th categories of U and the 2nd and 3rd categories of U. The result is Table 2 where now the dichotomous row variable U has category 1 denoting "not underemployed" and category 2 denoting "underemployed."

The usual time series models begin with a quantitative variable y, scores for which are observed at t = 1,...,T points in time. The across time variation in y is then "explained" by certain kinds of linear models [e.g., 14]. If $\underline{y}_T$ represents the vector of observations, the standard approach is to consider a model

$$y_t = f(\underline{y}_T) + e_t, \quad t = 1,\ldots,T,$$

where the e_t are assumed to be normally distributed error terms with constant variance and zero autocorrelation. The functional form f in those applications with which we are familiar is a linear function chosen in such a way to ensure that the e_t have regular properties. E.g., autoregressive-moving average models (ARMA models) reduce to certain variations on the linear model shown above. When a suitable function f can be found to purge the error term of undesirable properties, the forecast of y into the future for any t' > T is given as the projection along the trend curve fit to the original observations. Of course, we could also use the estimated function f to provide interpolated values of y for points t' < T, if there were sufficient reason to believe that f could be used to predict the trend in y for all points interior to the T points actually observed in the data. The forecasted score (or the interpolated score) $y_{t'}$ will represent an "optimal" prediction to the extent to which the chosen function f has ensured regular properties to the disturbances, and to the extent to which time-trend observed in the past can serve as a prediction of scores which are not yet known.

One kind of time series model appropriate for the attribute data in Table 2 is based upon a trending of logits. Let the observed frequencies in Table 2 be denoted as f_{it} and the expected frequencies given some model as F_{it}, i = 1, 2; t = 1,...,T. First consider a model for the U x T cross-classification whereby the expected logits $\phi_t = \log (F_{1t}/F_{2t})$ are related to the time scores t = 1,...,T by the following polynomial equation:

$$\phi_t = A + B_1 t + \ldots + B_{T-1} t^{T-1} = \underline{t}^{(T-1)'} \underline{B}, \qquad (2.1)$$

where $\underline{B}' = (A, B_1,\ldots,B_{T-1})$ and $\underline{t}^{(T-1)'} = (1, t,\ldots,t^{T-1})$. Equation (2.1) is a polynomial of degree T-1 linking the expected logits of U to the time scores, and it will be desirable to find models which fit the data and in which several of the B_i are zero. Models of this kind are considered by Bock [2, Ch. 8], Goodman [9], and Haberman [10, 11], but by considering the time series nature of (2.1) we shall obtain some new results. In this approach to the analysis of time series we have made weaker assumptions about the distribution of the dependent variable (i.e., it is binomial) and can appeal to maximum likelihood methods generally associated with log-linear models.

Given (2.1) above, a forecast for the logits of U for time points t' > T is straightforward. First we find a suitable representation of the time trend in our observed table. Suppose this model is

$$\phi_t = \underline{t}^{(p)'} \underline{B}^{(p)}, \qquad (2.2)$$

where $\underline{t}^{(p)}$, $\underline{B}^{(p)}$ are subsets of $\underline{t}^{(T-1)}$, $\underline{B}^{(T-1)}$, respectively. The predicted logit is then merely

$$\phi_{t'} = \underline{t}^{(p)'} \underline{b}^{(p)}, \quad t' = T + 1, T + 2,\ldots, \qquad (2.3)$$

where $\underline{t}^{(p)}$ is the same subset of $\underline{t}^{(T-1)}$ that appeared in (2.2) with the modification that t' replaces t. The vector $\underline{b}^{(p)}$ is the sample estimate of $\underline{B}^{(p)}$. The predicted proportions in the i-th category of U at t' are given by

$$P_{1t'} = \exp(\phi_{t'})/(1 + \exp(\phi_{t'}))$$
$$P_{2t'} = 1 - P_{it'}. \qquad (2.4)$$

If it were of interest to interpolate values of $\phi_{t'}$ for t' interior to the T sample points, then (2.3) and (2.4) are modified accordingly.

To estimate the model for the expected frequencies implied by (2.1), a model for the logits, several different strategies present themselves. For the column variable T a set of T-1 orthogonal polynomials are required to define the vector basis of variable T. Direct products of these with the simple deviation contrast vector (1/2, -1/2) for U define the appropriate interaction terms. For the case where the categories of T are equally spaced, standard computer programs such as the ECTA program of Goodman and Fay and the MULTIQUAL program of Bock and Yates [3] provide the necessary orthogonal polynomials. For cases where the number of time scores T is of moderate size, these may also be found in common statistical tables (e.g., [8]). For cases where the time scores are not equally spaced, the orthogonal polynomials can be obtained from formulae

reported by Bliss [1, pp. 2-27]. For the saturated model with zero degrees of freedom (where all of the B_i in (2.1) may be nonzero), the parameters can be calculated directly from formulae to be presented later. For the unsaturated model obtained by setting all of the B_i at zero, a model equivalent to the usual independence hypothesis for the two-way table, the constant A can also be estimated by elementary means. For various other models obtained from (2.1) by setting some (but not all) of the B_i at zero, the implied models for the frequencies are not equivalent to models based upon the fitting of marginals, and so computational methods for determining the F_{it} and the B_i different from the iterative proportional scaling algorithm have to be employed.

A Newton-Raphson algorithm can be used to find the maximum likelihood estimate of the F_{it} and B_i of (2.1). The approach suggests itself by considering the log-linear model for the frequencies implied by the linear model for the logits reported in (2.1). Letting $u = (\log F_{11}, \log F_{21}, \ldots, \log F_{2T})'$ we find that this model is

$$\underline{u} = X\underline{\beta} \qquad (2.5)$$

where $\underline{u}$ is 2T X 1, X is 2T X 2T (in the saturated model), and $\underline{\beta}$ is the 2T X 1 vector of coefficients. For various unsaturated models corresponding to (2.2), (2.5) will be modified by replacing the X matrix of contrasts by a corresponding 2T X (T + P) matrix of contrasts. The vector of logits $(\phi_1, \phi_2, \ldots, \phi_T)$ is obtained by premultiplying $\underline{u}$ in (2.5) by a matrix C with elements $C_{i,2i-1} = 1$, $C_{i,2i} = -1$, and all other $C_{ij} = 0$. That is,

$$\underline{\Phi} = C\underline{u}. \qquad (2.6)$$

From (2.1) we find that

$$\begin{aligned}\underline{\Phi} = {} & A + B_1 t_1 + \ldots + B_{T-1} t_1^{T-1} \\ & A + B_1 t_2 + \ldots + B_{T-1} t_2^{T-1} \\ & \vdots \\ & A + B_1 t_T + \ldots + B_{T-1} t_T^{T-1} \\ = {} & Z\underline{B}, \end{aligned} \qquad (2.7)$$

implying that $\underline{B}$ in (2.1) is given simply by

$$\begin{aligned}\underline{B} &= Z^{-1} C \underline{u} \\ &= Z^{-1} C X \underline{\beta}. \end{aligned} \qquad (2.8)$$

For unsaturated models corresponding to (2.2) Z will be of order T X p, but (2.8) will nonetheless provide the maximum likelihood estimate of $\underline{B}^{(p)}$ if $\underline{u}$ is a vector of maximum likehihood estimates. Equation (2.8) makes explicit some of the formulae which appear in Haberman [11], and shows how the coefficients in (2.1) can be estimated from computer output (e.g., MULTIQUAL output) providing $X\,\hat{\underline{\beta}}$.

The variance of a predicted logit $\phi_{t'}$ is easily seen to be

$$\text{Var}\,(\phi_{t'}) = \underline{t}^{(p)\prime}\ \text{Var}\,(\underline{b}^{(p)})\ \underline{t}^{(p)}, \qquad (2.9)$$

a formula familiar from regression analysis. Note that in (2.9) the $\underline{t}$ vector is composed of powers of powers of t', regardless of the value of t' (i.e., regardless of whether t' is an observed time score or an unobserved time score). Furthermore, the formula in (2.9) allows the variance of the predicted logit to depend on the distance of the prediction from the mean of the observed time scores, unlike some other asymptomatic variance formulae which might be used here. From (2.8) we have $\underline{b}^{(p)} = Z^{-1} C X \hat{\underline{\beta}}$ where the Z and X matrices are defined appropriately, and so

$$\begin{aligned}\text{Var}\,(\underline{b}^{(p)}) = {} & Z^{-1} C X\,(\text{Var}\,(\hat{\underline{\beta}})) \\ & X' C' Z'^{-1}. \end{aligned} \qquad (2.10)$$

As shown in [10, 11], $\text{Var}\,(\hat{\beta}) = (X' D(F) X)^{-1}$ where D(F) is the diagonal matrix with expected frequencies on the diagonal. Finally, the variance of predicted proportions in (2.4) can be approximated by application of the delta method. This shows how the polynomial time-trend model may be estimated and how the precision of a forecast can be obtained from it.

In sum, the approach to the time series analysis of qualitative attributes suggested here seems well suited to the interpolation of logits (or proportions) between time points actually sampled, and to the extrapolation or forecasting of logits (or proportions) into the future. While this approach has not to our knowledge been previously applied, there is little that is new in the log-linear time trend models suggested here.

As a first example consider the data in Table 2 where the 1973 sample is ignored. We consider the problem of forecasting the distribution in 1973 from the time trend 1969-1972. For simplicity, we assign scores -1.5, -.5, +.5, +1.5 to the four time-periods included. This choice of time scores only affects the value of the constant term A in (2.1). In Table 3 the degrees of freedom and the fit of various models are presented. The model H_0 where $\hat{\phi}_t = a$, equivalent to an hypothesis of independence between U and T, produces a likelihood-ratio Chi-square of 491.79 on 3 df, contradicting this simplest time-trend hypothesis. Introducing a linear term produces model H_1 where $\hat{\phi}_t = a + b_1 t$. With $L^2(H_1)$ of 13.69 on 2 df we have achieved a remarkable improvement in fit with addition of

only a single parameter. On such a large sample size as this (total n over 400,000), such a fit is certainly acceptable, even though the descriptive level of significance is approximately .001. We find by application of formulae presented earlier that a = 1.7344 and b_1 = -.0844, the latter term reflecting the decrease in economic opportunity 1969-1972.

We find when using H_1 and substituting the value t' = 2.5 (corresponding to 1973) in the equation ϕ_t = 1.7344 -.0844t that $\hat{\phi}_{1973}$ = 1.5264, implying a predicted proportion underemployed in 1973 of .1785. The observed logit and the observed proportion underemployed in 1973 were 1.5270 and .1642, respectively. (See Table 4.) We see that by virtue of the upturn in the economy during 1973 we have overestimated the number of underemployed persons by 1.42%.

Model H_2 in Table 3 corresponds to ϕ_t = $a + b_1t + b_2t^2$, and we see that this model does not significantly reduce Chi-square. Model H_3 corresponds to a linear and a cubic (but not a quadratic) term in the model. With $L^2(H_3)$ = .84 on 1 df we see that this model fits the data very well indeed. For H_3 we find a 1.7377, b_1 = -.1325, and b_3 = .0235. The predicted logit for 1973 is 1.5974, considerably worse than our first prediction. For these data the standard error of the forecasted proportion underemployed in 1973 would be virtually nil. Given the time-trend 1969-1972, the upturn in the economy during 1973 was totally unexpected.

Table 5 presents log-linear time trend models for the full 2 x 5 crosstable in Table 2. The Chi-square of 69.90 for the model with a linear time-trend parameter would be acceptable for most purposes. We see that addition of a quadratic term adds substantially, however, to the goodness-of-fit.

We now consider models for the 4 x 5 crosstable presented earlier in Table 1. Models for this table are generalizations of the one considered in (2.1), taking account of the polytomous U (dependent) variable. Models of the form

$$\phi_{it} = \log (F_{it}/F_{4t}) = A_i + B_{i1}t + \ldots + B_{i4}t^4, \; i = 1, 2, 3, \qquad (2.11)$$

are appropriate when U is unordered. To estimate models of the kind in (2.11) we generate the matrix X of contrasts in (2.5) by again using orthogonal polynomials for the T variable and using deviation contrasts implied by (2.11) for the U variable (see [3]). By following a hierarchy principle we might focus upon a subset of the wide range of models open to our choice where if B_{ik} = 0, then $B_{ik'}$ = 0 for k' > k, i = 1, 2, 3. The fit of some of these models is presented in Table 6. By restricting our attention to models of the kind in (2.11), interpolation or extrapolation is also straightforward, and can be carried out with the aid of the formulae presented earlier. We see from Table 6 that the model with only the linear terms B_{11}, B_{21}, B_{31}, is adequate (accounting for 79% of the variation in the data), but also that the inclusion of quadratic terms contributes in a substantial way to explaining time trend.

3. A Model Allowing Autocorrelation of the Logits

The models considered in the previous section linked the observed logits (or predicted logits) to scores reflecting the spacing of the time variable. For purposes of interpolation those models appear satisfactory. However, for purposes of forecasting (or extrapolation) beyond time points actually observed, the previous models can lead to unacceptable results. For example, in the analysis of the 2 x 4 crosstable presented in Tables 2 and 3, we found that $\hat{\phi}_t$ = 1.7374 -.0844t provided an acceptable summary of observed time trend. If we were to entertain this model seriously for purposes of forecasting, then the predicted logit for t' = 20.6 (=1.7374/.0844) would be zero, and the predicted distribution of the attribute would be a degenerate one. In this section we briefly consider a model which does not suffer this difficulty. This model is motivated by the simple autocorrelation model associated with the analysis of time series of quantitative variables [4], and suggests an alternative way of viewing time series attribute data.

A model where the expected logit at time t $\hat{\phi}_t$ depends only on the observed logit at time t-1, ϕ_{t-1} is the "first order autocorrelation of logits" model, viz.,

$$\hat{\phi}_t = \rho\phi_{t-1}, \qquad (3.1)$$

where ρ is the "autocorrelation" parameter. As in the usual time series approach to (3.1) (where the corresponding quantitative scores are substituted for the logits), the initial observation at t = 1 is considered as a given, and so we find that $\hat{\phi}_1 = \phi_1$, implying further that $f_{11} = F_{11}$, $f_{21} = F_{21}$. The model in (3.1) thus has some characteristics of a "quasi-independence" model, since the relation in (3.1) only pertains to a subset of the cells in the complete table. A least squares procedure, which in this case provides estimates almost equivalent to maximum likelihood, produced the results presented in Table 7. The model in (3.1) has an L^2 of 43.14 on two degrees of freedom, and provides an estimate of ρ of .9549. The predicted logit for 1973 is closer to the observed logit than was the case for the models considered in Section 2. (Cf. Table 4.) The advantage of model (3.1) is that forecasts of $\phi_{t'}$ for finite t' will result in nondegenerate predicted distri-

butions of the attribute. Because of this property these models deserve further consideration. Models of the kind in (3.1) can be modified to deal with certain other kinds of time series models (e.g., moving average models). We do not go into those details here. [Tables 5, 6 and 7 are available upon request from the writer]

REFERENCES

1. Bliss, C. I., Statistics in Biology, Vol. 2, New York: McGraw-Hill, 1970.

2. Bock, R. Darrell, Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill, 1975.

3. Bock, R. Darrell, and Yates, George, MULTIQUAL: Log-Linear Analysis of Nominal or Ordinal Qualitative Data by the Method of Maximum Likelihood, Chicago: National Educational Resources, 1973.

4. Box, G. E. P. and Jenkins, G. M., Time Series Analysis, Forecasting and Control, San Francisco: Holden-Day, Inc., 1970.

5. Clogg, Clifford C., "Measuring Underemployment: Demographic Indicators for the U.S. Labor Force, 1979-1973," Ph.D. Dissertation, University of Chicago, 1977. Forthcoming by Academic Press.

6. Cochran, W. G. and Cox, G. M., Experimental Design, 2nd ed., New York: Wiley, 1957.

7. Cox, D. R., The Analysis of Binary Data, London: Methuen, 1970.

8. Fisher, R. A. and Yates F., Statistical Tables for Biological, Agricultural, and Medical Research, 6th ed., New York: Hafner, 1963.

9. Goodman, Leo A., "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," Technometrics 13, (1971):33-61.

10. Haberman, Shelby J., The Analysis of Frequency Data, Chicago: University of Chicago Press, 1974.

11. ------, "Log-Linear Models for Frequency Tables with Ordered Classifications," Biometrics 30 (1973):589-600.

12. Kendall, M. G. and Stuart A., The Advanced Theory of Statistics, 3rd ed., New York: Hafner, 1973.

13. Land, Kenneth and Spilerman, Seymour, eds., Social Indicator Models, New York: Russel Sage Foundation, 1975.

14. Nelson, Charles R., Applied Time Series Analysis for Managerial Forecasting, San Francisco: Holden-Day, 1973.

15. Stone, Richard, Demographic Accounting and Model Building, Paris: Organization for Economic Co-Operation and Development, 1971.

Table 1. Labor Force Status Over Time, Civilian Population Aged 14 and Over, 1969-1973.

Source: March Current Population Survey

	YEAR				
Labor Force Status	1969	1970	1971	1972	1973
1. Adequate Employment	48017 (44.2%)	45299 (43.6%)	44373 (41.8%)	42811 (41.7%)	42350 (42.1%)
2. Mismatch	5640 (5.2%)	5560 (5.4%)	6219 (5.9%)	6363 (6.2%)	6766 (6.7%)
3. Economic Underemployment	8971 (8.3%)	9184 (8.9%)	10571 (10.0%)	10592 (10.3%)	9748 (9.7%)
4. Not-in-Labor-Force	45887 (42.3%)	43705 (42.1%)	44956 (42.3%)	42939 (41.8%)	41685 (41.5%)
Total	108,515	103,748	106,119	102,705	100,549

Table 2. 2 X 5 Cross-Classification of Labor Force Status Over Time

Source: Table 1

	YEAR				
	1969	1970	1971	1972	(1973)
Not Underemployed [a/]	93904	89004	89329	85750	(84035)
Underemployed [b/]	14611	14744	16790	16955	(16514)
Total	108,515	103,748	106,119	102,705	(100,549)

[a/] Not Underemployed = Adequately Employed or Not-in-Labor-Force.

[b/] Underemployed = Mismatched or Economic Underemployed

Table 3. Log-Linear Time-Trend Models for the 2 X 4 Table (Ignoring 1973)

Model	Likelihood-Ratio Chi-Square	Goodness-of-Fit Chi-Square	Degrees of Freedom
H_o: $B_1 = B_2 = B_3 = 0$	491.79	491.12	3
H_1: $B_2 = B_3 = 0$	13.69	13.71	2
H_2: $B_3 = 0$	12.69	12.95	1
H_3: $B_2 = 0$	.48	.48	1

Table 4. Observed Logits Log (p_{1t}/p_{2t}) and Expected Logits Log $(\hat{p}_{1t}/\hat{p}_{2t})$ [a/] From Model H_1.

	1969	1970	1971	1972	(1973)
Observed	1.8605	1.7978	1.6715	1.6209	(1.6270) [b/]
Expected	1.8640	1.7796	1.6952	1.6108	(1.5264) [c/]

[a/] Expected logits obtained from $\Phi_t = 1.7374 - .0844t$.

[b/] Proportion underemployed in 1973 = .1642.

[c/] Predicted proportion underemployed in 1973 = .1785.