AN ANALYSIS OF DATA FROM A Y2K HOUSEHOLD SURVEY

Terry L. Kissinger, David W. Chapman, Federal Deposit Insurance Corporation
Terry L. Kissinger, 550 17th St. NW, Washington, D.C. 20429

Key Words: Logistic Regression, Cumulative Logits, Proportional Odds, Latent Variables, SUDAAN

Introduction

In late 1999, The Federal Deposit Insurance Corporation (FDIC) and the Federal Reserve Board (FRB) co-sponsored a survey of U.S. adult residents that have accounts in banks, savings and loan institutions, and credit unions. The main purpose of the survey was to determine how concerned people were about banking-related problems associated with the year 2000 (Y2K) "computer bug" and how they planned to prepare for possible Y2K banking problems.

This survey, which was conducted over the telephone, began on October 19, 1999, and ended December 31, 1999. It was conducted in three separate periods, with weighting provided to yield estimates of population totals for each of the three periods.

Of particular interest in analyzing the survey data was whether certain demographic groups were especially concerned about Y2K banking problems. For such groups, this could have been an indication that they were more likely than other groups to take extreme action in preparing for Y2K, such as withdrawing all their money from their accounts. After such groups were identified, they could be targeted with ad hoc public relations campaigns to help allay their concerns.

This paper presents an analysis of the survey data using cumulative logits, with sampling design taken into account using the statistical software package SUDAAN (Shah, Barnwell, & Bieler, 1997). The sample design is described briefly, then the explanatory and response variables used in the analysis are defined. A presentation of the analytical methodology is given, indicating the usefulness of modeling with cumulative logits in this situation. Then the results are summarized, with suggestions for further analysis of the data provided.

Sampling Design

The sample for the Y2K survey of U.S. adult depositors was a random digit dialing (RDD) sample of ten-digit telephone numbers, conducted by the Gallup Organization under contract to FDIC. The specific method of RDD used was a "list-assisted 3+" procedure.

With this procedure, Gallup identified all clusters of 100 phone numbers (referred to as "100-banks") for which there were at least three listed numbers from U.S. residential telephone directories. Once a 100-bank was determined to contain at least three listed residential numbers, all 100 numbers in the 100-bank were eligible for selection. All telephone numbers included in the frame were assigned equal probabilities of selection.

After the sampling frame was assembled, systematic samples of telephone numbers for the following three periods were selected: (1) October 19 through November 12, 1999; (2) November 13 through December 12, 1999; and (3) December 13 through December 31, 1999. These three samples were selected as part of the same sampling process, with no overlap between any two periods.

Each ten-digit number selected for the sample was called to determine if it was a residential telephone number and, if so, whether the residence contained at least one adult aged 18 or more. One adult from each such residence was randomly selected for screening and interviewing, as identified by the most recent birthday. Adults who (1) did not have an account at a bank, savings and loan institution, or credit union or (2) reported having seen or heard "nothing at all" about the Y2K computer issue were not eligible for the survey.

The weight assigned to each survey respondent was the base sampling weight adjusted to align with estimates of population totals from the March 1998 Current Population Survey. The base sampling weight was the reciprocal of the initial selection probability. The initial selection probability was the product of three factors: (1) the initial sampling rate, used by Gallup to obtain the initial sample of ten-digit numbers; (2) the reciprocal of the number of adults in the household; and (3) the number of separate residential telephone numbers in the household.

The adjustments of the base sampling weights to align with estimates of population totals were derived separately in each of 48 cells, defined by four Census regions, two gender categories, two age categories, and three race/education categories. The estimate of the population total for each of these 48 cells was obtained by multiplying the March 1998 Current Population Survey estimate of the number of adults in the cell by the estimated eligibility rate. The estimated eligibility rate was derived by Gallup from fitting a logistic regression model to the survey data.

Response and Explanatory Variables

For the full data set, there were nine cumulative logit models fitted, relative to nine response variables regarding attitudes about the Y2K computer bug and related banking problems. Each response variable was a question or question subpart from the survey instrument. Table 1 gives the relevant question and question subparts and the corresponding number of ordinal levels used as response categories. (Each question or question subpart also had "don't know" and "refused" as possible answers.)

Table 1
Question and Question Subparts Used as Response Variables and Their Corresponding Numbers of Ordinal Response Levels

Questions and Question Subparts	Levels
Q3: Overall, how concerned are you about the Y2K computer issue? Would you say you are…	4
Q4: Compared to a month ago, would you say you are now ____ about the Y2K computer issue?	3
Q5: Some people say the Y2K computer issue might have an impact on banks. Please tell me how likely you think it is that each of the following banking problems will result from the Y2K computer issue.	(See Sub-parts)
Q5A: ATMs will not work.	5
Q5B: Direct deposit payments, such as social security checks, pension checks, or payroll checks, will not be properly credited to bank accounts.	5
Q5C: People will temporarily lose access to their money.	5
Q5D: Credit card systems will not work.	5
Q5E: Checks will not be accepted or processed properly.	5
Q7: How confident are you that your bank is ready for the year 2000?	4
Q13: How likely are you to keep some extra cash on hand because of the Y2K computer issue? Would you say…	5

For the ordinal response scales with an odd number of categories, low response categories indicated relatively high concern about Y2K, the middle response category was neutral, and high response categories indicated relatively low concern. For example, for each of the five subparts of Q5, the response categories were (1) definitely will happen; (2) probably will happen; (3) you are uncertain; (4) probably will not happen; and (5) definitely will not happen.

The ordinal response scales with an even number of categories were similar, only they did not have a neutral middle response category. For example, the response categories for Q7 were (1) not confident at all; (2) not too confident; (3) somewhat confident; and (4) very confident.

There were eight explanatory variables used in each cumulative logit model, seven corresponding to demographic variables and one corresponding to the three periods of the survey. Table 2 gives the explanatory variables used and their levels. (Race may be perceived as a combined race/ethnicity variable, since one level is Hispanic.)

Table 2
Explanatory Variables and Their Levels

Explanatory Variables	Levels
Gender	(1) Male; (2) Female
Age (in Years)	(1) 18-25; (2) 26-34; (3) 35-54; (4) 55-64; (5) 65 or more
Race	(1) African-American; (2) Hispanic; (3) White; (4) Other
Highest Level of Education Completed	(1) Less than High School; (2) High School Graduate; (3) Some College; (4) College Graduate
Total Annual Household Income	(1) < \$25,000; (2) \$25,000-49,000; (3) \$50,000-\$74,000; (4) \$75,000-\$99,000; (5) ≥\$100,000
Region	(1) Northeast; (2) Midwest; (3) South; (4) West
Urbanization	(1) Urban; (2) Suburban; (3) Rural
Period	(1) 10/19-11/12; (2) 11/13-12/12; (3) 12/13-12/31

The specific levels of each explanatory variable were chosen as being of special interest by a committee of staff and management of the two regulatory agencies co-sponsoring the survey. (Age, however, was recorded in number of years without grouping in the data set.) "Unknown" categories were also included for age, race, education, and income. There was no item nonresponse for gender, region, urbanization, or period.

Analytical Methodology

The logistic regression model fitted to the survey data was a cumulative logit model, also known as a proportional odds model because of a useful property of the model (McCullagh, 1980), discussed below. As noted by Fahrmeir and Tutz (1994), this model is based on the category boundaries or threshold approach, which dates back at least to Edwards and Thurstone (1952).

To understand the interpretation of this approach in the context of the survey data, following the presentation given by Allison (1999), suppose a response variable such as "Q5A: ATMs will not work" is based on some underlying latent variable Z that is continuous. This latent variable is not observed directly, but is recorded in terms of category boundaries or thresholds $\{\theta_1 > \theta_2 > \theta_3 > \theta_4\}$ that are used to transform Z into the observed variable Y with five ordinal categories.

Thus,

Y=1="Definitely will happen" if $\theta_1 < Z$,

Y=2="Probably will happen" if $\theta_2 < Z \le \theta_1$,

Y=3="You are uncertain" if $\theta_3 < Z \le \theta_2$,

Y=4="Probably will not happen" if $\theta_4 < Z \le \theta_3$,

and

Y=5="Definitely will not happen" if $Z \le \theta_4$.

The latent response variable Z is assumed to depend on the explanatory variables $\mathbf{X} = \left[X_1 X_2 \ldots X_p\right]^T$ according to the linear model $Z = \alpha^* + \left(\mathbf{B}^*\right)^T \mathbf{X} + \sigma\varepsilon$. Here ε is an error term following a standard logistic distribution, such that $f(\varepsilon) = \dfrac{e^{\varepsilon}}{\left(1+e^{\varepsilon}\right)^2}$ and $F(\varepsilon) = \dfrac{e^{\varepsilon}}{1+e^{\varepsilon}}$. Both α^* and σ are scalar parameters, and $\mathbf{B}^* = \left[\beta_1^* \beta_2^* \ldots \beta_p^*\right]^T$ is a vector of parameters.

The observed response variable Y has five categories, as defined by the four thresholds, with probability p_i pertaining to the probability of an observation falling into the i[th] category. Cumulative probabilities are given by $F_i = \sum_{j=1}^{i} p_j$ for i=1,2,3,4,5.

Using a cumulative link function, the cumulative logit model is then given by $\ln\left(\dfrac{F_i}{1-F_i}\right) = \alpha_i + \mathbf{B}^T \mathbf{X}$, where the parameters given by $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and the vector $\mathbf{B} = \left[\beta_1 \beta_2 \ldots \beta_p\right]^T$ are related to the parameters for the latent variable Z as follows:

$$\alpha_i = \frac{\alpha^* - \theta_i}{\sigma} \text{ for i=1,2,3,4}$$

and

$$\mathbf{B} = \frac{\mathbf{B}^*}{\sigma}.$$

Hence, testing the null hypothesis that β_j is equal to 0 is equivalent to testing whether β_j^* is equal to 0.

A nice feature of the cumulative logit model is that the parameters $\mathbf{B} = \left[\beta_1 \beta_2 \ldots \beta_p\right]^T$ have a convenient and useful interpretation, particularly when there are no interaction terms. Specifically, for a dichotomous explanatory variable X_j and for any i=1,2,3,4, B_j is the natural logarithm of the ratio of the odds of being in category i or less when $X_j = 1$ to the odds when $X_j = 0$, keeping other explanatory variables constant. Thus, e^{B_j} is the cumulative odds ratio for X_j for any choice of cumulative odds, adjusted for the effects of other explanatory variables in the model. The fact that the interpretation of B_j does not depend on i is known as the proportional odds assumption of this model.

A related feature of this model is that $\mathbf{B} = \left[\beta_1 \beta_2 \ldots \beta_p\right]^T$ does not depend on the placement of the thresholds $\{\theta_1, \theta_2, \theta_3, \theta_4\}$. Thus, with response variables that have ordinal (but not interval) scales, the cumulative logit model is a relatively parsimonious model with easily interpretable parameters.

The standard logistic distribution may be shown to be related to the exponential distribution, the Gumbel distribution, the Pareto distribution, and the power function distribution (Evans, Hastings, & Peacock, 1993). If a standard normal distribution is assumed for ε instead of a standard logistic distribution, a cumulative probit model is obtained.

Cumulative logit and cumulative probit models provide similar fits, due to the similarity of logistic and normal distributions. Parameters in cumulative logit models are simpler to interpret, however (Agresti, 1990). Nevertheless, the cumulative probit model is often preferred in Bayesian analyses because sampling from its posterior distribution is particularly efficient (Johnson & Albert, 1999).

If a standard Gumbel distribution is assumed for ε, a cumulative complementary log-log model is obtained. This model is often referred to as a grouped Cox model because it may be derived as a grouped version of the proportional hazards model in survival analysis (Fahrmeir & Tutz, 1994). Agresti (1990) gives conditions under which this type of model may be preferred.

While the cumulative logit model may be the most commonly used model for an ordinal response variable, other choices for defining the logits for a multinomial response variable are given by Fienberg (1980), Christensen (1997), and Hosmer and Lemeshow (2000). Cumulative logit models making the proportional odds assumption are not equivalent to loglinear models (Agresti, 1990).

Results

The numbers of completed interviews for the three periods were 1,326, 1,748, and 1,283, respectively. The unit response rate for the entire survey was about 53 percent. The item nonresponse rate for each response variable was 4 percent or less, with Q5A and Q5D having the largest item nonresponse rates. (Item nonresponse for a response variable was defined as a response of "don't know" or "refusal.")

Among the explanatory variables, there were no missing data for gender, region, urbanization, or period. The item nonresponse rate was 3 percent or less for race, age, and education. Income, with an item nonresponse rate of 12 percent, was the only variable used in the analysis with a relatively large item nonresponse rate.

Nine cumulative logit models were fit to the data, using each of the eight explanatory variables. The reference levels were as follows: male (gender), 65 years or older (age), white (race), college graduate (education), $100,000 or more (income), Midwest (region), rural (urbanization), and Period 3 (period). These reference levels were chosen because in general they were the levels indicating the least concern about Y2K-related banking problems. For each model, observations for which there was item nonresponse for the response variable were omitted.

Parameters were estimated using the Taylor linearization method with the SUDAAN statistical package. Region and period were treated as stratification variables, using the stratified with replacement (STRWR) design option, as recommended in such analytical problems by Lehtonen and Pahkinen (1996).

Tests for an association between each explanatory variable and each response variable were conducted, using the modified Wald test statistic for an appropriate contrast. Table 3 on the following page gives the p-value for each test.

The modified Wald test was used instead of the likelihood ratio test because SUDAAN does not provide likelihood ratio statistics adjusted for design effects. Likelihood ratio tests typically yield more power than Wald tests and generally perform better, especially in small samples or samples with unusual data patterns (Hauck and Donner, 1977; Jennings, 1986). Also, SUDAAN uses only an approximation to the likelihood function, as shown by Hosmer and Lemeshow (2000) for a binary response setting.

Additionally, SAS provides a score test for the proportional odds assumption, but this test is not provided by SUDAAN. Allison (1999) notes that the score test in SAS may reject the null hypothesis that the proportional odds assumption is appropriate more often than is warranted, typically producing p-values less than 0.05 when testing with many explanatory variables and a large sample size.

Backward elimination using the modified Wald test statistics was also used. With a 0.05 level of significance for each test, this method led to the same explanatory variables being significantly associated with each response variable, except that region was also significantly associated with Q4.

It may be seen from Table 3 that gender, age, race, education, and income were associated with nearly every response variable. Additionally, period was associated with Q7. An examination of cumulative odds ratios for the model with Q7 as the response variable indicated that people were more confident that their bank was ready for the year 2000 in successive periods, possibly the result of mailings by financial institutions to their depositors during that time.

Table 4 on the next page gives estimated cumulative odds ratios for gender, age, race, education, and income, comparing reference levels with other levels using the full models. (Cumulative odds ratios for unknowns are not given.)

It is apparent from Table 4 that certain demographic groups were especially concerned about Y2K banking problems. Specifically, females, young age groups, minority racial and ethnic groups, non-college attendees, and low income groups all had relatively high levels of concern about Y2K-related banking problems.

Table 3
P-Values of the Modified Wald Test for an Association between Each Response Variable and Each Explanatory Variable

Explanatory Variable	Q3	Q4	Q5A	Q5B	Q5C	Q5D	Q5E	Q7	Q13
Gender	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Age	<0.001	0.676	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Race	<0.001	0.004	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.176
Education	0.509	0.594	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.300
Income	0.039	0.757	0.033	<0.001	0.001	0.004	0.001	<0.001	0.017
Region	0.639	0.091	0.942	0.253	0.146	0.978	0.758	0.313	0.348
Urbanization	0.116	0.106	0.231	0.365	0.511	0.549	0.819	0.305	0.218
Period	0.244	0.777	0.102	0.319	0.178	0.310	0.055	<0.001	0.633

Table 4
Estimated Cumulative Odds Ratios for Each Response Variable, Comparing the Cumulative Odds of Levels of Five Explanatory Variables to the Cumulative Odds of Reference Levels

Level of Explanatory Variable	Q3	Q4	Q5A	Q5B	Q5C	Q5D	Q5E	Q7	Q13
Gender: Female	1.573*	1.322*	1.794*	1.486*	1.368*	1.655*	1.523*	1.390*	1.391*
Age: 18-25 Years	1.824*	1.212	1.651*	2.835*	2.271*	1.882*	1.969*	3.591*	2.684*
Age: 26-34 Years	1.716*	1.115	1.912*	3.065*	2.447*	1.889*	2.115*	3.339*	2.307*
Age: 35-54 Years	1.413*	1.027	1.203	2.148*	1.643*	1.438*	1.510*	2.384*	1.905*
Age: 55-64 Years	1.367*	1.177	0.955	1.675*	1.034	1.084	1.136	1.344	1.369*
Race: African-American	2.157*	1.523*	1.741*	1.860*	1.944*	2.006*	2.032*	2.092*	1.307*
Race: Hispanic	1.456*	1.853*	1.463*	1.554*	1.121	1.529*	1.416	1.884*	1.238
Race: Other	1.471*	1.493	1.483*	1.212	1.529*	2.022*	1.361	2.443*	1.125
Education: Less Than High School	1.304	1.289	1.842*	1.889*	2.065*	1.604*	2.078*	2.160*	1.142
Education: High School Graduate	1.062	1.018	1.647*	1.461*	1.454*	1.534*	1.574*	1.501*	1.093
Education: Some College	0.968	0.941	1.126	0.984	1.080	1.086	1.024	0.929	0.923
Income: Less Than $25,000	1.108	1.117	1.209	1.897*	1.956*	1.611*	1.786*	1.898*	0.878
Income: $25,000-$49,000	1.251	0.980	1.336*	1.756*	1.637*	1.488*	1.446*	1.407*	0.995
Income: $50,000-$74,000	1.131	1.091	1.314*	1.674*	1.634*	1.285*	1.348*	1.375*	1.158
Income: $75,000-$99,000	0.889	1.151	0.942	1.271	1.408*	1.064	1.086	1.086	0.952

* indicates that a 95 percent confidence interval for the cumulative odds ratio does not include 1.

Small sample sizes for some levels of the explanatory variables precluded testing for most two-factor interactions. Collapsing would be required to do meaningful testing for most interaction effects.

Nonetheless, models with two-factor interaction terms involving period were fit, and it was found that region interacted with period for response variables Q5A, Q5D, Q5E, and Q13 at a 0.05 level of significance. Although differences among regions regarding the level of concern about Y2K-related banking problems were not great, the Northeast region went from being one of the regions with the most concern in the first two periods to being one of the regions with the least concern in the third period.

An interaction effect for income and period was also significant at a 0.05 level for Q5E. This appeared to be the result of a much greater difference in the level of concern between younger age groups and older age groups in the first period than in subsequent periods regarding whether checks would be processed properly.

Suggestions for Further Analysis

The survey data were subjected to very thorough analyses on an ongoing basis under strict time constraints prior to the turn of the century. However, more analytical work could be done to further understand the attitudes of the various demographic groups regarding possible Y2K banking problems before the century date change.

To reduce item nonresponse bias (particularly regarding household income), some type of imputation could be done. Furthermore, to account for the component of variance due to imputation, multiple imputation, such as that described by Rubin (1987), could be used. One possible approach for implementing multiple imputation would involve deleting some responses to create a monotone pattern of missingness prior to imputing.

Additionally, further tests for interaction effects could be performed after collapsing levels for some of the demographic variables, such as race, age, education, and income. Since collapsing would be done on the basis of sample size limitations rather than formal testing for the appropriateness of collapsing, the results may be difficult to relate to those of the present analysis.

References

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application*, Cary, NC: SAS Institute Inc.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Wiley, New York.

Edwards, A.L., and Thurstone, L.L. (1952), "An Internal Consistency Check for Scale Values Determined by the Method of Successive Intervals," *Psychometrika*, 17, 169-180.

Evans, M., Hastings, N., and Peacock, B. (1993). *Statistical Distributions*, Wiley, New York.

Fahrmeir, L., and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models,* Wiley, New York.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*, the MIT Press, Cambridge, MA.

Hauck, W.W., and Donner, A. (1977), "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association*, 72, 851-853.

Hosmer, D.W., and Lemeshow, S. (2000). *Applied Logistic Regression,* Wiley, New York.

Jennings, D.E. (1986), "Judging Inference Adequacy in Logistic Regression," *Journal of the American Statistical Association*, 81, 471-476.

Johnson, V.E., and Albert, J.H. (1999). *Ordinal Data Modeling,* Wiley, New York.

Lehtonen R., and Pahkinen E.J. (1994). *Practical Methods for Design and Analysis of Complex Surveys*, Wiley, New York.

McCullagh, P. (1980). "Regression Model for Ordinal Data (with discussion)," *Journal of the Royal Statistical Society*, B 42, 109-127.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1997). *SUDAAN User's Manual, Release 7.5,* Research Triangle Park, NC: Research Triangle Institute.