

Tim Futing Liao
Department of Sociology
University of Essex
Colchester C04 3SQ
United Kingdom
44 1206 872663 (phone)
44 1206 873410 (fax)
tfliao@essex.ac.uk
Word count: 2,659

Categorical Data Analysis. Social science data, when quantified, can be subject to a variety of statistical analyses, most common of which is REGRESSION ANALYSIS. However, the requirements of a continuous dependent variable and of other REGRESSION ASSUMPTIONS make linear regression sometimes a less desirable analytic tool because a lot of social science data are categorical.

Social science data can be categorical in two common ways. First, a variable is categorical when it records nominal or discrete groups. In political science, a political candidate in the United States can be a Democrat, Republican, or Independent. In sociology, one's occupation is studied as a discrete outcome. In demography, one's contraceptive choice such as the pill or the condom is categorical. Education researchers may study discrete higher education objectives of high school seniors: university, community college, or vocational school.

Furthermore, a variable can take on an ordinal scale such as the LIKERT-TYPE SCALE or a scale that resembles it. Such a scale is widely used in psychology in particular and in the social sciences in general for measuring personality traits, attitudes, and opinions and typically has five ordered categories ranging from *most important* to *least important* or from *strongly agree* to *strongly disagree* with a neutral middle category. An ordinal scale has two major features: There exists a natural order among the categories, and the distance between a lower-positioned category and the next category in the scale is not necessarily evenly distributed throughout the

scale. Variables not measuring personality or attitudes can also be represented with an ordinal scale. For example, one's educational attainment can be simply measured with the three categories of "primary," "secondary," and "higher education." These categories follow a natural ordering, but the distance between the "primary" and "secondary" and that between "secondary" and "higher education" are not necessarily equal. It is apparent that an ordinal variable has at least three ordered categories. While there is no upper limit for the total number of categories, researchers seldom use a scale of more than seven ordered categories.

The examples of categorical data above illustrate their prevalence in the social sciences. Categorical data analysis in practice is the analysis of categorical *response* variables.

HISTORICAL DEVELOPMENT

The work by Karl Pearson and G. Udny Yule on association between categorical variables at the turn of the 20th century paved the way for later development in models for discrete responses. Pearson's contribution is well-known through his namesake statistic, while Yule was a strong proponent of the ODDS RATIO in analyzing association. However, despite important contributions by noted statisticians such as R. A. Fisher and William Cochran, categorical data analysis as we know it today did not develop until the 1960s.

The postwar decades saw a rising interest in explaining social issues and a burgeoning need for skilled social science researchers. As North American universities increased in size to accommodate postwar baby boomers, so did their faculties and the bodies of university-based social science researchers. Categorical scales come naturally for measuring attitudes, social class, and many other attributes and concepts in the social sciences. Increasingly in the 1960s, social surveys were conducted and otherwise quantifiable data were obtained. The increasing

methodological sophistication in the social sciences satisfied the increasing need for analytic methods for handling the increasingly available categorical data. It is not surprising that many major statisticians who developed regression-type models for discrete responses were all academicians with social sciences affiliations or ties, such as Leo Goodman, Shelby Haberman, Frederick Mosteller, Stephen Fienberg, and Clifford Clogg. These methodologists focused on loglinear models while their counterparts in the biomedical sciences concentrated their research on regression-type models for categorical data. Together, they and their biomedical colleagues have taken categorical data analysis to a new level.

The two decades before the 21st century witnessed a greater concern with models studying association between ordinal variables especially when their values are assumed to be latent or not directly measurable. The same period also saw a resurgence of interest in latent variable models studied by Lazarsfeld in the 1950s (e.g., latent class analysis) and their extension to regression-type settings. At the forefront of these two areas were Leo Goodman, Clifford Clogg, and Jacque Hagenars, who represented the best of both North American and European social science methodologists.

Another tradition in categorical data analysis stems from the concerns in econometrics with a regression-type model with a dependent variable that is limited in one way or another (in observed values or in distribution), a notion popularized by G. S. Maddala. When only two values (binary) or a limited number of ordered (ordinal) or unordered (nominal) categories can be observed for the dependent variable, LOGIT and PROBIT models are in order. During the past quarter century, computer software mushroomed to implement these binary, ordinal, and multinomial logit or probit models. In statistics, many models for categorical data such as the

logit, the probit, and the Poisson regression as well as the linear regression can be more conveniently represented and studied as the family of GENERALIZED LINEAR MODELS.

TYPES OF CATEGORICAL DATA ANALYSIS

There are several ways to classify categorical data analysis. Depending on the parametric nature of a method, two types of categorical data analysis arise.

1. *Nonparametric methods.* These methods make minimal assumptions and are useful for hypothesis testing. Examples include PEARSON CHI-SQUARE, FISHER'S EXACT TEST (for small expected frequencies), and the MANTEL-HAENSZEL TEST (for ordered categories in linear association).
2. *Parametric methods.* This model-based approach assumes random (possibly stratified) sampling and is useful for estimation purposes and flexible for fitting many specialized models (e.g., symmetry, quasi-symmetry). It also allows the estimation of the STANDARD ERROR, COVARIANCE, and CONFIDENCE INTERVAL of model parameters as well as predicted response probability.

The second type is the most popular in the social sciences, evidenced by the wide application of parametric models for categorical data. Variables in social science research can occupy dependent or independent positions in a statistical model. They are also known as response and explanatory variables, respectively. The type (categorical or not) and the positioning of a variable (dependent or independent) give rise to the statistical model(s) for (categorical) data, providing another way to classify categorical data analysis. For simplicity, we view all variables as either categorical or continuous (i.e., with interval or ratio scales). In some models, all variables can be considered dependent as only association among them is of interest

to the researcher. For the models that distinguish between dependent and independent variables, we limit our attention to those with single dependent variables. That is, in such models there is only one dependent variable per model regardless of the number of independent variables, which can be all continuous, all categorical, or a mixture of both.

[TABLE 1 ABOUT HERE]

Models of association do not make a distinction between response and explanatory variables in the sense that the variables under investigation depend on each other. When both variables in a pair are continuous, we use correlation analysis; when they are categorical, we use contingency table and loglinear analysis. The situation can be generalized to multiple variables as in the case of partial correlation analysis and multiway contingency tables.

When the two variables in a pair are of different types (i.e., one categorical and the other continuous), a causal order is assumed with the separation of the independent and the dependent variables. For analyzing the distribution of a continuous response variable in explanatory categories, we apply ANOVA or linear regression, which can be extended into including additional explanatory variables that are continuous. In that case, we use ANCOVA and linear regression. The flexibility of logit and probit models is evidenced by their applicability to all the three cells for a categorical response variable regardless of the type of explanatory variables—continuous, categorical, or mixed. Often the same data may be analyzed using different methods, depending on the assumptions the researcher is willing to make and the purposes of the research.

Thus, the cells in the column under the heading of categorical dependent variables comprise the methods of categorical data analysis. A further consideration of these methods is whether the response variable has ordered categories as opposed to pure nominal ones. Excluded

from the table is Poisson regression that models an integer dependent variable following a Poisson distribution, which is also a limited-dependent-variable model.

Depending on whether a variable is discrete or ordinal, a model for categorical data can be further classified. For example, a logit (or probit) model is binary when the dependent variable is dichotomous, is ordinal (or ordered) when the dependent variable has ordered categories, and is multinomial when it has more than two discrete categories. Loglinear models for association between ordinal variables can be extended into linear-by-linear association and log-multiplicative models, to name just two examples. See `LOGLINEAR MODEL` and `ASSOCIATION MODEL` for further details.

We have so far examined categorical data analysis with observed variables only. Social scientists also employ in categorical data analysis the `LATENT VARIABLE` to represent concepts that cannot be measured directly. Here we have yet another way of classifying categorical data analysis, depending on whether the latent variables or the observed variables or both in a model are categorical.

[TABLE 2 ABOUT HERE]

Specifically, categorical data analysis is concerned with the second row in Table 2 where the observed variables are categorical. `LATENT CLASS ANALYSIS`, which estimates categorical latent classes using categorical observed variables, has received most attention in the social sciences.

Several major computer programs have made the latent class model much more accessible by many social scientists. These include C. C. Clogg's mainframe program `MLLSA`, which is currently available in S. Eliason's `Categorical Data Analysis System` at the University

of Minnesota, S. Haberman's NEWTON and DNEWTON, J. Hagenaars's LCAG, and J. Vermunt's LEM, among others.

MODEL FITTING AND TEST STATISTICS

Yet one more way to view and classify categorical data analysis looks at how model fitting is defined. Much of current social science research using categorical data derives from three general yet related traditions: one that fits expected frequencies to observed frequencies, another that fits expected response values to observed response values, and a last that fits expected measures of association through correlation to the observed counterparts. The third tradition describes how categorical data are handled in factor-analytic and structural equation modeling approaches, and it receives separate attention elsewhere in the encyclopedia. We focus here on the first two traditions.

The first tradition fits a particular statistical model to a table of frequencies. Let f_i indicate observed frequency in cell i and F_i indicate expected frequency in cell i in the table. Then our purpose in model fitting is to reduce some form of the difference between the observed and the expected frequencies in the table. Thus, we obtain the PEARSON χ^2 STATISTIC and the LIKELIHOOD RATIO STATISTIC L^2 (or G^2):

$$\chi^2 = \sum_i \frac{(f_i - F_i)^2}{F_i} \quad \text{and} \quad L^2 = 2 \sum_i f_i \log\left(\frac{f_i}{F_i}\right).$$

The summation is over all cell i except the ones with STRUCTURAL ZEROS. These statistics can be used alone in a nonparametric method for a table of frequencies or in a parametric method such as a loglinear model, which typically expresses $\log(F_i)$ as a linear combination of the effects of the rows, columns, layers, and so on, and their interactions, of a table. The two test statistics given above are actually special cases of a family of test statistics known as power divergence

statistics, which includes a third member, the CRESSIE-READ STATISTIC, that represents a compromise between χ^2 and L^2 .

The second tradition is that of the generalized linear model, which attempts to model the expected value μ of the response variable y as a function of a linear combination of the explanatory variables and their parameters. The data are expected to follow a form of exponential distribution, and the function linking $g(\mu)$ to the linear combination of the independent variables and parameters may take on various forms, such as the widely used function of the logit.

The two traditions are closely related. For example, when all variables—response and explanatory—are categorical, $F_i = n_i \mu_i$, where n_i is the number of observations in cell i . The test statistics described above also apply to the assessment of model fitting of generalized linear models.

MODEL COMPARISON AND SELECTION

This is a prominent issue in categorical data analysis. Researchers often feel the need to compare statistical models and to choose a better fitting model from a set of models. The latter situation describes hierarchical loglinear model fitting when a number of loglinear models embedded in one another are considered. These models have different degrees of freedom because they include a different number of parameters. Differences in L^2 give comparative likelihood ratio statistics for drawing conclusions.

Sometimes researchers need to compare models that are nonhierarchical. This can be a comparison of the model fitting of two (or more) subgroups of observations or two (or more) models that include different sets of variables. While sometimes treating these subgroups or

submodels as members of a single combined supermodel is a useful exercise, this approach is not always realistic because the submodels under comparison can be just incompatible. Therefore, a statistical criterion of a model's goodness of fit that applies across different models with different degrees of freedom is necessary. Two widely applied criteria belonging to this category are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

Relying on information-theoretic considerations, the AIC adjusts the likelihood-ratio statistic with twice the number of degrees of freedom in a given model:

$$\text{AIC} = L^2 + 2(df),$$

where df represents the number of degrees of freedom in the model. Between two models, hierarchical or not, the model with the smaller AIC value is the better one.

Based on Bayesian decision theory, the BIC further adjusts the likelihood ratio statistic with the total sample size (N):

$$\text{BIC} = L^2 + \log(N)(df).$$

The BIC gives an approximation of the logarithm of the Bayes factor for model comparison. It has been shown that the BIC penalizes the model with a larger degree of freedom more so than the AIC if sample sizes are not very small. When comparing hierarchical models where L^2 differences are also applicable, the BIC usually favors a parsimonious model more so than the other goodness-of-fit statistics.

MISSING DATA IN CATEGORICAL DATA ANALYSIS

Missing data can be a serious concern in any multivariate analysis and deserve particular attention in models of frequency tables that may have zero cells. Of special interest are the so-called partially observed data. If an observation has missing information on all variables, the

observation cannot be used at all. But if an observation has missing information on some of the variables in the model, the conventional approach of listwise deletion would waste useful information. Added to this problem is the mechanism with which missing information occurs. Of the three mechanisms of missing data—missing completely at random, missing at random, and missing not at random—the last may create bias in drawing inference about relationships between variables in the model. Developments in recent years have sought to deal with estimation of statistical models of categorical data under various missing-data mechanisms. The LEM software by J. Vermunt is the most flexible for such purposes.

TIM FUTING LIAO

References

- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.
- Clogg, C. C., and Shihadeh, E. S. (1994). *Statistical Models for Ordinal Variables*. Thousand Oaks, CA: Sage.
- Everitt, B. S. (1992). *The Analysis of Contingency Tables*. London: Chapman & Hall.
- Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Table 1: Correspondence between Variable Types and Statistical Models

		Dependent	
		Continuous	Categorical
Dependent	Continuous	Correlation	
	Categorical		Contingency table analysis, Loglinear models
Independent	Continuous	Linear regression	Logit and probit models
	Categorical	ANOVA, linear regression	Logit and probit models
	Mixed	ANCOVA, linear regression	Logit and probit models

Table 2: Types of Latent Variable Models According to the Type of Variables

	Latent continuous	Latent categorical
Observed continuous	Factor analytic model; SEM	Latent profile model
Observed categorical	Latent trait model; SEM	Latent class model