Modeling Individual Utility Functions Using SAS PROC LOGISTIC: A Case Study Using the NES Data Set

Godofredo Gatus, American Express Financial Advisor,
West Los Angeles, California

Abstract

The purpose of this paper is to illustrate how the SAS System may be used to develop a model which can estimate and predict individual utility functions. Numerous SAS multivariate procedures are used such as PROC REG, PROC AUTOREG, PROC CORR and PROC LOGISTIC. The unique data set of interest to be used is the American National Election Survey (NES) of 1996.

Introduction

This paper presents results using the SAS PROC LOGISTIC procedure for determining several attitudinal utility preferences. For the sake of comparison, PROC REG is also used with the same variables. These results show how the SAS System may be used to analyze a large-scale questionnaire. While numerous models were estimated, three are presented. The three selected for inclusion have as their dependent variables: the respondents view on U.S. isolationism, whether or not the respondent believes that the economy will be better off next year, and the respondents' view on social security spending. Independent variables used in each model are: age, highest grade completed, gender, apparent intelligence and income level of respondent.

The Questionnaire

The 1996 American National Election Study is the twenty-fourth in a series of studies of American national elections conducted by the Center for Political Studies and the Survey Research Center at the University of Michigan. The contents of the 1996 Election study are both a traditional presidential election year time-series data collection and a panel study. The number of cases in the file numbers 1714, with complete codebook documentation for 1657 variables. The file is available in both SAS and SPSS format for researchers. In addition to questions of a political nature, there are numerous substantive themes such as: preferences on personal and national economic well-being; positions on social welfare issues including abortion, women's roles, prayer in schools, the death penalty; social altruism and social connectedness and opinions about the nation's most important problems.

Statistical Methodology

The primary statistical tool for analyzing the data is logistic regression analysis. Regression analysis is the study of the relationship between a dependent variable (Y), and one or more independent variables (X's). A linear regression equation may be expressed as:

$$Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .. + \beta_k X_k + \mu_i$$

where

Y_i is the dependent variable

X_i is the independent variables

β_i is the coefficient of variation

μ_i is the error term or residual

Regression analysis allows the researcher to determine the

influence which the independent variables have on the dependent variable. In addition, a correctly specified regression model allows for numerous statistical tests and summary statistics, such as R-square, which indicates how well the model fits the data. For most of these tests to be mathematically correct, there are a number of assumptions that are imposed on the model (Please see Gujarati for a complete list of these assumptions). For applications of the regression model where the dependent variable is a multi-value quantitative variable, the assumptions should be tested. Pending results from these tests, there are a variety of methods to deal with assumption violations. When the dependent variable is dichotomous, however, as in the case of many of the responses in this NES questionnaire, there are widely-known statistical problems when the linear probability model is applied. Primarily, the error term μ_i is no longer normally distributed, rather it is binomially distributed. Also, the error term is heteroscedastic - which is a violation of the assumptions of the ordinary least squares (OLS) model. Although regression coefficients, β_i, may still be linear and unbiased, they will not be efficient in that their standard errors will not be minimum. The result of using OLS regression in the presence of heteroscedasticity is that the conclusions derived from the analysis may be suspect from a statistical point of view. An additional problem resulting from using OLS regression in the current context of this questionnaire study is that there is no guarantee that the predicted values of the Y response variable will lie within the interval (0,1), which is where we would like it to be in order to be able to interpret the predicted value of Y as a probability. Thus an alternative approach to analyzing the questionnaire is called for.

Logistic Regression

The use of logistic regression in the famous Framingham heart study paper by Truett, Cornfield and Kannel(1967) made logistic regression the standard method for analyzing dichotomous data.[1] The logistic or logit model for k explanatory variables may be expressed as:

$$\text{Log}\left[\frac{P_i}{1-p_i}\right] = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_k X_k + \mu_i$$

where p_i is the probability that $y_i = 1$

log is the natural logarithm

The expression on the left side of the equation is referred to as the logit or log- odds. If p_i is taken as the probability of an event happening, then P_i / (1 - p_i) is said to be the odds of the event happening.

The above logit equation may be solved for p_i , resulting in the following equation:

$$P_i = 1 / (1 + \exp(-\beta_o - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - .. - \beta_k X_k)$$

No matter what values we use for the β_s and X_s, p_i will always lie between

[1] Hosmer, David W. and Lemeshow, Stanley, Applied Logistic Regression, New York: John Wiley & Sons, 1989., p. vii.

0 and 1. Thus we may interpret the Y or response variable as a probability.[2]

Empirical Results

The following empirical results were derived for the three models selected. Both the logistic regression and for comparison purposes OLS regression are presented.

The LOGISTIC Procedure
Response Profile

Ordered Value	V960410	Total Frequency
1	1. Agree	346
2	5. Disagree	1016

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1545.768	1460.413
SC	1550.984	1491.713
-2 Log L	1543.768	1448.413

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	95.3550	5	<.0001
Score	93.4693	5	<.0001
Wald	85.8342	5	<.0001

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept1	1	0.2182	0.6587	0.1098	0.7290		
V960605	1	0.000948	0.00382	0.0615	0.7979	0.00914	1.001
V960607	1	-0.1249	0.0318	15.4005	<.0001	-0.1802	0.883
V960066	1	0.00130	0.1418	0.0001	0.9932	0.000330	1.001
V960071	1	0.2746	0.1048	6.9695	0.0083	0.1177	1.316
V960702	1	-0.0366	0.0117	9.7589	0.0018	-0.1344	0.964

Analysis of Maximum Likelihood Estimates

Variable	Label
Intercept	Intercept
V960605	96PR:R'S AGE
V960607	96PR:HIGHEST GRADE COMPLETED
V960066	96PR:R'S GENDER
V960071	96PR:R'S APPARENT INTELLIGENCE
V960702	96PR:(IN FAMILY) R SELF INCOME 1995

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.6	Somers' D	0.337
Percent Discordant	32.9	Gamma	0.339
Percent Tied	0.5	Tau-a	0.128
Pairs	351536	c	0.669

The REG Procedure
Model: MODEL1
Dependent Variable: V960410 96PR:U.S. ISOLATIONISM PRO/CON
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	251.69218	50.33844	19.98	<.0001
Error	1356	3877.24746	2.83666		
Corrected Total	1361	4129.54464			

Root MSE	1.68418	R-Square	0.0686
Dependent Mean	3.98385	Adj R-Sq	0.0652
Coeff Var	42.27521		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.69532	0.47101	6.15	<.0001
V960605	96PR:R's AGE	1	-0.00119	0.00277	-0.43	0.6671
V960607	96PR:HIGHEST GRADE COMPLETED	1	0.09161	0.02296	4.16	<.0001
V960066	96PR:R'S GENDER	1	-0.00879	0.09962	-0.09	0.9297
V960071	96PR:R'S APPARENT INTELLIGENCE	1	-0.17988	0.07268	-2.48	0.0134
V960702	96PR:(IN FAMILY) R SELF INCOME 1995	1	0.02556	0.00836	3.06	0.0023

The LOGISTIC Procedure
Response Profile

Ordered Value	V960387	Total Frequency
1	1. Get better	344
2	3. Get worse	174
3	5. Stay about the same	846

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
4.5041	5	0.4793

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2476.515	2482.218
SC	2486.951	2518.745
-2 Log L	2472.515	2468.218

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2969	5	0.5075
Score	4.3114	5	0.5055
Wald	4.2683	5	0.5115

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept1	1	-0.5547	0.5604	0.9795	0.3223		
Intercept2	1	0.0433	0.5601	0.0060	0.9384		
V960605	1	-0.00388	0.00334	1.3504	0.2452	-0.0364	0.996
V960607	1	-0.0258	0.0272	1.8993	0.1924	-0.0512	0.965
V960066	1	0.0884	0.1187	0.5545	0.4565	0.0242	1.092
V960071	1	0.0242	0.0867	0.0783	0.7797	0.0104	1.025
V960702	1	0.00232	0.00999	0.0492	0.8245	0.00810	1.002

Analysis of Maximum Likelihood Estimates

Variable	Label
Intercept1	Intercept Parameter 1
Intercept2	Intercept Parameter 2
V960605	96PR:R's AGE
V960607	96PR:HIGHEST GRADE COMPLETED
V960066	96PR:R'S GENDER
V960071	96PR:R'S APPARENT INTELLIGENCE
V960702	96PR:(IN FAMILY) R SELF INCOME 1995

Association of Predicted Probabilities and Observed Responses

Percent Concordant	51.8	Somers' D	0.062
Percent Discordant	45.6	Gamma	0.064
Percent Tied	2.6	Tau-a	0.033
Pairs	498084	c	0.531

The REG Procedure
Model: MODEL1
Dependent Variable: V960387 96PR:U.S. ETR/WRS OFF NEXT YEAR
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	11.56137	2.31225	0.78	0.5618
Error	1358	4008.42406	2.95245		
Corrected Total	1363	4020.98834			

Root MSE	1.71827	R-Square	0.0029
Dependent Mean	3.73607	Adj R-Sq	-0.0008
Coeff Var	45.99134		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.34819	0.47677	4.93	<.0001
V960605	96PR:R's AGE	1	0.00259	0.00274	1.05	0.3703
V960607	96PR:HIGHEST GRADE COMPLETED	1	-0.02931	0.02339	-1.25	0.2107
V960066	96PR:R'S GENDER	1	-0.07691	0.10154	-0.76	0.4495
V960071	96PR:R'S APPARENT INTELLIGENCE	1	-0.03123	0.07402	-0.39	0.7221
V960702	96PR:(IN FAMILY) R SELF INCOME 1995	1	-0.00032403	0.00853	-0.04	0.9715

The LOGISTIC Procedure
Response Profile

Ordered Value	V960560	Total Frequency
1	1. Increased	632
2	2. Decreased	90
3	3. Kept about the same	648
4	7. Cut out entirely [VOL]	3

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
116.6972	10	<.0001

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2487.057	2362.294
SC	2502.731	2404.092
-2 Log L	2481.057	2346.294

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	134.7630	5	<.0001
Score	130.8850	5	<.0001
Wald	119.3886	5	<.0001

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Odds Ratio
Intercept1	1	2.2419	0.5698	15.4787	<.0001		
Intercept2	1	2.5299	0.5707	19.6503	<.0001		
Intercept3	1	8.7159	0.8165	113.9419	<.0001		
V960605	1	-0.0127	0.00331	12.4787	0.0004	-0.1104	0.988
V960607	1	-0.1696	0.0283	35.8232	<.0001	-0.2424	0.844
V960066	1	-0.3594	0.1187	9.1627	0.0025	-0.0984	0.698
V960071	1	0.2672	0.0871	9.4207	0.0021	0.1144	1.306
V960702	1	-0.00458	0.00993	0.1941	0.6595	-0.0160	0.996

Analysis of Maximum Likelihood Estimates

Variable	Label
Intercept1	Intercept Parameter 1
Intercept2	Intercept Parameter 2
Intercept3	Intercept Parameter 3
V960605	96PR:R's AGE
V960607	96PR:HIGHEST GRADE COMPLETED

---

[2] See Gujarati, Damodar (1995). Basic Econometrics, Third Edition, New York: McGraw-Hill, Inc. chapter 16 for a discussion of the logistic model.

V960066 96PR:R'S GENDER
V960071 96PR:R'S APPARENT INTELLIGENCE
V960702 96PR:(IN FAMILY) R SELF INCOME 1995

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.4	Somers' D	0.331
Percent Discordant	33.3	Gamma	0.333
Percent Tied	0.4	Tau-a	0.186
Pairs	528846	c	0.666

The REG Procedure
Model: MODEL1
Dependent Variable: V960560 96PR:SOC SEC SPEND INCR/DECR

Analysis of Variance

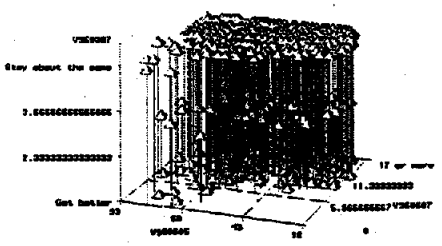
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	129.90036	25.98007	29.01	<.0001
Error	1367	1224.39973	0.89568		
Corrected Total	1372	1354.30007			

Root MSE	0.94641	R-Square	0.0959
Dependent Mean	2.02258	Adj R-Sq	0.0926
Coeff Var	46.79204		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.99708	0.26516	3.76	0.0002
V960066	96PR:R'S SEX	1	0.09584	0.08153	1.18	0.2400
V960067	96PR:HIGHEST GRADE COMPLETED	1	0.07594	0.01597	4.75	<.0001
V960071	96PR:R'S GENDER	1	-0.16568	0.05549	-2.99	0.0029
V960560	96PR:R'S APPARENT INTELLIGENCE	1	-0.12998	0.04878	-2.66	0.0078
V960702	96PR:(IN FAMILY) R SELF INCOME 1995	1	0.00743	0.00976	0.76	0.4470

The following diagram illustrates the use of the SAS procedure PROC G3D. It shows the three-dimemsional relationship between age, level of education, and the respondents' expectation regarding the future state of the economy.



Conclusion

The SAS System was used to develop utility models for three preferences of the respondents in the National Election Survey of 1996. Respondents' views on U.S. isolationism, the future state of the economy and social security spending were modeled using PROC LOGISTIC and PROC REG. For the isolation question, education, income and apparent intelligence were important factors in revealing the respondents' preference. For the future state of the economy, no particular variables were statistically significant at the traditional levels of significance. Age, race, gender and intelligence were shown to be important determinants of the respondents' preference regarding social security spending. The SAS System was shown to be a very efficient tool for analyzing this important data set.

References

Addison, Paul D., Logistic Regression Using the SAS® System: Theory and Application Cary, N.C.: SAS Institute Inc. 1999.

Gujarati, Damodar (1995). Basic Econometrics, Third Edition, New York: McGraw-Hill, Inc.

Hatcher, Larry. A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling, Cary, N.C. The SAS Institute Inc. 1994.

Hosmer, David W. and Lemeshow, Stanley, Applied Logistic Regression, New York: John Wiley & Sons, 1989.

NES website: www.umich.edu/~nes

SAS Institute Inc.(1995), Logistic Regression Examples Using the SAS® System, Version 6, First Edition. Cary, N.C.: SAS Institute, Inc.

SAS Institute Inc., SAS/ETS User's Guide, Version 6, Second Edition, Cary, NC: SAS Institute Inc., 1988.

SAS Institute Inc., SAS/STAT User's Guide, Vols I and II, 1990 Version 6, Fourth Edition. Cary, N.C.: SAS Institute Inc.

SAS Institute Inc., SAS/ETS Software: Applications Guides 1 and 2, Version 6, First Edition, 1993.

SAS Institute Inc., SAS/STAT User's Guide,
Vols I and II, 1990 Version 6, Fourth Edition. Cary, N.C.: SAS Institute Inc.