Estimating and Interpreting Ordered Regression Models Using PROC PROBIT

Kenneth E. Fernandez
University of California, Riverside

ABSTRACT

Most survey research regarding public opinion (from product satisfaction to presidential approval) provide ordinal response data. There are many ways to analyze ordinal outcomes using SAS but one of the most useful methods is PROC PROBIT. Although this procedure is comprehensive, the coefficients cannot be interpreted directly because they reflect the arbitrary assumptions used in identifying the model. Instead, predicted probabilities and partial and discrete change in probabilities should be used to interpret the relationship between independent dependent variables. Unfortunately calculating these probabilities using PROC PROBIT is not an intuitive process and the estimated intercept and thresholds are often different from other statistical packages because of different identifying assumptions.

This paper will explain the necessary steps to estimate an ordered regression model and how to produce and interpret predicted probabilities. These methods are illustrated with survey data from a two county area in California (the Inland Empire).

INTRODUCTION

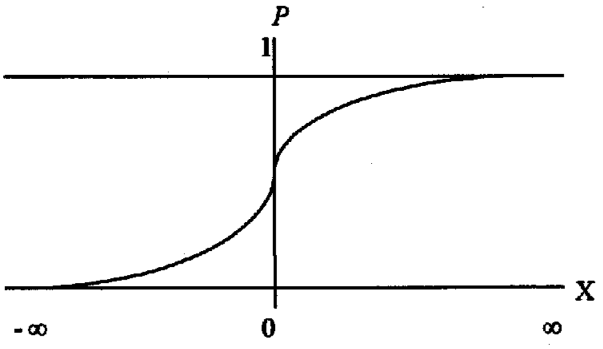
One of the most powerful and useful analytic statistical tools available to researchers is the regression analysis. It allows us to examine multivariate relationships and create models for forecasting and prediction. Ideally it would be nice if we could use this tool to analyze all types of data and often we try to. Like the old saying goes, if you are a hammer everything looks like a nail. But different nails need different hammers. The Linear Regression Model (LRM) commonly used by researches does not provide us with the analytic power we desire when the data does not fit the statistical assumptions used to estimate the parameters of the model.

Often the data that we have on the questions we are interested in do not appear in continuous, normally distributed form. Many researchers (commercial and academic) are interested in public opinion (from product satisfaction to presidential approval) and such data often come from questions with ordinal response categories. These categories can be ranked from low to high but the distances between each category is unknown. There are many ways to analyze ordinal outcomes using SAS; one of these is PROC PROBIT. This paper explores this procedure and presents a method to better understand and interpret the results of PROC PROBIT.

MODEL SPECIFICATION

When the Linear Regression Model is applied to ordinal data (also know as Linear Probability Model, LPM) several statistical assumptions are violated and problems occur. These include Heteroscedasticity, normally distributed errors, and nonsensical predictions. These problems bias and reduce the efficiency of the estimates (Gujarati 1995; Long 1997). Beyond the statistical problems met with the use of the LPM on ordinal data, there are also some theoretical problems. The LPM assumes that the effect of x on y remains constant throughout the range of x. This is unrealistic when modeling how a variables influences the probability of an event. For example, if we are interested in the probability of someone purchasing a new car, the income level variations (x1) at the mid-range will greatly influence this probability, but changes in income at the low levels of income will only marginally increase the probability of purchasing a car. If you are very poor an increase in income by $1000 still won't increase your probability of purchasing a car by very much – the same phenomenon occurs at the high income levels, where a decrease in a millionaire's income by $1000 will not greatly deter him or her from purchasing a new car.

What is needed is a model that allows the probability of y to change slower as x_i gets very small and also as x_i gets very large. This relationship is well modeled by the normal cumulative distribution function (CDF) of a random variables.



The sigmoid shaped distribution of the normal CDF indicates the probability a random variables will take on a specific value (0, 1 for this binary example).

The normal CDF is not the only distribution that can be used to model Ordered Regression Models (ORM). SAS PROC PROBIT can specify the normal, logistic or Gompertz. For this paper I have chosen to focus on the normal CDF for two reasons: 1) the logistic distribution has received a large amount of attention in the statistical and SAS literature (see Knoke & Bohrnstedt 1994; Stokes, Davis, & Koch, 1995) and 2) because the probit model is popular with political science due to its relationship with utility theory or rational choice theory (See Gujarati 1995; McFadden 1973).

In the probit model, the errors are assumed to have a normal distribution with a conditional mean of 0 and variance of 1. This results in the cumulative distribution function:

$$\phi(\varepsilon)=\int_{-\infty}^{\varepsilon}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)dt$$

Where t is a standardized normal variables, i.e. ~ N (0, 1).

Once the distribution of the errors is specified maximum likelihood estimation (MLE) can be used to estimate the regression coefficients. SAS computes maximum-likelihood estimates using a modified Newton-Raphson algorithm for the probit equation:

$$\Pr(y=1)=F(x'\beta)$$

Where F is a cumulative distribution function (SAS Institute, 1990). See Long 1997 (pp. 54-57) for alternative numerical methods of MLE.

INTERPRETING THE COEFFICIENTS

Assessing the effects of the independent variables on the dependent variable from the coefficients produced by the probit model is not as intuitive as the interpretation of the LRM coefficients. Stating that a unit change in x1 is associated with a reduction in the z-score or the cumulative normal function of the probability that Y = 1 may lose something in the translation for those with little statistical background . In addition, identification of the coefficients require that we make assumptions about the mean and variance of the error term (ε). This makes the interpretation of the coefficients themselves problematic since they reflect both the relationship between the independent variables and the ordinal dependent variable AND the assumptions we hold. All of this means that interpretation of the coefficients in isolation is limited to the direction (positive or negative) and statistical significance and even the direction of the coefficients can lead to some confusion (see next section).

Instead, one can use the coefficients to compute the probabilities of observing values of y given a specific value of x (perhaps the mean) while holding all other ind. variables constant and then recalculating the probability given x + 1. The difference between these two probabilities is a measure of "elasticity". This measure is easily interpreted since a unit change in x is associated with a change in the probability that a value of y will occur. These measures can be compared across independent variables in order to provide a measures of the magnitude of the effects.

COMPUTING PROBABILITIES OF OBSERVED VALUES

The formula used to calculate the probabilities of the values of dependent variable for a set of observed values of the independent variables can vary depending on the statistic package you use or the literature you read. The formula generally referenced in text books like that of Long (1997) or in the statistical package like STATA is similar to the following:

$$\Pr(y_i=\mathrm{m}\mid x_i)=F(\tau_m-x_i\beta)-F(\tau_{m-1}-x_i\beta)$$

Where F represents the Cumulative normal distribution with a mean of 0 and a standard deviation of 1 and the tau's (τ) represent the thresholds or cutpoints.

SAS's model specification is different. For a model that has a dependent variable with 4 observed outcomes (such as the one in my data) the probability of a specific outcome to be observed for a given set of x's are modeled as follows:

$$\Pr(y_i=1\mid x_i)=\Phi(\tau_1+a+\beta x_i)$$

$$\Pr(y_i=2\mid x_i)=\Phi(\tau_2+a+\beta x_i)-\Phi(\tau_1+a+\beta x_i)$$

$$\Pr(y_i=3\mid x_i)=\Phi(\tau_3+a+\beta x_i)-\Phi(\tau_2+a+\beta x_i)$$

$$\Pr(y_i=4\mid x_i)=1-\Phi(\tau_3+a+\beta x_i)$$

Where ϕ represents the Cumulative normal distribution with a mean of 0 and a standard deviation of 1. *a* represents the intercept which is assumed to be zero in order to fully identify the model.

As you can see at first glance, the signs of the betas (β) will be reversed. In addition, the estimates of τ_2 and τ_3 will be different. This makes interpretations across the literature somewhat confusing. In SAS a negative coefficient indicates that the probability that the lesser response value occurs decreases in STATA it indicates an increased likelihood that the lesser value will occur.

Even if this confusion was not a problem the coefficient estimates themselves do not tell the researcher much about the probabilities. In order to truly get a feel of how the independent variables influence the probability that a response value will occur the probabilities need to be calculated for a range of values for the independent variables.

The OUTPUT statement in PROC PROBIT can be specified to compute the cumulative probability estimates and export this information into a SAS dataset, but only for the observations that occur in the data set. This limitation in addition to the less then useful structure of this output prevents this information from being easily readable, useable, or interpretable. In addition, probabilities for the highest response level are left out. These can be calculated by hand by simple addition and subtraction or a simple SAS program could be written, but this can be tedious. This doesn't mean that this output is useless. This output can be separated by _Level_ (response category) and then sorted and plotted to examine the range of probabilities and the corresponding values of x_i within the data. The limitation is that the specific effect of a particular explanatory variable cannot be examined because the other independent variable values cannot be held constant.

Instead, it is very useful if the probability of the values of y can be seen as x_i changes by some discrete unit (lets say 1) while the other independent variables are held constant. Then the change in the probabilities can be compared within and across the independent variables.

The information you need to compute the probabilities are the parameter estimates for the independent variables, the cutoff points (tao$_1$-tao$_j$), and the minimum, mean, maximum, standard deviation depending on where you want to set the baseline probability score and what value you want to change x_i. In the example in the next section I simply set the base at the average score (or zero for dichotomous variables) for the independent variables and change each variable by 1 while holding the other variables constant.

The SAS program below will calculate probability estimates of y for a given set of x_i values. It should be noted that because of the structure of the data a CLASS statement must be used in the PROC PROBIT program. This means that the OUTEST option cannot be used to save the estimates to a SAS dataset. Instead, I use an ODS statement to do this.

SAS PROGRAM

```
/* the ods output places the     */
/* parameter estimates in        */
/* temporary SAS dataset ests    */

PROC PROBIT;
ods output parameterestimates=ests;
class y1;
model y1 = x1 x2 x3 x4/ d=normal;


/* delete unnecessary variables  */
/* and transpose the SAS dataset */

DATA ests2;
SET ests;
keep variable estimate;


PROC TRANSPOSE data = ests2 out =
transpos;
id Variable;


/* code creates a dataset with   */
/* the descriptive statistics of */
/* the ind. variables            */


%LET varlist=x1 x2 x3 x4;
%LET stdlist=stdx1 stdx2 stdx3 stdx4;
%LET minlist= minx1 minx2 minx3 minx4;
%LET maxlist= maxx1 maxx2 maxx3 maxx4;
%LET mulist= mux1 mux2 mux3 mux4;


PROC MEANS data=probit noprint;
var &varlist;
output out=stats std=&stdlist
min=&minlist
max=&maxlist mean=&mulist;


/* merge the 2 datasets for      */
/* probability calculation       */


DATA merged;
merge stats transpos;
run;

/* this macro program extract    */
/* necessary parameters and      */
/* computes the probablities     */
/* for various values of x1-x4   */
```

```
%MACRO calcprobs (file, n, var,
deltax);
data &file;
set merged;
n=&n;
b1=&x1;
b2=&x2;
b3=&x3;
b4=&x4;
x1=mu&x1;
x2=mu&x2;
x3=mu&x3;
x4=mu&x4;
if x1 < 1 then x1 = 0;
if x2 < 1 then x2 = 0;
if x3 < 1 then x3 = 0;
if x4 < 1 then x4 = 0;
i1=intercept;
i2=inter_2;
i3=inter_3;

&var = &var + &deltax ;
xb=(x1*b1)+(x2*b2)+(x3*b3)+(x4*b4)+i1;

p1=cdf('Normal',xb,0,1);
p2=cdf('Normal',i2+xb,0,1)-p1;
p3=cdf('Normal',i3+xb,0,1)-
cdf('Normal',xb+i2,0,1);
p4=1-cdf('Normal',i3+xb,0,1);
keep n x1-x4 p1-p4 ;
output;
run;

%MEND calcprobs;

%calcprobs (filex1a, 1, x1, 0)
%calcprobs (filex1b, 2, x1, 1)
%calcprobs (filex2a, 3, x2, 0)
%calcprobs (filex2b, 4, x2, 1)
%calcprobs (filex3a, 5, x3, 0)
%calcprobs (filex3b, 6, x3, 1)
%calcprobs (filex4a, 7, x4, 0)
%calcprobs (filex4b, 8, x4, 1);


/* merge the 2 datasets & export */
/* to MS Excel for placement in  */
/* report                        */


DATA merged2;
merge filex1a filex1b filex2a filex2b
filex3a filex3b filex4a filex4b;
by n;
run;

PROC PRINT data=merged2;

PROC EXPORT DATA= WORK.merged2
            OUTFILE=
"C:\oprobit\probs.xls"
            DBMS=EXCEL2000 REPLACE;

run;
```

*A copy of this code can be found at http://pages.zdnet.com/fernandezk/WUSS2001code.html

EXAMPLE: FEAR OF CRIME

For an illustrative example I will examine a survey question that asked approximately 1000 residents of Riverside and San Bernardino County how fearful they were of becoming a victim of a serious crime.

How Fearful are you of being a victim of a serious crime?	n
1= Very Fearful	55
2= Somewhat Fearful	253
3= Not too Fearful	409
4= Not at all Fearful	200

Four explanatory variables were chosen for this example. They include ECONSEC which is an additive scale (10 points) that assess a respondent's perception of his or her economic well-being. Respondents that score high on this measure on more economically insecure. GENDER, which simply identifies the sex of the respondent (1=male). HISP90, which measures the size of the Hispanic community within the respondent's zip code (in %). And Ideology, which places the respondent on a 7 point liberal/conservative scale. The data for this paper are based on a survey conducted by the Center for Social and Behavioral Sciences Research at the University of California, Riverside. The survey was implemented between December, 1999 and January, 2000.

Analysis of Parameter Estimates

Var	Estimate	SE	Chi-Sq.	Pr >	Label
Interc	-2.39201	0.2148	123.972	<.000	Intercept
x1	0.06163	0.0214	8.2937	0.004	econsec
x2	-0.17127	0.0732	5.4684	0.019	gender
x3	0.0079375	0.0028	7.5167	0.006	hisp90
x4	0.07607	0.0325	5.4622	0.019	ideology
Inter.2	1.15604	0.0659			2
Inter.3	2.37875	0.0769			3

Positive coefficients can be interpreted as increasing the likelihood of observing the lowest value (1) which in this case represents "Very Fearful". Beyond this direction and the statistical significance not much else can be extrapolated.

Using the program listed in the previous section, the probabilities of an observed value of y given a set of values of x1-x4 can be calculated, printed, and exported into a Microsoft Excel spreadsheet. The resulting dataset shown below is created with this program. The odd number rows shows the predicted probabilities of y given that all the independent variables are at their mean. Row number 2 shows the probabilities of y given a change in x1 by 1 unit, while row 4 shows the probabilities of y when x2 changes by 1 unit

x1	x2	x3	x4	n	p1	p2	p3	p4
7.204	0	25.56	3.165	1	0.066	0.298	0.445	0.191
8.204	0	25.56	3.165	2	0.075	0.313	0.438	0.175
7.204	0	25.56	3.165	3	0.066	0.298	0.445	0.191
7.204	1	25.56	3.165	4	0.047	0.255	0.457	0.241
7.204	0	25.56	3.165	5	0.066	0.298	0.445	0.191
7.204	0	26.56	3.165	6	0.067	0.300	0.444	0.189
7.204	0	25.56	3.165	7	0.066	0.298	0.445	0.191
7.204	0	25.56	4.165	8	0.077	0.316	0.436	0.171

With this information the difference in the calculated probabilities can easily be computed. In a causal relationship, this elasticity measures the change in the probability of y = i with respect to a unit j change in x holding all other variables constant.

Elasticity Y = 4		x
x1	ECONSEC	0.016322
x2	GENDER	-0.05004
x3	HISP90	0.002153
x4	IDEOLOGY	0.020014

The elasticity measure is useful because it provides the researcher with a method to estimate the magnitude of the effects rather than just simply noting the directions and statistical significance. Since the statistical significance is a function of the sample size its use as a measure of magnitude of effect can be misleading. The elasticity measures allows the researcher to examine the "practical" significance an independent variable has on the dependent variable. Note that the range of x1-x4 should also be examined in conjunction with the elasticity score in order to better comprehend the magnitude of the effects. Although GENDER has a larger elasticity score, ECONSEC has a 10 point range.

SUMMARY

Because many of our problems include ordinal data Ordered Regression Models are an important tool for researchers. This paper only scratched the surface of the many ways interpret these models. Calculating the elasticity is only one of many ways to manipulate and interpret the predicted probabilities of values of y in a manner easily understood by a broad audience. With some slight manipulation of the SAS program listed above, probabilities can be produces so as to enable the researcher to create plots of predicted and cumulative probabilities, tables of predicted probabilities, and explore partial and discrete change in probabilities of y depending on the level of detail needed (see Long 1997, pp. 130-37).

ACKNOWLEDGEMENTS

I would like to thank Professor Max Neiman of the Department of Political Science, University of California, Riverside for the use of the Inland Empire Opinion Poll data.

REFERENCES

Gujarati, D. N. 1995. *Basic Econometrics*, 3^{rd} edition. New York: McGraw-Hill.

Knoke, D., and G. W. Bohrnstedt. 1994. *Statistics for Social Data Analysis*. Itasca, IL: F.E. Peacock.

Long, S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

McFadden, D. 1973 "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics*. New York: Academic Press.

SAS Institute Inc. 1990. *SAS/STAT User's Guide, Version 6, 4^{th} edition*. Cary, NC: SAS Institute Inc.

Stokes, M. E., C. S. Davis, and G. G. Koch. 1995. *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Kenneth E. Fernandez, Graduate Student
Department of Political Science
University of California
Riverside, CA 92521
(909) 781-9507
politicalscience@email.com