

Presenting Binary Logit/Probit Models Using the SAS/IML

Hun Myoung Park, Indiana University, Bloomington, Indiana

ABSTRACT

The binary logit/probit regression models can be more difficult to interpret the results than linear regression models. The difficulty is due largely to their nonlinear nature. Just reporting parameter estimators is not helpful for understanding the impacts of independent variables. Like STATA and SPSS, however, the SAS procedures such as LOGISTIC, PROBIT, and GENMOD focus on producing parameter estimators and goodness-of-fit measures. This paper illustrates how to effectively present binary logit/probit models using the several methods extensively discussed by J. Scott Long (1997). A SAS/IML program, similar to Long's SPost module, was written to conduct the necessary follow-up analyses of binary logit/probit models.

1. INTRODUCTION

The binary response (choice) regression model (BRM) has been one of the most commonly used data analysis techniques in social sciences. It is no longer surprising to encounter many articles in journals that apply the logit/probit models to various cases. Despite this popularity, it is not easy to find literature that presents the result in an easy to understand manner. It is the nonlinearity of the BRMs that underlies the difficulty in interpretation.

SAS has the LOGISTIC, PROBIT, GENMOD, and CATMOD procedures to estimate binary logit/probit models. Like STATA and SPSS, however, all these procedures focus on estimating parameters and goodness of fit. Parameter estimators in BRMs do not tell much about the impacts of independent variables. Thus, effective interpretation of the binary logit/probit models calls for more than model estimation.

J. Scott Long (1997) extensively discussed various ways of presenting BRMs such as predicted probabilities of ideal types, discrete changes, marginal changes, and factor changes in odds. He also wrote a module for STATA, called SPost, which is devoted to conducting these follow-up analyses in an easy manner (Long and Freese 2001).¹

The purpose of this paper is to develop a SAS/IML program similar to the SPost, and to illustrate effective ways of presenting binary logit/probit models using the program. This paper consists of four parts. The next section compares BRMs with ordinary least squares (OLS). The third section explains the problems associated with interpreting BRMs. The fourth section discusses interpretation methods of BRMs and demonstrates how effectively these methods can interpret the BRMs. Then, I will briefly explain software issues and the SAS/IML program for the binary logit/probit models. Finally, I conclude with suggestions that two statements be added to the LOGISTIC procedure.

2. NONLINEARITY OF THE BINARY RESPONSE MODEL

The linear probability model (LPM) refers to an OLS that is applied to binary response models (BRMs). Since the dependent variable of BRMs by definition has only two choices (0 or 1, or yes or no), the LPM violates OLS assumptions (Maddala 1983; Long 1997; Green 2000). First, a linear functional form does not fit the data. Instead, a S-shaped curve like the cumulative standard logistic distribution function may be more relevant. Second, the variance of errors is not constant for a given independent variable. This is heteroscedasticity. These violations of LPM result in with biased and inefficient estimators. Statistical inferences based on these estimators are neither reliable, nor valid. Third, the linear regression lines predict unreasonable values greater than 1 or smaller than 0. Consequently, the LPM is just inappropriate for a dichotomous dependent variable case. Table 1 summarizes the differences between the LPM and BRMs (binary logit/probit models).

Table 1. Comparison between LPM and BRMs

Functional Form	Dependent Variable Independent Variable Linearity	LPM (OLS)	BRMs (logit/probit)
		Interval Linear combination Linear	Binary (logit) Linear combination Nonlinear
Estimation	Prob. Distribution		Logistic or normal distribution
	Estimation	Moment-based/MLE	MLE
	Sample size		Large N preferred
Estimators	Goodness-of-fit	F and R ²	LR (χ^2), Pseudo R ²
	Coefficient	Slope	Change in the logit
	Discrete change	Constant (slope)	It varies
	Marginal change	Constant (slope)	It varies
	Factor change in odds	N/A	Constant

Like LPM, the BRMs also have a linear combination of independent variables in the right-hand side. But, the dependent

¹ <http://www.indiana.edu/~jslsoc/spost.htm>

variable in BRMs is not a typical interval variable Y . In the binary logit model (BLM), logit, log of odds ratio $\ln\Omega$, is the dependent variable (see the first equation below). Equivalently, the probability of an event ($Y=1$) is formulated in the cumulative standard logistic probability density function (see the second equation below). As a result, BRMs are called as nonlinear probability models as opposed to LPM of $Y=X\beta$.

$$\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \ln\Omega = X\beta$$

$$P(Y=1|X) = \Lambda(X\beta) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

Note that the Λ of the logit model denotes the cumulative standard logistic probability distribution function. The probit model uses a cumulative normal probability distribution function referred to as Φ . Despite different probability distribution functions used, the logit and probit models produce almost the same results in reality (Long 1997). The logit model produces roughly 1.8 times (standard deviation of the standard logistic function) larger coefficients than the probit model. But z and Wald (chi-squared) statistics are equivalent (almost the same p -values), indicating that the binary logit/probit models present the similar results in different ways. This paper focuses primarily on the binary logit model.

BRMs use the maximum likelihood estimation (MLE) method that requires an assumption about probability distributions. By contrast, LPM (OLS) often uses the moment based estimation method that does not require such an assumption. As a result, BRMs prefer a large sample size to take advantages of the asymptotic properties of MLE (King 1998; Allison 1991). LPM is not as sensitive to sample sizes as BRMs.

The nonlinearity of BRMs results in difficulty interpreting the effects of the independent variables (Long 1997). More specifically, the slope of the probability curve is not constant, but depending on the values of an independent variable of interest and other variables. The coefficients just tell the direction of the impacts and the extent that the standard logistic distribution function curve is stretched. In LPM, by contrast, the coefficients are constant slopes of independent variables (slope=marginal change=discrete change). So, marginal (partial) changes and discrete changes are necessary in BRMs to interpret results effectively. In short, parameter estimators do not provide readers with any intuitive idea about how individual independent variables affect the probability of an event in which they are most interested.

3. INTERPRETATION PROBLEM

Let me compare the binary logit model with LPM to emphasize the interpretation problem in BRMs. The sample model here has four variables.² The dependent variable, `owncar`, is whether a student owns cars (1 for owner). The `budget` is his/her monthly budget (interval). The `credits` is the number of credit hours that he/she took (interval). Finally, `offcamp` is whether he/she lives off campus (1 for off campus housing). Suppose we wish to know how these three variables affect car ownership.

The REG procedure estimates LPM and returns the following results (Table 2). F statistic says that the model is not bad despite the poor R^2 . .0467. The `budget` and `offcamp` turned out statistically significant. We can interpret these coefficients in a straightforward manner. For additional \$100 increase in monthly budget, we can expect .03 (= .0003 X 100) increase in `owncar`, holding all other variables constant. Similarly, living off campus is expected to increase `owncar` by .4986, holding all other variables constant. However, what the .03 and .4986 mean? They are not the actual changes in predicted probabilities of an event, owning cars. Then, what can we say about car ownership using "these numbers"? This is a fundamental problem of LPM.

```
PROC REG; MODEL owncar = budget credits offcamp; RUN;
```

Table 2. Results of the Linear Probability Model (OLS)

Parameter	Estimator	SE	T	P-value
Intercept	.0299	.1159	.26	.7964
budget	.0003	.0001	4.41	.0001
credits	.0007	.0036	.19	.8505
offcamp	.4986	.0994	5.02	.0001

N=1,000, F=16.25 ($p<.0001$), $R^2=.0467$, Adjusted $R^2=.0438$

Now, consider the following LOGISTIC procedure for the binary logit model and its output (Table 3). We know (1) the model fits the data well by looking at Likelihood Ratio (LR) and Wald test, (2) the `budget` and `offcamp` are statistically significant, while the `credits` is not. Finally, (3) all three independent variables have positive impacts on the dependent variable `owncar`.

```
PROC LOGISTIC DESCENDING COVOUT OUTEST=cov;
  MODEL owncar = budget credits offcamp /COVB STB CTABLE PPROB=.5;
  UNITS budget=sd credits=sd;
RUN;
```

² This data set was provided for the class by David Good at Indiana University. The data were manipulated to maintain confidentiality.

However, there is no way to directly interpret the coefficients substantively. A coefficient just represents the expected change in logit, not in `owncar`, when the independent variable increases by one unit. What does the .0017 change in logit for a dollar increase in `budget` mean? How is the change in logit related to the change in the likelihood of an event? Little is said by coefficients about to what extent and how an independent variable affects the dependent variable in conjunction with other variables. Here is a rationale to develop various interpretation methods for binary logit/probit models.

Table 3. Results of the Binary Logit Model

Parameter	Estimator	SE	Wald	P-value
Intercept	-2.3505	.6577	12.7711	.0004
<code>budget</code>	.0017	.0004	18.7116	.0000
<code>credits</code>	.0037	.0189	.0383	.8439
<code>offcamp</code>	2.2495	.5687	15.6439	.0001

N=1,000, LR=44.43 (p<.0001), Wald=36.05 (p<.0001), McFadden's R^2 =.0377

4. EFFECTIVE INTERPRETATION METHODS

In observance of the difficulty in interpretation of BRMs, J. Scott Long (1997) extensively discussed several methods for effective interpretations. Long and Freese (2001) show how to conduct these follow-up analyses of BRMs using their `SPost` module.³ The major follow-up analyses include predicted probabilities of ideal types, plotting the predicted probabilities, marginal changes, discrete changes, and factor changes in odds.

4.1 PREDICTED PROBABILITIES FOR IDEAL TYPES

The interpretation using predicted probabilities is an intuitive approach. The predicted probabilities are computed as,

$$\hat{P}(Y=1 | X) = \Lambda(X\hat{\beta}) = \frac{\exp(X\hat{\beta})}{1 + \exp(X\hat{\beta})}$$

SAS LOGISTIC procedure can produce estimated (predicted) probabilities with the `OUTPUT` statement (e.g., `OUTPUT OUT=pred PREDICTED=p;`). Predicted probabilities are based on the actual values of independent variables of all observations in a data set. The LOGISTIC procedure also provides association measures (e.g., Somers' D, Gamma, and Tau-a) of predicted probabilities and observed responses. Although not theoretical, we may simply report the percentage of correct predictions by splitting predicted probabilities at .5 and assigning 1 if greater than .5 and 0 otherwise. However, this method may be a measure of goodness of fit and less useful especially when responses are extremely unbalanced.⁴

Long (1997) suggested the predicted probabilities for several ideal types, not for all possible value combinations of independent variables. This approach, often used in mass media, is quite useful in that it can focus on several interesting ideal types, and that it intuitively presents the results to econometricians as well as those without a rigorous understanding of quantitative methods. From the sample model, we can compute the predicted probabilities for ideal types, holding `credits` at its mean.

Table 4. Predicted Probabilities of Ideal Types

Ideal Types	Predicted Probabilities	Difference (Off-On campus)
On campus students with low budget level (\$500)	.1883	
Off campus students with low budget level (\$500)	.6875	.4992
On campus students with middle budget level (\$1,000)	.3507	
Off campus students with middle budget level (\$1,000)	.8366	.4859
On campus students with high budget level (\$1,500)	.5570	
Off campus students with high budget level (\$1,500)	.9226	.3656

Table 4 shows how the `budget` and `offcamp` affect the probability of owning cars. For example, 19 percent of students who live on campus with the low level of monthly budget (\$500) will own cars, whereas 56 percent of on campus students with high level of budget (\$1,500) will drive. Living off campus increases the predicted probabilities of owning cars by roughly 40-50 percent, depending on monthly budget level. This tabulation of predicted probabilities can provide audience with the big picture of the model efficiently. However, this approach does not tell the impacts of individual independent variables.

4.2 PLOTTING PREDICTED PROBABILITIES

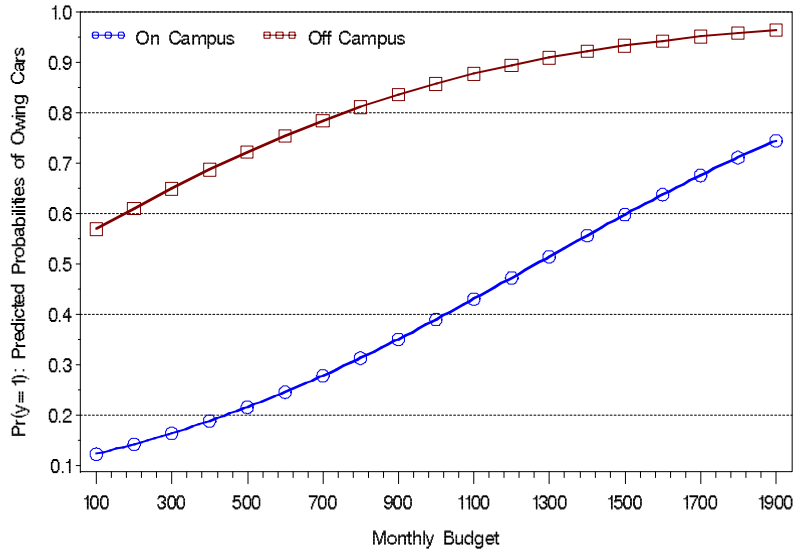
Plotting predicted probabilities over the range of an interval independent variable is also an effective presentation method. Like the tabulation of the ideal types, it is quite intuitive, but it also present the impact of individual independent variables in a more easily interpreted graphic form.

³ STATA users need to download the `SPost` from <http://www.indiana.edu/~jslsoc/stata/> using the `.net` command.

⁴ Suppose 99 percent of responses are yes. Assuming no problem in random sampling, we are able to predict results almost correctly by just saying "yes" without considering anything (theory and model) else.

The substantial gap between the two curves in Figure 1 shows that off campus students are more likely to own cars than on campus students, although the difference decreases as students' monthly budget increases. Of course, students' budget has positive effects on the probability of owning cars. This plot is consistent with the above predicted probabilities of ideal types. This presentation method effectively illustrates how individual independent variables affect the probabilities, appealing to even those who are not familiar with the binary logit model.

Figure 1. Plotting Predicted Probabilities



Source: the Binary Logit Model for Owning Cars (2004)

4.3 MARGINAL (PARTIAL) CHANGES

A marginal change (marginal effect) in the binary logit/probit models is comparable to the slope (coefficient) in OLS in the light that it is the slope of the probability curve relating an independent variable to the probability of an event, holding all other variables constant. Marginal changes are computed by taking the first partial derivative with respect to corresponding independent variables.

$$\frac{\partial \Lambda(X\beta)}{\partial x_k} = \frac{e^{X\beta}}{(1 + e^{X\beta})^2} \beta_k = \Lambda(X\beta)[1 - \Lambda(X\beta)]\beta_k$$

While marginal changes in LPM are coefficients (scalars), the marginal changes in BRMs are not constant. The marginal effects vary depending on the values of the independent variable and other independent variables. Accordingly, it is important to choose appropriate baseline values of x_k and other independent variables.⁵

Table 5. Marginal Changes and Discrete Changes

Parameter	0→1	-1/2→1/2	-sd/2→sd/2	Marginal
budget	.0001	.0003	.0604	.0003
credits	.0006	.0007	.0025	.0007
offcamp	.5091			

* Baseline: mean values except offcamp=0

The marginal change can be interpreted as follows. For one unit increase in x_k from the baseline, the probability of an event is expected to increase/decrease by the magnitude of marginal change, holding all other variables constant. For instance, for \$100 increase in students' budget, the probability of owning cars is expected to increase by 3 percent ($= .0003 \times 100$), holding all other variables constant (see Table 5). However, marginal changes are not appropriate for interpreting the impacts of dummy variables, *offcamp* in this case.

4.4 DISCRETE CHANGES

The discrete change tells how much predicted probabilities will change when we increase x_k by δ from a baseline.⁶ Unless δ approaches zero (infinitely small), the discrete change and marginal change are not the same. It is useful especially when we interpret coefficients of dummy variables and when we wish to focus on the predicted probability changes for a particular range of an independent variable. Discrete changes are computed as,

⁵ Mean values are often used. But median is more informative when variables are substantially skewed.

⁶ The δ represents a range of value changes, such as from 0 to 1, from minimum to maximum, from -1/2 from the baseline to 1/2, and from -1/2 standard deviation from the baseline to 1/2 standard deviation.

$$\frac{\Delta\Lambda(X\beta)}{\Delta x_k} = P(y=1 | X, x_k + \delta) - P(y=1 | X, x_k)$$

The discrete change can be interpreted as follows. For a change in x_k from x_k to $x_k + \delta$, the probability of an event is expected to change by the magnitude of discrete change, holding all other variables at the given values. From Table 5, a standard deviation change in budget centered around the mean will increase the probability of owing cars by 6 percent, holding other variable at the given values. For the dummy variable, if a student lives off campus, his probability of owning cars is .51 greater than a student who lives on campus, holding all other variables at the given values.

Like marginal changes, the discrete changes depend upon the baseline values of x_k , the values of other independent variables, and the magnitudes of δ . Note that the discrete changes of centered unit for dummy variables do not make any sense (see the empty cells in Table 5).

4.5 FACTOR CHANGES IN ODDS

The odds ratios are computed by taking the exponential of the logit equation.⁷ The factor change in odds of x_k is,

$$\ln \Omega(X, x_k + \delta) - \ln \Omega(X, x_k) = \frac{\Omega(X, x_k + \delta)}{\Omega(X, x_k)} = e^{\beta_k \delta}$$

Note that one or a standard deviation of the independent variable is often used as δ . In fact, it is not difficult to compute factor change in odds even by hand. The LOGISTIC procedure by default reports odds ratio estimates for one unit increase of corresponding independent variables. The UNITS statement specifies the units of change of interval variables (e.g., `UNITS budget=sd credits=10;`).

Interpretation using odds ratio is as follows (see Table 6). For a standard deviation increase in budget, the odds of owning cars are 40.7 percent (1.407 times) greater, holding all other variables constant. Living off campus increases the odds of owning cars by a factor of 9.483, holding all other variables constant.

Table 6. Odds Ratio Changes

Parameter	Estimator	e^{β_k}	$e^{\beta_k SD_k}$	SD of IV
budget	.0017	1.0017	1.4065	201.8442
credits	.0037	1.0037	1.0143	3.8283
offcamp	2.2495	9.4827		

Unlike marginal or discrete changes, factor changes in odds do not depend on the baseline value of x_k and the values of other variables; they depend on the magnitudes of change in x_k . However, a constant factor change in the odds does not correspond to a constant change or constant factor change in the probability of an event (Long 1997). Accordingly, the interpretation using odds ratio is less useful in the sense that most people are interested in the probabilities rather than the odds ratio. Table 7 recapitulates the five approaches with their pros and cons.

Table 7. Pros and Cons of the Approaches

	Pros	Cons
Predicted Probability	Intuitive	No impact of IV
Plotting P.P.	More intuitive	Not easy to draw
Marginal Changes	Useful for interval IV	Invalid for binary IV
Discrete Changes	Useful for binary and interval IV	
Factor changes in odds	Easy to compute	Less intuitive

5. A SAS/IML PROGRAM

This section discusses software issues and briefly describes the SAS/IML program for the follow-up analyses mentioned above.

5.1 SOFTWARE ISSUES

SAS by default uses the compiler mode of executing commands. SAS 8.2 and 9.1 also provide the point-and-click (menu-driven) mode and beginners can use the SAS Enterprise Guide. But writing SAS programs for the batch jobs is common for most SAS users. This mode contrasts with STATA's interactive mode. Table 8 summarizes the comparison between SAS and STATA interfaces.

Table 8. Comparison between SAS and STATA Interfaces

	SAS/Stat 8.2 and 9.1	STATA 8.2 (SE)
Noninteractive mode	Default	Yes (.do file)

⁷ Since $\ln \Omega(X) = X\beta$, $\Omega(X) = e^{X\beta}$

(Compiler/batch job)		
Interactive mode	No	Default
Point-and-Click	Yes	Yes
Follow-up Analysis	Difficult	Relatively Easy

STATA is command-driven software which uses an command interpreter rather than a compiler, although it provides the noninteractive and the point-and-click mode. STATA's command interpreter keeps analysis results (e.g., a covariance matrix and a coefficient matrix) in memory, making it possible to use them afterward without running the analysis again. Accordingly, STATA is very handy in terms of conducting follow-up analyses.

In the SAS LOGISTIC procedure, it is possible to store parameter estimators and their variance-covariance matrix into a data set using the COVOUT and OUTEST options (see the LOGISTIC procedure in section 3). But we need to perform matrix operations to conduct follow-up analyses using SAS/IML.

5.2 THE SAS/IML PROGRAM

The SAS/IML program is designed for the effective interpretation methods discussed in section 4. It adopts the Fisher scoring method, an iteratively reweighted least squares algorithm, illustrated in the manual (SAS 1990 and 2004). The program consists of five parts: model estimation, hypothesis testing and goodness of fit measures (`wald` module), predicted probabilities (`prtab` module), discrete and marginal changes (`prchange` module), and generating predicted probabilities for plotting (`prgen` module). The probit model can be estimated by changing variables `p` and `pdf` to indicate a normal probability distribution. The predicted probabilities over the range of an interval variable were plotted by the GPLOT procedure. This program was written and tested in SAS 8.2 and 9.1.

In order to use this program, users must specify a data set, the dependent variable, and independent variables using macro variables (see `dataset`, `depend`, `independ` in the program). In addition, independent variable names need to be defined in a macro variable `var`. For hypothesis testing, users have to check `p` and `q` matrices in the `wald` module. In the `prtab` module, `r` and `row` indicate the location and values of the independent variable appearing in the row, while `c` and `col` are for the column variable. Note that the 1 by K matrix `baseline` contains baseline values of independent variables. In the `prgen` module, the variable `interval` indicates the location of an interval variable; `start`, `jump` and `cut` are used to define cut points of X-axis of the plot; the `category` indicates the location of a categorical variable; finally, the `group` specifies values of the categorical variable to be compared. See the appendix for the details.

6. CONCLUSION

The nonlinearity of the BRMs results in interpretation problems. This paper discusses effective methods of interpretations such as predicted probabilities, plotting the predicted probabilities, marginal changes, discrete changes, and factor changes in odds. The appropriateness of these approaches largely depends on research questions and audience. For instance, predicted probabilities of ideal types and plotting are more suitable for the audience lacking necessary econometrics backgrounds than other approaches.

The LOGISTIC procedure needs to support the follow-up analyses discussed here. So, I suggest that the procedure have at least two additional statements: say `PREDICT` (for predicted probabilities) and `PCHANGE` (for marginal and discrete changes). These statements seem useful for other categorical dependent variable models (e.g., ordinal regression model and multinomial regression model) and other SAS procedures (e.g., `REG`, `PROBIT`, `NLIN`, and `SYSLIN`) when researchers wish to know predicted values of hypothetical observations. Of course, the `PRCHANGE` statement seems less useful in linear models of the `REG` and `SYSLIN` procedure.

```
PROC LOGISTIC;
  MODEL owncar = budget credits offcamp;
  PREDICT budget=500 credits=10 offcamp=1;
  PCHANGE budget=1000 offcamp=0;
RUN;
```

The usage of the statements is straightforward. Users can specify particular values of independent variables to be used as a baseline. If a variable is not explicitly specified in the statement, its mean is used as the baseline value in computation.

REFERENCES

- Allison, Paul D. 1991. Logistic Regression Using the SAS System: Theory and Application. Cary, NC: SAS Institute.
- Greene, William H. 2000. Econometric Analysis (Fourth edition). Prentice Hall.
- King, Gary. 1998. Unifying Political Methodology: The Likelihood Theory of Statistical Inference. Ann Arbor, MI: University of Michigan Press.
- Long, J. Scott and Jeremy Freese. 2001. Regression Models for Categorical Dependent Variables Using STATA. College Station, TX: STATA Press.
- Long, J. Scott. 1997. Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences. Sage Publications.
- Maddala, G. S. 1983. Limited Dependent and Qualitative Variables in Econometrics. New York: Cambridge University Press.

SAS Institute. 2004 and 1990. SAS/IML Software: Usage and Reference (User's Guide). Cary, NC: SAS Institute.
SAS Institute. 2004 and 1999. SAS/STAT User's Guide, Version 8 and 9.1. Cary, NC: SAS Institute.

ACKNOWLEDGMENTS

I am grateful to J. Scott Long of Indiana University for his insightful and enthusiastic lecture. I would also like to thank John Samuel, Director, Jefferson Davis, Kevin Wilhite, and Takuya Noguchi of the UITs Center for Statistical and Mathematical Computing, Indiana University. Finally, I would like to acknowledge Korea University Computer Club (KUCC) that provided me with valuable opportunity to learn SAS for the first time in 1987.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hun Myoung Park
Indiana University at Bloomington
410 North Park Avenue
Bloomington, IN 47408
812-855-4724
Fax: 812-855-8802
kucc625@indiana.edu
<http://mypage.iu.edu/~kucc625>

APPENDIX: SAS/IML PROGRAM

```

/*****
Binary Logit Regression Model (SAS LOGISTIC and SAS/IML)
Date: July 15, 2004 | Data set: logit.car | Program: logit.sas
Written by Hun Myoung Park (Indiana University at Bloomington)
kucc625@indiana.edu http://mypage.iu.edu/~kucc625

Dependent variable: 1 for owing cars and 0 otherwise
*****/
OPTIONS PAGESIZE=50 LINESIZE=80 NOCENTER NODATE;
LIBNAME logit 'c:\temp';

/*****
*** Linear Probability Model (OLS) *****/
PROC REG DATA=logit.car;
    MODEL owncar = budget credits offcamp;
RUN;

/*****
*** Using the LOGISTIC procedure *****/
PROC LOGISTIC DATA=logit.car DESCENDING COVOUT OUTEST=logit.cov;
    MODEL owncar = budget credits offcamp /COVB STB CTABLE PPROB=.5;
    UNITS budget=sd credits=sd;
RUN;

/*****
*** Macros to define the data set, DV and IV variables, and labels */
%LET dataset=logit.car;
%LET depend=owncar;
%LET independ= budget credits offcamp;
%LET var={"Budget", "Credits", "Offcamp" }; /* Variable names */

/*****
*** Beginning SAS/IML */
PROC IML;
USE &dataset; /* To convert variables of the data set to matrices */
READ ALL VAR{&depend} INTO y; /* To define the dependent variable */
READ ALL VAR{&independ} INTO x; /* To define the IVs */
CLOSE &dataset; /* Close the data set */

parameter={"Interceptor"}//&var; /* Lables including the intercept */
variable=&var; /* Lables without the intercept */

N = NROW(x); /* the number of observations */
x = REPEAT(1, n, 1)||x; /* N by K matrix */
K = NCOL(x); /* To count the number of parameters */
beta = REPEAT(0, K, 1); /* K by 1 matrix */
beta0 = beta+1; /* To check the extent of coefficient changes */
weight = REPEAT(1, n, 1); /* For weighting */

x_bar=x[:]; /* Means of independent variables (1 by K) */
x_sd=SQRT((x[##,]-N*(x_bar#x_bar))/(N-1)); /* Standard deviations of IVs */
x_max=J(1,K,0); /* 1 by K matrix for maximum values of IVs */
x_min=J(1,K,0); /* 1 by K matrix for minimum values of IVs */
DO i=2 TO K; /* To assign max and min values to corresponding matrices */
    x_max[1, i]=MAX(x[,i]); x_min[1, i]=MIN(x[,i]);
END;

/* Maximum Likelihood Estimation Module */
DO i=1 TO 50 WHILE (MAX (ABS (beta-beta0))>1e-8);
    beta0 = beta;
    xb = x*beta; /* N by 1 */

```

```

p = 1/(1+EXP(-xb)); /* N by 1 P(Y=1|X) from the cumulative standard logistic */
pdf = p#p#EXP(-xb); /* N by 1 from the standard logistic function */

*p = PROBNORM(xb); /* For probit model */
*pdf = EXP(-xb#xb/2)/SQRT(8*ATAN(1)); /* For probit model */

lnL = SUM((y=1)#LOG(p) + (y=0)#LOG(1-p))#weight); /* log likelihood (scaler) */
w = weight/(p#(1-p)); /* N by 1 weight matrix */
g = pdf#x; /* N by K matrix of partial derivative of p with respect to betas */
covb = INV(T(g)*(w#g)); /* K by K covariance matrix of betas */
beta = beta + covb*(T(g)*(w#(y-p))); /* K by 1 matrix of next betas */
END;

p0 = SUM((y=1)#weight)/SUM(weight); /* Average response */
lnL0 = SUM((y=1)#LOG(p0) + (y=0)#LOG(1-p0))#weight); /* log Likelihood */
LR = 2*(lnL-lnL0); /* Likelihood ratio (Chi-squared) */
df = K-1; /* Degree of freedom */
p_LR = 1-PROBCHI(LR, df); /* P-value of the Likelihood ratio */
AIC=(1/N)*(-2*lnL+2*K); /* Akaike's Information Criterion */
BIC=-2*lnL-(N-K)*LOG(N); /* Bayesian Information Criterion */
McFadden=1-(lnL/lnL0); /* McFadden's Pseudo R squared */
PRINT ,, '<> Likelihood Ratio Test on the Null Hypothesis of BETA=0',,
      lnL LR df p_LR;
PRINT ,, '<> Goodness of Fit',, Mcfadden AIC BIC,,;

se = SQRT(VECDIAG(covb)); /*Standard Errors of coefficients */
Z = beta/se; /* Z values of coefficients */
Wald = Z#Z; /* Wald Chi-squared statistics of coefficients */
p_B = 2*(1-PROBNORM(ABS(z))); /* P-values of coefficients */
odds=beta|p_B||EXP(beta)||EXP(beta#T(x_sd))||T(x_sd); /* Odds ratio */

PRINT ,, '<> Maximum Likelihood Estimates',, parameter beta se z Wald p_B ;
PRINT ,, '<> Odds Ratio Estimates',, odds[ROWNAME=parameter
      COLNAME={"Beta" "P>|z|" "e^b" "e^(b*sd)" "x_sd"}];
PRINT ,, '<> Estimated Covariance Matrix',,
      covb[ROWNAME=parameter COLNAME=parameter];

/*****
*** Hypothesis Testing: Wald test *****/
START wald; /* Beginning the module */
  W=J(1,3,0);
  Qbr=q*beta-r;
  W[1,1]=T(Qbr)*INV(q*covb*T(q))*Qbr;
  W[1,2]=NROW(r);
  W[1,3]=1-PROBCHI(W[1,1], W[1,2]);
  PRINT ,, '<> Linear Hypothesis Testing',,
    W[COLNAME={"Wald" "df" "P>Chi2"}];
FINISH wald; /* Ending the module */

q={0 1 0 0, 0 0 1 0, 0 0 0 1}; /* Q matrix for constraints */
r={0, 0, 0}; /* R matrix for restrictions */
RUN wald; /* To conduct the Wald test */

/*****
*** Predicted probabilities for ideal types *****/
START prtab; /* Beginning the module */
ideal=J(NROW(row)+1, NCOL(col)+1, 0);
DO i=1 to NROW(row); /* For the variable in row */
  baseline[1,r]=row[i,1]; /* Adjusting the baseline */
  ideal[i+1,1]=row[i,1]; /* Label from the row values */
  DO j=1 to NCOL(col); /* For the variable in column */
    baseline[1,c]=col[1,j]; /* Adjusting the baseline */

```

```

        IF i=1 THEN ideal[1,j+1]=col[1,j]; /* Label from the column values */
        ideal[i+1, j+1]=EXP(baseline*beta)/(1+EXP(baseline*beta));
    END; /* End of column */
END; /* End of row */
PRINT ,, '<> Predicted Probabilities of the Ideal Types',, ideal;
FINISH prttab; /* Ending the module */

baseline=x_bar; /* Baseline using the mean values of IV */
r=4; row={0, 1}; /* Location and values of the row variable */
c=2; col={500 1000 1500}; /* Location and values of the column variable */
RUN prttab; /* To compute the predicted probabilities for the ideal types */

/*****
*** Discrete and marginal changes *****/
START prchange; /* Beginning the module */
dc=REPEAT(0, df, 5); /* Initializing the discrete change matrix */
p1=EXP(baseline*beta)/(1+EXP(baseline*beta));
DO i=1 TO 5; /* Discrete changes */
DO j=1 TO df; /* For individual independent variables */
    x0=baseline; /* Initializing the baseline vector of Xk */
    x1=baseline; /* Initializing the baseline vector of Xk + delta */
    IF i=1 THEN; DO; x0[1,j+1]=x_min[1,j+1];
        x1[1,j+1]=x_max[1,j+1]; END; /* Min-->Max */
    ELSE IF i=2 THEN;
        DO; x0[1,j+1]=0; x1[1,j+1]=1; END; /* 0-->1 */
    ELSE IF i=3 THEN; DO;
        x0[1,j+1]=x0[1,j+1] -.5;
        x1[1,j+1]=x1[1,j+1] +.5; END; /* -1/2-->1/2 */
    ELSE IF i=4 THEN; DO;
        x0[1,j+1]=x0[1,j+1] -(x_sd[1,j+1]/2); /* -sd/2-->sd/2 */
        x1[1,j+1]=x1[1,j+1] +(x_sd[1,j+1]/2); END;

    /* Actual computation of marginal and discrete changes */
    IF i=5 THEN;
        DO; dc[j,i]=p1*(1-p1)*beta[j+1,1]; END; /* Computing marginal changes */
    ELSE; DO; /* Computing discrete changes */
        xb0=x0*beta; dc0=EXP(xb0)/(1+EXP(xb0)); /* Starting */
        xb1=x1*beta; dc1=EXP(xb1)/(1+EXP(xb1)); /* Ending */
        dc[j,i]=dc1-dc0; /* Difference between the starting and ending */
    END; /* End of computing marginal and discrete changes */
    END; /* End of computing a column */
END; /* End of row (independent variables) */

PRINT ,, '<> Discrete and Marginal Changes';
PRINT dc[ROWNAME=variable COLNAME={"Min->Max" "0->1" "+-1/2" "+-sd/2" "Marginal"}];
PRINT "Prob(y=1|x) = " p1; /* Predicted probability given the baseline */
PRINT baseline[COLNAME=parameter]; /* Baseline used in computation */
PRINT x_bar[COLNAME=parameter]; /* Mean of X */
PRINT x_sd[COLNAME=parameter]; /* Standard deviation of X */

FINISH prchange; /* Ending the module */

baseline=x_bar; baseline[1,4]=0; /*baseline={1 400 12.005 0} */
RUN prchange; /* To conduct computation */

/*****
*** Generating predicted probabilities over a range of an IV *****/
START prgen; /* Beginning the module */
nGroup=NCOL(group); /* The number of curves */
pr_gen=J(cut, nGroup+1, 0); /* Initiating the output matrix */
DO i=1 TO nGroup; /* For all values of the group variable */
    baseline[1, category]=group[1,i]; /* To adjust baseline for the group variable */
    DO j=1 TO cut;

```

```

        baseline[1, interval]=start+j*jump; /* Adjusting baselines */
        IF i=1 THEN pr_gen[j,1]=start+(j-1)*jump;
        pr_gen[j,i+1]=EXP(baseline*beta)/(1+EXP(baseline*beta));
    END; /* End of computation for cut point j */
END; /* End of computation for the group i */
PRINT pr_gen;
CREATE WORK.prgen FROM pr_gen; /* To export the matrix into a data set */
APPEND FROM pr_gen;
CLOSE WORK.prgen; /* To close the data set */

FINISH prgen; /* Ending the module */

baseline=x_bar; /* baseline = {1 400 12.005 *} */
interval=2; /* To indicate the locate of the interval variable (budge) */
start=100; jump=100; cut=20; /* Starting value, step, cut points */
category=4; /* To indicate the location of the group variable (offcamp) */
group={0 1}; /* The number of curves (off versus on campus) */
RUN prgen; /* To execute the module */

QUIT; /* Terminate SAS/IML */

DATA WORK.prgen; /* Labeling Variables */
SET WORK.prgen;
LABEL col2='On Campus' col3='Off Campus';
RUN;

GOPTIONS RESET=global GUNIT=pct NOBORDER HTITLE=4 FTITLE=swissb
          HTEXT=3 FTEXT=swiss;
PROC GPLOT DATA=WORK.prgen;
    PLOT (col2 col3)*col1 /OVERLAY LEGEND=legend1 HAXIS=axis1 HMINOR=4
          VAXIS=axis2 VMINOR=1 CAXIS=black
          VREF=0 to 1 by .2 LVREF=2;
    FOOTNOTE1 COLOR=black HEIGHT=3 JUSTIFY=left LSPACE=1
        'Source: the Binary Logit Model for Owning Cars (2004)' ;
    LEGEND1 LABEL=none FRAME CBORDER=white POSITION=(top left inside)
        MODE=share SHAPE=symbol(4,2) VALUE=(justify=center);
    SYMBOL1 COLOR=blue INTERPOL=spline VALUE=circle HEIGHT=4 WIDTH=3;
    SYMBOL2 COLOR=maroon INTERPOL=join VALUE=square HEIGHT=4 WIDTH=3;
    AXIS1 LABEL=(ANGLE=0 'Monthly Budget') ORDER=(100 TO 2050 BY 200)
        MAJOR=(HEIGHT=2) MINOR=(HEIGHT=1) OFFSET=(2,2) WIDTH=1 ;
    AXIS2 LABEL=(ANGLE=90 'Pr(y=1): Predicted Probabilities of Owning Cars')
        OFFSET=(1,0) WIDTH=1;
RUN;

```