"Logistic Regression Analysis in Higher Education: An Applied Perspective

Chapter · January 1994

CITATIONS
74

READS
1,986

1 author:

Alberto F Cabrera
University of Maryland, College Park
85 PUBLICATIONS **3,420** CITATIONS

SEE PROFILE

Cabrera, A. F. (1994). *Logistic regression analysis in Higher Education: An applied perspective.* In John C. Smart (ed.), Higher Education: Handbook of Theory and Research (225-256). Volume 10. New York: Agathon Press.

Logistic Regression Analysis in Higher Education: An Applied Perspective*

Alberto F. Cabrera
School of Education, SUNY—Albany

Deciding to attend college, choosing a particular institution over another, dropping out before completing a degree, majoring in a particular academic discipline, and transferring from one institution to another are examples of some college outcomes educational researchers constantly deal with when addressing issues affecting higher education institutions. As the seasoned educational researcher already knows, how well one answers these research questions can substantially affect how effective institutional programs will be in addressing such college behaviors as enrollment and retention (Hossler, 1991). In dealing with these behaviors, however, the researcher faces two main problems. First, college outcomes are the product of an array of factors in which both student characteristics and those of the institution interact among themselves (Pascarella and Terenzini, 1991). Second, many college outcomes are dichotomous in nature. There are no interval scales to describe such behaviors. Either an individual attends college or not, majors in hard sciences or not, stays or leaves the institution, or obtains a bachelor degree or not.

While a growing number of conceptual frameworks can assist the researcher in identifying what factors are relevant to the particular college behavior under consideration (see Pascarella and Terenzini, 1991), the task of quantifying the effect these factors have may be constrained by the nature of the college behavior under consideration. Although several statistical techniques are available, only a few of them conform to the specific dichotomous nature of outcome measures such as enrollment, persistence, and degree attainment. These include: structural modeling for dichotomous dependent variables (e.g., Bentler, 1989; Jöreskog and Sörbom, 1988; Muthén, 1988), log-linear analysis (e.g., Christensen, 1990; Hinkle, Austin and McLaughlin, 1989; Marascuilo and Serlin, 1988), discriminant analysis (e.g., Marascuillo and Levin, 1983), probit regression and logistic regression (e.g., Fienberg, 1983; Hanusheck and Jackson, 1977).

*The author would like to thank Maria B. Castañeda, SUNY-Albany, Barbara Zusman, University of Illinois Research Laboratory, Amaury Nora, University of Illinois at Chicago, W. Paul Vogt, SUNY-Albany, and Ernest Pascarella, University of Illinois at Chicago, for their invaluable comments and suggestions.

The application of logistic regression in higher education to deal with dichotomous dependent variables is not new. Its use can be traced back to the late '60s and early '70s. During these decades efforts were made at developing econometric models to explain college choice (see Manski and Wise, 1983). Bishop (1977), for instance, employed this technique to study college enrollment decisions and how responsive these decisions were to tuition and student aid programs. Likewise, Manski and Wise (1983) relied on multivariate logistic regression in addressing the role of the Basic Educational Opportunity Grant (BEOG) in facilitating college choice and enrollment for the high school class of 1972. More recently, St. John (1990, 1991) and St. John and Noell (1989) relied on logistic regression to address the role of tuition on college attendance and to estimate the effect of student aid in facilitating college attendance on the part of minorities. The application of logistic regression has not been restricted to college enrollment. Behaviors such as college persistence, transfer decisions, and degree attainment have also elicited the use of this technique.

Stage (1989), for instance, used a combination of logistic regression and LISREL for validating Tinto's (1975, 1987) model of college persistence. Stampen and Cabrera (1986, 1988) used logistic regression for aggregate data in exploring the extent to which student aid equalized opportunities to persist in college. St. John, Kirshstein and Noell (1991) utilized logistic regression to document the effects of financial aid on year-to-year college persistence for the 1980 high school senior cohort. Cabrera, Stampen and Hansen (1990) used this method to explore the effects of ability to pay on the persistence process. Dey (1991) relied on logistic regression to explore determinants of persistence to graduation for a national sample of college students.

The purpose of this chapter is to provide a basic introduction to the use of logistic regression. The chapter was written with a specific audience in mind: educational researchers seeking basic information about how logistic regression can be used in addressing policy questions involving dichotomous outcomes[1]. The focus of the chapter is not on the mathematical foundations underlying logistic regression. These foundations are discussed in detail by Aldrich and Nelson (1986), Backer and Nedler (1988), Christensen (1990), Collett (1991), Fienberg (1983), Freeman (1987), Hanusheck and Jackson (1977), and Maddala (1987), among others. Rather, the emphasis is on the practical applications of logistic regression. The chapter is organized into three sections. Section I introduces the reader to the underlying assumptions associated with the logistic re-

[1]Logistic regression need not to be confined to the analysis of dichotomous dependent variables. Weiler (1987), for instance, relied on nested multinomial logistic regression to model the effect of factors affecting decisions of nonattendance, attendance at a four-year college or university, attendance at a community college, or attendance at a technical institute. Examples on the use of multinomial logistic regression can be found in Aitken, Anderson, Francis and Hinde (1990) and in Aldrich and Nelson (1984).

gression model. Section II presents a step-by-step illustration of the use of logistic regression. This section also introduces the reader to two statistical packages that handle logistic regression: the Generalised Linear Interactive Modelling (GLIM) and SPSS-X (Backer and Nedler, 1988; Norusis, 1990). In illustrating the results, particular effort is placed on discussing methods available for assessing alternative models, indicators of goodness of fit, procedures for testing the statistical significance of variables, and interpretation of the results in practical terms. Finally, Section III briefly summarizes the controversy surrounding the use of Ordinary Least Square (OLS) over Logistic regression.

I. ASSUMPTIONS UNDERLYING THE USE OF LOGISTIC REGRESSION

There are two main assumptions underlying the use of logistic regression. The first deals with the nature of the distribution associated with the binary outcome; the second deals with the nature of the relationship between the outcome variable (e.g., persistence) and the independent variable(s) (e.g., financial aid).

Under the first assumption, it is presumed that each of the potential values of the outcome variable Y (0 or 1), has a corresponding expected probability that varies as a function of the values that the independent variable(s) can take for each subject. Statistically this statement can be expressed as follows:

$$E[Y_{ith} = 1 / X = x] = P(Y_{ith} = 1)$$

where $P(Y_{ith} = 1)$ represents the probability of observing the condition of success (i.e., persisting) for the ith subject given a particular value of X. These probabilities are assumed to follow a binomial distribution (see Plane and Oppermann, 1977). A unique characteristic of this type of distribution is that although the probability distribution has an overall mean referred to as P (i.e., the proportion of subjects that meet the criterion of success), the variance changes as a function of the subject under consideration. The variance for each subject ($V_{(ith)}$) is expressed as follows:

$$V_{(ith)} = P(Y_{ith} = 1) * [1-(P_{ith} = 1)]$$

where $P(Y_{ith} = 1)$ represents the probability of observing the condition of success, say persisting, for the ith subject and $[1-(P_{ith} = 1)]$ represents the probability of not observing the condition of success, say dropping-out, for the ith subject.

As far as the nature of the relationship between a binary outcome and a given independent variable is concerned, the logistic regression model presumes that this association can be accounted for by a logistic function (Collett, 1991; Hanusheck and Jackson, 1977). In the case of one independent variable, the logistic function takes the following form:

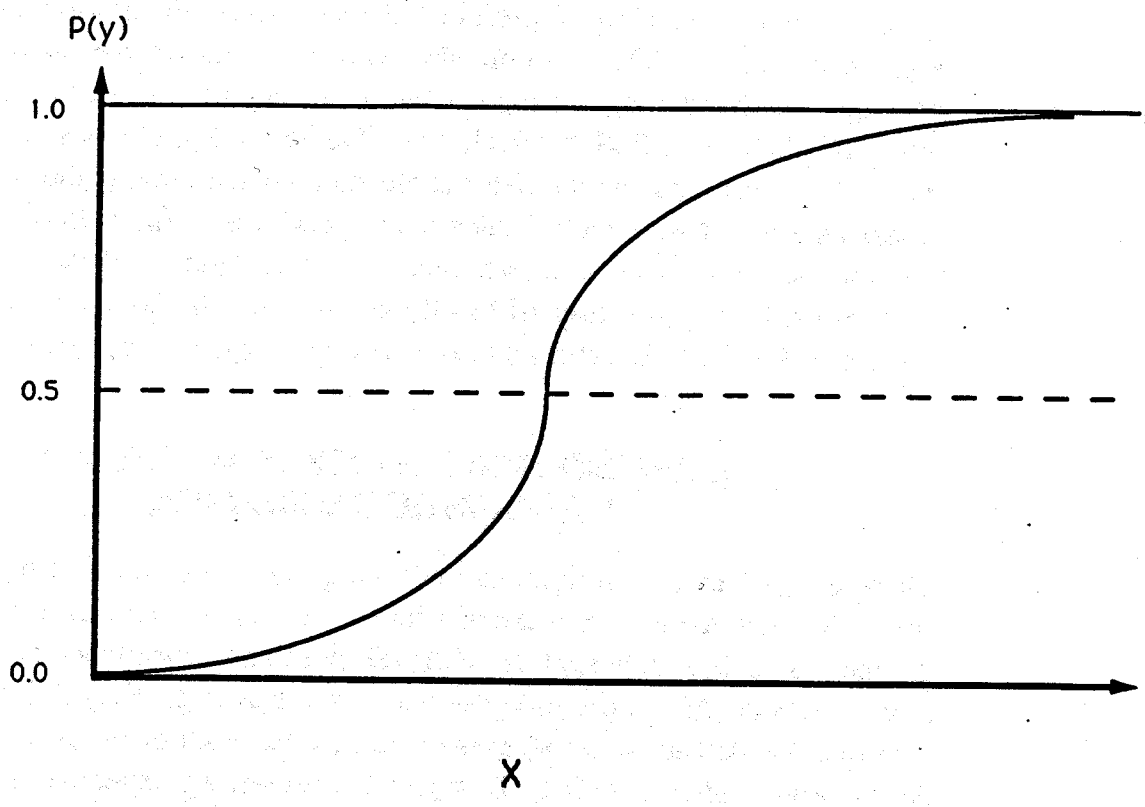

FIGURE 1. The logistic distribution.

$$L = \ln \frac{P(Y)}{1 - P(Y)} = B_0 + B_1 X_1 \quad (1)$$

where L is called the logit or the natural logarithm of the odds ratio, B_0 and B_1 refer to the familiar intercept and beta weight and $P(Y)$ stands for the expected probability of Y across different values of X. Since probabilities are the focus of analysis, equation (1) can be restated as follows:

$$P(Y) = \frac{\exp(B_0 + B_1 X_1)}{1 + \exp(B_0 + B_1 X_1)} \quad [2]$$

As shown by Aldrich and Nelson (1986), the logistic distribution is *S*-shaped and has values ranging from .0 to 1.0 as $B_1 X_1$ changes from negative infinity to positive infinity. Figure 1 displays a logistic distribution function.The estimation of the equation's parameters (intercept and betas) rests upon a method called Maximum Likelihood (ML). ML basically assumes that the underlying relationship between an independent variable and a dichotomous dependent variable (Y) follows a probability function (also called likelihood function)[2] which is

[2]A detailed discussion of the nature of the probability or likelihood function and its properties can be found in Christensen (1990), Collett (1991), Hanusheck and Jackson (1977) and Fienberg (1984).

S-shaped in nature (see figure 1). There are two types of probability distributions that can be employed to study the relationship between the dichotomous dependent variable *Y* and the independent variable(s). These are the probit distribution and the logistic distribution (Hanusheck and Jackson, 1977). Of the two, the logistic distribution is the most frequently used[3]. Analyses based upon the logistic distribution are usually referred to as logistic regression. Based on the assumption that the relationship between the dichotomous dependent variable and an independent variable can be represented by a logistic distribution (see Figure 1 and equation 1), the probability of the dependent variable (P) is estimated for each group (in the case of grouped data) or for each subject (in case of individual data). The natural logarithm of the odds (logit) is employed to transform the probability estimates into a continuous variable. Next, the ML approximation to logistic regression seeks to assess the effects of the independent variable upon the probability function. This is accomplished through an iterative process of estimation whereby estimates for the intercept and for betas are chosen so as to maximize the likelihood of reproducing the observed probability value for *Y*. Thus, the principle underlying ML approximation can be simply summarized as deriving those estimates for the intercept and the slopes that would make the *likelihood* of having observed *Y* the highest (Hanusheck and Jackson, 1977).

As noted by Aldrich and Nelson (1984), the principle followed by ML for estimating parameters for the intercepts and the slopes is quite similar to that employed by the Ordinary Least Squares (OLS) regression model, with a major conceptual difference. While OLS is concerned with choosing those parameter estimates that would minimize the sum of the squared errors between the observed and predicted *Y*s, ML estimation seeks to choose those estimates that would yield the highest probability or likelihood of having obtained the observed probability *Y*.

Several computer programs are available to handle logistic regression models. These programs rest on different approximations to the Maximum Likelihood method and consequently are likely to yield slightly different results. The most popular are GLIM and SPSS-X[4]. These programs are available for both main frame and PC environments. The illustration of the use of logistic regression for individual data will be based on GLIM version 3.77 (Backer and Nedler, 1988) and on SPSS-X version 4.0 (Norusis, 1990) both for the PC environment.

[3]The probit distribution and the logistic distributions are very similar. The difference between the two functions shows up in the tails of the distributions with the probit distribution approaching the axes faster than the logistic distribution does (see Hanusheck and Jackson, 1977). The logistic approximation is usually preferred over probit given its convenient mathematical properties (Hanusheck and Jackson, 1977; Norusis, 1990).

[4]Options to handle the logistic regression model are also available in SAS and in BMDP.

II. HANDLING OBSERVATIONS: A MICRODATA APPROACH

The logistic regression model is quite flexible as to the unit of analysis it can handle (Christensen, 1990; Hanusheck and Jackson, 1977). The logistic regression model can be employed when the unit of analysis is the individual subject, or in those circumstances in which the group[5], rather than the individual subject, is the unit of analysis. The application of logistic regression for individual observations (also called logistic regression for microdata) is adaptable to the metric of the independent variable under consideration. The microdata approach can be applied where the independent variables are truly categorical, such as gender, ethnicity, type of major, or where the variable has been categorized, such as low, middle and high income, or where continuous and categorical variables are mixed together such as type of major and GPA[6].

Several examples can be found in the literature that include applications of these two approaches in higher education. Manski and Wise (1983), for instance, employed the aggregate approach to study the effect of SAT (broken down into two categories), family income (two categories), region (two categories), ethnicity (two categories), and being a high school leader (two categories) on the likelihood of applying for and being admitted to college, for a representative sample of the high school class of 1972. More recently, St. John (1991) employed the microdata approach to study the effect of academic preparation, academic ability, region, social background, and degree aspirations on the probability of attending college for a representative sample of the high school class of 1982. Moreover, St. John's study provides an excellent illustration of how logistic regression for microdata data can bring together continuous and categorical variables in attempting to predict a dichotomous dependent variable.

There are several procedures that can be utilized for model testing, assessing the fit of a given model, estimating the statistical significance of the parameters and selecting alternative models. The reader familiar with OLS will find that these procedures bear a striking resemblance to those employed for assessing models under the OLS approach. These methods, as they apply to microdata, will be illustrated with a data base on the outcomes of doctoral work.

An Illustration

Background. This illustration utilizes a data base on graduate students created by a major midwestern doctoral granting institution. The data base is the product of an ongoing project aimed at studying determinants of graduate related behaviors (see Baird and Smart, 1991; Ethington and Bode, 1992; Smart, Baird and

[5]The reader interested in the use of logistic regression for aggregate data is referred to Fienberg (1984), Freeman (1988) and Christensen (1990). Aitken et al. (1990) also provide several illustrations.

[6]For a comprehensive discussion of the differences between the two approaches, see Collett (1990) and Hanusheck and Jackson (1977).

Rode 1991). The doctoral student subsample employed by Nora, Cabrera and Shiville (1992) for testing a model of engagement in graduate related behaviors will be employed. The specific focus of the example is one of those behaviors, namely, determinants of research engagement.

Variables. The dependent variable under study is research engagement, this is a dichotomous variable; doctoral students who indicated that they participated with faculty or peers in research projects were coded as 1. Those who indicated otherwise were coded 0. Consequently, the dependent variable Y had two potential values. In order to facilitate the illustration of logistic regression only eleven of the independent variables employed in the Nora et al. (1992) study were analyzed. Two indicators were employed to measure background variables. These were father's education, an ordinal variable made up of five levels ranging from 1 = high school diploma or less to 5 = PhD or advanced degree, and gender (1 = female, 2 = male). Pre-Commitments were measured through the Pre-institutional commitment scale, a composite averaged across two items, and Pre-professional commitment, made up of a single item measured in a Likert scale ranging from 1 (not important) to 5 (very important). Academic Contacts with Faculty, a composite scale averaged across five items, and Interactions with Peers, a composite scale averaged across four items, provided measures of a student's integration with the academic and social components of the department in which the doctoral student was enrolled. Final vocational and professional commitments were measured through a single Likert-scaled item. Academic growth and development was measured via a Conceptualizing Skills scale, a composite averaged across six items, and a Research Skills scale, a composite averaged across four items. These scales basically measure the extent to which students reported gains in conceptualization and research skills resulting from experiences with their respective academic departments. Using the Biglan's subject matter classification scheme (Biglan, 1973a, 1973b), students were also classified as either majoring in soft majors (major = 1) or majoring in hard majors (major = 2). This variable was included to account for the effect that a particular discipline may have on scholarly behaviors. Table 1 displays summary statistics and alpha reliabilities for each variable.

Creating the Data Base. Preparing data bases for individual cases for GLIM and SPSS-X is similar to the process employed for other canned programs. Columns are used to represent variables while rows stand for individual cases across variables. Table 2 displays the case number, the columns representing the respective variable, values for each independent variable (X_1 throughout X_{11}) and the observed probability of conducting research (Y).

Coding Schemes for Categorical Variables. In GLIM, the specification of categorical and ordinal variables is quite simple, for the program contains options allowing the researcher to specify the categorical variable with its respective levels. In GLIM, such a categorical variable as Father's education can be iden-

TABLE 1. Descriptive Statistics and Marginal Distributions

Variables	Count	Cell %	Mean	S.D.	Alpha
Dependent Variable:					
Research Engagement (RENG)					
1. Yes	263	49.4	–	–	–
0. No	269	50.6	–	–	–
Independent Variables:					
Father's Education (FAED)					
1. High School or less	224	42.1	–	–	–
2. Some college	79	14.8	–	–	–
3. Bachelor's degree	95	17.9	–	–	–
4. Master's or equivalent	55	10.3	–	–	–
5. Ph.D. or advance	79	14.8	–	–	–
Sex					
1. Female	284	53.4	–	–	–
2. Male	248	46.6	–	–	–
Pre-Commitments					
Pre-Institutional (PINT)	–	–	2.33	0.87	0.864
Pre-Professional (PPRF)	–	–	3.31	0.84	–
Integration:					
Faculty (FACU)	–	–	3.47	0.98	0.848
Peer	–	–	3.04	0.99	0.627
Final Commitments:					
Professional (PROF)	–	–	4.50	0.89	–
Vocational (VOCA)	–	–	3.72	1.25	–
Academic Growth:					
Conceptualizing Skills (CSKL)	–	–	2.91	0.59	0.811
Research Skills (RSKL)	–	–	3.01	0.68	0.748
Major (MJR)					
1. Soft	282	53.0	–	–	–
2. Hard	250	47.0	–	–	–

tified by a single variable made up of five categories. Unlike SPSS-X, the use of the coding 0 and 1 in GLIM is reserved for the dependent variable, while categorical variables can be represented by any digit provided that 0's are not included. In the case of the variable *MAJOR*, for instance, the code 1 represents those subjects majoring in soft disciplines while code 2 represents those subjects majoring in hard disciplines. For the dependent variable, the code 1 represents students that reported being engaged in research ($y = 1$) while 0 is reserved for those students who indicated otherwise ($y = 0$). Displayed in Table 1 is the coding system employed for analyzing the data in GLIM.

In SPSS-X, two-category variables—also called dummy or indicator variables—are coded as 0 or 1 to signify the presence or absence of the classification

TABLE 2. Layout of Data for Logistic Regression

Case #	FAED (X1)	SEX (X2)	PINT (X3)	PPRF (X4)	FACU (X5)	PEER (X6)	PROF (X7)	VOCA (X8)	CSKL (X9)	RSKL (X10)	MJR (X11)	RENG (Y)
1	3.00	1.00	3.00	3.00	4.00	4.67	5.00	4.00	3.00	3.50	1.00	0.00
2	1.00	2.00	2.00	3.00	3.60	2.00	5.00	2.00	3.17	3.50	1.00	1.00
3	1.00	2.00	2.00	4.00	1.00	2.67	5.00	5.00	3.67	3.75	2.00	0.00
4	1.00	1.00	3.00	2.00	2.60	2.67	4.00	3.00	3.17	3.50	1.00	1.00
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
532	2.00	1.00	2.00	1.00	3.20	4.33	3.00	3.00	4.00	4.00	1.00	1.00

condition. In the case of the variable *MAJOR*, for instance, the code 0 represents those subjects majoring in soft disciplines while the code 1 represents those subjects majoring in hard disciplines. When the categorical independent variable under consideration is made up of more than two categories, new variables need to be created to represent the categories. In SPSS-X, the handling of the five-level Father's education variable involves the creation of four dummy variables. These are: *COLLEGE* = 1 for students whose father has some college and 0 otherwise, *BACHELOR* = 1 for students whose father has a bachelor degree and 0 otherwise, *MASTER* = 1 for students whose father has a master degree and 0 otherwise, and *PHD* = 1 for those students whose fathers have a Ph.D. degree or equivalent and 0 otherwise (see Table 4). Since the categorical independent variables were created to meet GLIM specifications, these variables were recoded to meet SPSS-X requirements as shown in Table 4.

Reading the Data Base. There are two procedures to supply GLIM and SPSS-X with the data. The first is iterative; that is, the data can be provided once GLIM or SPSS-X is invoked. The second procedure relies on creating an ASCII file and storing it under the directories containing the GLIM and SPSS-X programs. The latter is recommended particularly in those cases where large data bases are in place. The complete set of GLIM commands for reading and performing the series of logistic regressions employed in this illustration is displayed in Table 3. The equivalent SPSS-X commands are listed in Table 4.

GLIM uses the terminology "**$UNITS**" to refer to cases or individuals. The command **$UNITS 532 $** instructs GLIM that information on 532 subjects is to be processed. The command **$FACTOR** identifies the categorical independent variables along with their respective number of categories. As noted, the command identifies three categorical variables; namely, Father's education (FAED), Gender (SEX) and Major (MJR). In the case of FAED, the command

TABLE 3. Commands for Logistic Regression Using GLIM

```
$UNITS 532 $
$FACTOR FAED 5 SEX 2 MJR 2 $
$VARIATE PINT PPRF FACU PEER PROF VOCA CSKL RSKL $
$FORMAT FREE $
$DATA FAED SEX PINT PPRF FACU PEER PROF VOCA CSKL RSKL MJR
RENG $
$DINPUT 13 $
  File name? PHD.DAT
$CALC N = 1.0 $
$YVARIATE RENG $
$ERROR B N $
$LINK G $
$FIT FAED+SEX $
$DISPLAY D E R $
$FIT FAED+SEX+PINT+PPRF $
$DISPLAY D E R $
$FIT FAED+SEX+PINT+PPRF+FACU+PEER $
$DISPLAY D E R$
$FIT FAED+SEX+PINT+PPRF+FACU+PEER+PROF+VOCA $
$DISPLAY D E R$
$FIT FAED+SEX+PINT+PPRF+FACU+PEER+PROF+VOCA+CSKL+RSKL $
$DISPLAY D E R$
$FIT FAED+SEX+PINT+PPRF+FACU+PEER+PROF+VOCA+CSKL+RSKL+MJR $
$DISPLAY D E R $
$FIT SEX+PINT+PPRF+FACU+PEER+PROF+VOCA+CSKL+RSKL+MJR $
$DISPLAY D E R $
$FIT 1 $
$DISPLAY D $
$STOP
```

identifies a categorical variable made up of five values or levels. The command **$VARIATE** specifies the continuous independent variables; namely, Pre-Institutional Commitment (PINT), Pre-Professional Commitment (PPRF), Interactions with Faculty (FACU), Interactions with Peers (PEER), Professional Commitment (PROF), Vocational Commitment (VOCA), Conceptualization Skills (CSKL) and Research Skills (RSKL). The command **$DATA** identifies the particular order in which information for each variable was stored. In this case: FAED, SEX, PINT, PPRF, FACU, PEER, PROF, VOCA, CSKL, RSKL, MJR, and RENG. In GLIM, like in SPSS-X, data may be read in either fixed or free format. In either mode, the data are read by GLIM on a row basis. Under the **FIXED** format, Fortran commands are used to identify the specific number of variables and their location in the file. If a **FREE** format is used, it is advisable that the variables be separated at least by one space. The command

TABLE 4. Commands for Logistic Regression Using *SPSS-X*

```
DATA LIST FILE 'PHD.DAT' FREE /FAED SEX PINT PPRF FACU PEER PROF
                                VOCA CSKL RSKL MJR RENG.
COMPUTE COLLEGE = 0.
COMPUTE BACHELOR = 0.
COMPUTE MASTER = 0.
COMPUTE PHD = 0.
IF (FAED EQ 2) COLLEGE = 1.
IF (FAED EQ 3) BACHELOR = 1.
IF (FAED EQ 4) MASTER = 1.
IF (FAED EQ 5) PHD = 1.
COMPUTE NSEX = 0.
IF (SEX EQ 2) NSEX = 1.
COMPUTE NMJR = 0.
IF (MJR EQ 2) NMJR = 1.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX /CASEWISE = PRED RESID ZRESID.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX PINT PPRF/CASEWISE = PRED RESID ZRESID.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX PINT PPRF FACU PEER/CASEWISE = PRED RESID ZRESID.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX PINT PPRF FACU PEER PROF VOCA/CASEWISE = PRED RESID ZRESID.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX PINT PPRF FACU PEER PROF VOCA CSKL RSKL/CASEWISE = PRED
RESID ZRESID.
LOGISTIC REGRESSION RENG WITH COLLEGE BACHELOR MASTER PHD
NSEX PINT PPRF FACU PEER PROF VOCA CSKL RSKL NMJR/CASEWISE =
PRED RESID ZRESID. LOGISTIC REGRESSION RENG WITH NSEX PINT.
PPRF FACU PEER PROF VOCA CSKL RSKL NMJR/CASEWISE = PRED RESID
ZRESID.
FINISH.
```

$DINPUT 13 $ instructs GLIM that the raw data file is located under the directory containing the GLIM programs and prompts GLIM for a file name. In this case, **PHD.DAT**. The **$YVARIATE** directive identifies the dependent variable; namely, Research Engagement (RENG).

The specification of the data for SPSS-X is varied and simpler as compared with GLIM. The data can be read from a file already defined for SPSS-X—the SPSS-X system file—through the command **GET FILE** or from an ASCII file via the command **DATA LIST FILE**. In the example (see Table 4), the command **DATA LIST FILE** indicates that the data are to be read on a **FREE** format basis from a file named **PHD.DAT** whose variables are listed in the sequence

identified in the command. That is, FAED SEX PINT PPRF FACU PEER PROF VOCA CSKL RSKL MJR RENG (see Table 4).

Model Testing. To fit logistic regression models to binary dependent variables in GLIM, it is necessary to establish that the probability distribution associated with the dependent variable is binomial (B). In turn, the binomial distribution depends on both n, number of trials in which Y = 1 is observed and p, the observed probability. The commands **$CALC N = 1.0 $** and **$ERROR B N $** meet these two objectives; that is, the identification of n for which the observed probability is to be computed, the value of n for which the condition in the dependent variable is met ($Y = 1$) and the underlying distribution associated with the dependent variable (B or Binomial). The command **$LINK G $** makes reference to the likelihood function to be employed in estimating the slopes and the intercept. In this particular case, **$LINK G $** instructs GLIM to use the ML for logistic distributions[7]. In GLIM, the **$FIT** command is employed in specifying the models to be estimated. In this example, seven alternative models were specified. For each model under estimation, the **$DISPLAY** command specifies the output. The **D** command prints the scaled deviance and degrees of freedom for the model. Option **E** displays the estimates of the parameters and the standard deviations for each parameter[8]. Option **R** lists for each subject the observed probability of engaging on research activities, the predicted probability of engaging on research activities, and the standardized residual (see Table 3). The standardized residuals are obtained by scaling the difference between observed probabilities and predicted probabilities under the model[9]. Standardized residuals provide an indication of how well the model fits the data. In general, standardized residuals greater than two (2) in absolute value signify that the model produced a poor prediction for the particular observation under consideration (see Aitkin et al., 1990; Fienberg, 1984). The use of these options will become evident as the results of the logistic models are discussed.

[7]There are two possibilities that the educational researcher can use to define the relationship between the independent and the categorical dependent variable. These are G for the logit model and P the probit model. For a technical discussion of the available models see Backer and Nedler (1988) and Aitken et al. (1990).

[8]There are many options that the user can employ when estimating a logistic model. Option S, for instance, requests for standardized residuals and fitted values which are particularly useful when assessing the extent to which the model under or over-predicts the observed frequencies for the dependent variable. For a detailed discussion of these options see Backer and Nadler (1988).

[9]The formula to compute the standardized residuals is as follows:

$$e_{ith} = \frac{Y_{ith} - P_{ith}}{P_{ith} * (1 - P_{ith})}$$

where e_{ith} represents the standardized residual for the ith subject, Y_{ith} the observed probability for the ith subject and P_{ith} the predicted probability under the logistic model for the ith subject. The expression $P_{ith} * (1 - P_{ith})$ represents the variance.

The equivalent GLIM program for SPSS-X is listed in Table 4. Fitting logistic regression models in SPSS-X is accomplished with the command **LOGISTIC REGRESSION**[10]. In this command, the dependent variable is listed first while the independent variables are specified after the directive **WITH** (see Table 4). By default, SPSS-X treats variables coded as either 0 or 1 as dummy variables. The options **DEVIATION, CATEGORICAL** and **CONTRAST** within the SPSS-X's procedure **LOGISTIC REGRESSION** can be used to identify independent dummy variables originally coded with values other than 0 and 1 or to handle categorical variables made up of more than two categories or levels[11] (see Norusis, 1990). In the example, categorical variables originally coded with values other than 0 or 1 were recoded to meet the SPSS-X default option for categorical variables (see lines 3 through 13 in Table 4). The command **CASEWISE** in SPSS-X is employed to specify options to examine the adequacy of the resulting logistic regression models. The options **PRED RESID ZRESID** estimate for each subject the observed probability of engaging on research activities, the predicted probability of engaging on research activities, and the standardized residual[12].

Hierarchical Testing of Models

Testing Alternative Models. In OLS, testing alternative models basically rests on assessing whether or not adding or deleting variables accounts significantly for changes in the proportion of variance in the dependent variable (Pedhazur, 1982). Such tests basically rest on a comparison between the coefficients of determination (R^2) associated to each model under consideration, and an assessment of the statistical significance of any observed change in R^2s.

A similar test is available for assessing alternative logistic regression models. The core of this test rests on the maximum likelihood function (usually referred to as G^2 or scaled deviance[13]) associated with a particular logistic regression model (Aldrich and Nelson, 1986; Collett, 1990; Fienberg, 1983; Freeman, 1987). The maximum likelihood function statistics (G^2) provides an overall

[10]To estimate probit regression models, the **PROBIT** command is also available.

[11]Care should be exercised when using these options. By default, these options can create new classification schemes that may not conform to the classification scheme originally intended by the researcher. Furthermore, these different classification schemes can generate different regression coefficients that can lead the researcher to reach incorrect conclusions as to the effect of the variables under consideration (see Norusis 1990, pp. 52–56).

[12]The output generated by this program is available upon request.

[13]The manner in which the maximum likelihood function is reported varies. Some of the terms most commonly used are: the −2 Log L (Aldrich and Nelson, 1984; St. John, 1991), the −2 Log-likelihood value (Manski and Wise, 1983; Norusis, 1990), and the G^2, also called "scaled deviance" (Christensen, 1990; Collett, 1991; Feinberg, 1984; Freeman, 1987). In GLIM, the maximum likelihood function is reported as "scaled deviance" while in SPSS-X this measure is reported as −2 Log Likelihood.

indication of how well the estimates for the parameters in the model fit the data[14]. Nevertheless, unlike the R^2, the maximum likelihood function per se has little value in judging whether or not the model is a valid one. Unlike the R^2 in OLS, the maximum likelihood function does not represent the proportion of variance explained in the dependent variable. As already noted, the focus of analysis in logistic regression is not the matrix of interrecorrelations among the dependent and independent variables but, rather, the probability of a given outcome, and finding those estimates for the slope and the intercept that maximize the likelihood of reproducing the observed probability. Therefore, the utility of the maximum likelihood function statistics lies on assessing alternative models (Fienberg, 1984; Collett, 1991).

The test basically involves a comparison of the likelihood function between two alternative models. In logistic regression, the best fitting model is the one that yields a significantly smaller G^2. This test is carried on by comparing the differences in G^2s between two given models to a Chi-square distribution table with degrees of freedom equal to the difference in degrees of freedom between two alternative models[15]. In logistic regression, reductions in G^2 with an associated p-value less than .05 indicate that the model accounts for a significant improvement of fit.

Forward and Backward Stepwise Processes. As in OLS, the estimation of alternative models in logistic regression can follow a forward stepwise process or a backward stepwise process (Pedhazur, 1982). Under the forward stepwise approach, individual variables or groups of variables are added in a sequential manner, and the validity of the added variables is judged in terms of significant improvements of fit. The backward stepwise process basically consists of deleting groups or individual variables and assessing the extent to which their deletion significantly worsens the model (Fienberg, 1983). Both approaches have been applied to study dichotomous dependent variables in higher education (see St. John, 1991; Cabrera et al., 1990; Mallete and Cabrera, 1991).

The use of the forward stepwise approach is illustrated in Table 5[16]. Each column displays the parameter estimates for the specific model under consideration. For each model, the scaled deviance G^2 and corresponding degrees of freedom are also presented. As shown in Table 5, six models, referred as steps, were sequentially estimated. The sequence of this testing was dictated by the

[14]By convention, a good fit is evidenced when the maximum likelihood function is close to the degrees of freedom for the respective model (see Cabrera et al., 1990; Mallette and Cabrera, 1991; St. John, 1991).

[15]The chi-square test is used based on the fact that the difference in G^2s follows a Chi-square (X^2) distribution (see Feinberg, 1984).

[16]This table is based on results generated by GLIM. The interested reader will notice that SPSS-X produces slightly different results. These differences are attributable to the fact that the programs take different approaches to the maximum likelihood solution.

TABLE 5. Effects of Background, Prior Commitments, Integration, Academic Growth, and Major on Research Engagement

Factor	Step 1 (Beta)	Step 2 (Beta)	Step 3 (Beta)	Step 4 (Beta)	Step 5 (Beta)	Step 6 (Beta)
Father's Education:						
Some College	0.3007	0.3169	0.2770	0.2906	0.3248	0.2322
Bachelor	0.3989	0.3529	0.3440	0.3462	0.3613	0.2562
Master	0.4049	0.3963	0.3416	0.3475	0.4728	0.2847
Ph.D.	0.0669	0.0335	−0.0720	−0.0691	0.0302	0.0495
Gender (Male)	0.4478**	0.4173**	0.5134**	0.5182**	0.5356**	0.3980*
Pre-Institutional		−0.1753	−0.2639**	−0.2675**	−0.2968**	−0.2867**
Pre-Professional		−0.0028	−0.0770	−0.0821	−0.0649	−0.0966
Faculty			0.4172**	0.4128**	0.3278**	0.3361**
Peer			0.2313**	0.2255**	0.1869*	0.1553
Professional Comit.				0.0624	0.0633	0.0758
Vocational Comit.				−0.0048	−0.0475	−0.0475
Conceptualization					−0.2325	−0.1831
Research					0.7712**	0.7276**
Major (Hard Major)						0.6314**
Intercept	−0.3991	0.0444	−1.6690	−1.8790	−2.9890	−3.0960
G^2	727.070	724.170	696.120	695.770	673.610	663.290
df	526.000	524.000	522.000	520.000	518.000	517.000
G^2 */df*	1.382	1.382	1.334	1.338	1.300	1.283
"R^2"	0.019	0.024	0.072	0.073	0.107	0.122
PCP	56.02	57.33	59.4	59.77	63.16	64.85
X^2, *df*	10.37, 5	13.27, 7	41.32**, 9	41.67**, 11	63.83**, 13	74.15**,14

* = $p < .05$ ** = $p < .01$

pattern suggested in the model advanced by Nora et al. (1992). Model 1 (step 1) represents the effect of background characteristics on the likelihood of engaging in research activities. Model 2 represents the added effect attributable to Pre-Commitment factors while Model 6 represents the incremental effect on the outcome variable attributable to major while taking into account background characteristics, pre-commitments, interactions with faculty and peers, and academic growth.

As shown in Table 5, the addition of the groups of variables in each of the six steps appears to increase the ability to predict research engagement. This is suggested by the reduction in the G^2 across the six models. At this point, however, the G^2 associated with model 6 only provides an indication that this model appears to fit the data better than its predecessors. A more rigorous test of the statistical significance of the alternative models is displayed in Table 6.

TABLE 6. Effects of Adding Factors on the Fit of the Model

Model	*df*	G^2	Change in *df*	Change in G^2	Improvement of Fit *p*-value
1. Background Only	526	727.07			
2. Adding Pre-Commitments	524	724.17	$df_1 - df_2 = 2$	$G^2_1 - G^2_2 = 2.90$	.2466
3. Adding Integration	522	696.12	$df_1 - df_3 = 4$	$G^2_1 - G^2_3 = 30.95$	.0005
4. Adding Commitments	520	695.77	$df_1 - df_4 = 6$	$G^2_1 - G^2_4 = 31.30$	.0005
5. Adding Academic Growth	518	673.61	$df_1 - df_5 = 8$	$G^2_1 - G^2_5 = 53.46$	.0005
6. Adding Major	517	663.29	$df_1 - df_6 = 9$	$G^2_1 - G^2_6 = 63.78$	.0005

Column 1 in Table 6 represents the model under estimation. Columns 2 and 3 display the degrees of freedom and scaled deviance (G^2) for the respective model while columns 4 and 5 represent changes in degrees of freedom and in G^2s between a given model, say Pre-Commitments, and the alternative model (Background only). In logistic regression, it is customary to use the first model (the Background only model in this case) as the *baseline* model or null model when comparing alternative models (see Feinberg, 1983).

As indicated in Table 6, results of the hierarchical inclusion of variables suggest that the variable Major contributed the most to the model's fit followed by Academic Growth and Final Commitments. Results also indicate that the addition of Pre-Graduate Commitments variables, as a group, made a small and nonsignificant contribution to the fit of the model (*p*-value = .25). Comparisons across models need not be merely constrained to the background model as the only reference model. This strategy can be used to compare models among themselves. For instance, results indicate that model 6 gives a more plausible representation of the data than model 5 (observed $X^2 = G^2_5 - G^2_6 = 673.61 - 663.29 = 10.32$; $df = df_5 - df_6 = 518\text{–}517 = 1$; *p*-value = .001). Likewise, results suggest that model 5 constitutes an improvement in relation to model 4 (observed $X^2 = G^2_4 - G^2_5 = 695.77 - 673.61 = 22.16$; $df = df_4 - df_5 = 520 - 518 = 2$; *p*-value = .0005). However, no evidence is found as to the relative improvement of model 4 over model 3 (observed $X^2 = G^2_3 - G^2_4 = 696.12 - 695.77 = .35$; $df = df_3 - df_4 = 522 - 520 = 2$; *p*-value = .8187).

Judging alternative models under the backward elimination test rests on the extent to which deleting variables actually worsens the fit of the model. The test is accomplished by estimating a model in which a group of variables is deleted

TABLE 7. Reduced Model

Factor	Beta	S. E.	Change in P
Gender (Male)	0.3755*	0.1929	0.0930
Pre-Institutional	−0.2918**	0.1138	−0.0723
Pre-Professional	−0.0901	0.1138	−0.0225
Faculty	0.3397**	0.1102	0.0843
Peer	0.1560	0.0967	0.0390
Professional Comit.	0.0711	0.1088	0.0178
Vocational Comit.	−0.0490	0.0791	−0.0122
Conceptualization	−0.1818	0.2033	−0.0453
Research	0.7178**	0.1739	0.1728
Major (Hard Major)	0.6834**	0.1921	0.1651
Intercept	−2.9650	0.7749	
Baseline P_o: .494			

G^2, df = 664.91, 521; X^2, df = 72.53**, 10; G^2/df = 1.276;
pseudo R^2 = 0.120; PCP = 64.85%

*$p < .05$ **$p < .01$

and the G^2 of the reduced or trimmed model is compared against the G^2 for the original model. An inspection of Table 5 indicates that the variable Father's education had no significant effect across the six models. It stands to reason, then, that the exclusion of this variable would not affect the predictive power of the models. In order to test this hypothesis a new model (model 7), in which the variable Father's education was eliminated from model 6, was estimated. Results support this expectation. The reduction of parameters ($df_7 - df_6 = 521 - 517 = 4$) did not significantly worsen the fit of the model (observed $X^2 = G^2_7 - G^2_6 = 664.91 - 663.29 = 1.62$; p-value = .8088). The resulting trimmed model (Model 7) will be kept to continue illustrating the interpretation of the output. Results for Model 7 are displayed in Table 7.

Goodness of Fit

Several indicators are available for assessing the goodness of fit of a given logistic regression model. Some of them are: the pseudo "R^2", the proportion of cases correctly predicted (PCP) by the model, the G^2 / df ratio and The X^2 statistics for overall fit. As in hierarchical modeling, most of these tests rely on the maximum likelihood function associated with a particular model. *The pseudo "R^2"*. An equivalent formulation of R^2 is also available in logistic regression. This indicator is usually referred as the pseudo "R^2". In the OLS context, the coefficient of determination (R^2) has the interesting property of providing an indicator of how well a set of independent variables explains the variance observed in the dependent variable (Draper and Smith, 1981). No equivalent in-

terpretation, however, is available in the logistic regression context. As demonstrated by Christensen (1990) and by Aldrich and Nelson (1986), the pseudo "R^2" represents, at most, the proportion of error variance that an alternative model reduces in relation to a null model. As a standard practice, the null model is usually specified as the one in which all the slopes with the exception of the intercept are set to zero. In GLIM, this model is specified by the following command **$FIT 1 $.** In SPSS-X the maximum likelihood function for the null model is produced by default and it is referred as the "Initial Log Likelihood Function."

Several formulas are available to estimate the pseudo "R^2" (see Aldrich and Nelson, 1986; Christensen, 1990; Maddala, 1987; Mare, 1980; Taylor, 1983). Aldrich and Nelson recommend the following formula:

$$\text{pseudo } R^2 = \frac{X^2}{(N + X^2)} \qquad [3]$$

where X^2 is the chi-square statistic for the overall fit for the model as described below, and N is the total sample size. As is the case of the OLS' R^2, this formula produces values ranging from 0 to 1, and produces a conservative estimate of the reduction in error variance. As shown in Table 5, each successive model accounted for a reduction in unexplained variance. In the case of Model 7 (see Table 7), the corresponding pseudo R^2 indicates that the model accounted for a 12 percent reduction in error variance [$R^2 = 72.53/(532+72.53)$].

Regardless of the method employed in the computation of the pseudo "R^2", the reader is cautioned against relying on this indicator as the sole criterion in judging the validity of a particular model. As noted above, the pseudo "R^2" has no equivalent interpretation to the R^2 in OLS. Moreover, controversy exists as to the most appropriate formula to be employed in the computation of the pseudo "R^2" (see Maddala, 1987; Aldrich and Nelson, 1986). Furthermore, there are no tests to assess the statistical significance of this measure (Maddala, 1987). This controversy is also compounded by the fact that different formulas can yield large or small pseudo R^2s. In view of these problems, Aldrich and Nelson (1984) recommend that different measures of fit be simultaneously taken into account when judging the fit of a particular model, while Fienberg (1983) encourages that the selection of models be also based on the hierarchical testing procedures already described.

Proportion of cases correctly predicted (PCP). Aldrich and Nelson (1983) note that the proportion of cases correctly predicted (PCP) by the logistic model provides an overall indicator of fit much in line with the OLS' R^2. This measure basically involves a comparison between the number of cases that the model predicted as being either 0 (i.e., not engaged in research activities) or 1 (i.e., engaged in research activities) against the total sample size. In GLIM, predicted

probabilities are obtained by the following command: **$DISPLAY R $** (see Table 3). In SPSS-X, the option **PRED** accomplishes the same purpose (see Table 4). Since predicted probability values range from 0 to 1, cut-off scores are employed to identify correctly predicted cases. Subjects are usually identified as correctly predicted by the model whenever the model predicts a probability .5 or greater (see Aldrich and Nelson, 1984). Using this criterion, results indicate that Model 7, for instance, yielded correct predictions for 65 percent of the subjects (see Table 7). By default, SPSS-X produces the classification tables needed in the computation of PCPs.

As is the case in the pseudo "R^2", no known procedures can be found to assess the statistical significance of this indicator of overall fit. Furthermore, the manner in which the cut-off score is set can substantially increase or decrease the estimates of the proportion of cases correctly predicted under the model (see Aldrich and Nelson, 1984). In view of these problems, it would be advisable that judgments on the validity of this indicator be based on PCPs reported by the extant literature applicable to the particular outcome under consideration. For instance, Dey (1991) reported PCPs for college enrollment, dropout and degree attainment obtained from a large national data base. Dey's estimates, then, can assist the educational researcher in judging institutional based models dealing with similar college behaviors. In the absence of such a literature, PCPs associated to alternative models can aid in such an evaluation. As noted in Table 6, the PCPs increased as groups of variables were added. The highest is the one associated with Model 6 (PCP = 65%). Deleting the variable Father's education, as shown in Table 7, did not substantially worsen the proportion of cases correctly predicted (PCP = 65%).

The G^2/df ratio. Similar to chi-square (X^2) tests for LISREL models (see Stage, 1990), the ratio of the G^2 to its degrees of freedom provides an additional indicator of how well the model fits the data (see Cabrera, Stampen and Hansen, 1990; Mallette and Cabrera, 1991; St. John, 1991). The degrees of freedom are given by the following formula: df = sample size−# of parameters fitted. The degrees of freedom for model 7, for instance, are 521 since we are estimating eleven parameters for a sample of 532 PhDs ($df = 532 - 11 = 521$). Since there are no known procedures to test the statistical significance of the G^2/df ratio, rules of thumb such as those suggested by Stage (1990) for LISREL can also be applied to the logistic regression context. Stage (1990) recommends that a particular model be accepted whenever the G^2/df ratio is less than 2.5. Model 7 meets such a criterion ($G^2/df = 1.28$; see Table 7).

The X^2 statistic for overall fit. In OLS, the F ratio for the regression equation is commonly employed to assess the omnibus hypothesis that the independent variables as a group have no effect on the dependent variable (Pedhazur, 1982). The X^2 statistic for overall fit plays the same role in logistic regression. Aldrich and Nelson (1984) recommend the following formula:

$$X^2 = G^2_o - G^2_a \text{ [17]} \quad (4)$$

where G^2_o represents the scaled deviance associated with a model in which only the intercept is fitted (also referred as the null model), and G^2_a represents the scaled deviance for the full model (also referred as the alternative model). In GLIM, the scale deviance for the null model (G^2_o) can be estimated by the following command: **$FIT 1 $.** In SPSS-X, the X^2 statistic for overall fit is produced by default and it is referred as **"Model Chi-Square."** SPSS-X also automatically estimates the degrees of freedom and corresponding significance level for this indicator.

Support for the full model will be evidenced whenever there is a substantial reduction in the X^2. Furthermore, unlike the measures of fit described so far, there are procedures to assess whether the X^2 statistics for overall fit is statistically significant. This test is based on the fact that the difference in G^2s follows a X^2 distribution (see Feinberg, 1984). The degrees of freedom for the X^2 statistics for overall fit are K-1; where K-1 represents the number of coefficients constrained to be zero under the null model (see Aldrich and Nelson, 1984; Collett, 1991). The test, then, involves comparing the computed "X^2" to a critical value drawn from a X^2 distribution with K-1 degrees of freedom at a given significance level. For model 7 (see Table 7), the observed $X^2 = 737.44 - 664.91 = 72.53$ with corresponding $df = 10$ yielded a p-value less than .01 which indicates that the model fits the data.

Testing Individual Regression Coefficients

Individual Coefficient Estimates. Table 7 presents the beta weights and corresponding standard errors for the model under consideration. There are several similarities between OLS and logistic regression concerning the estimation of individual coefficients (Aldrich and Nelson, 1986). As in OLS regression analysis, one of the purposes of logistic regression is to estimate the relationships between a set of independent variables and the dependent variable as well as the statistical significance of such relationships. These relationships, as in OLS, are expressed in terms of beta weights associated with each independent variable. The ML approximation to logistic regression also estimates standard errors for each coefficient which can be employed for testing the null hypothesis that a particular coefficient, say B_{major}, has no effect on the dependent variable. Resembling the test employed in OLS for assessing the significance of individual parameters, the test employed in logistic regression is defined as a t-test. This test

[17]Statistical packages differ in terms of the reported statistics for this joint hypothesis test. Some of them report the equivalent statistics $-2\log(L_o/L_1)$. Aldrich and Nelson (1984) provide a detailed description of alternative ways in which this test is reported.

is obtained by the ratio of the estimated parameter over its standard error (i.e., t-test = B_{major}/SE_{major}). As shown in Column 1 in Table 7, the t-test reveals that only five variables in Model 7 were found to exert a significant effect at a *p*-value < .05.

As in OLS, the sign associated to the beta weights indicates the direction of the effect that a particular independent variable has on the dependent variable. In the case of categorical variables, the interpretation of the coefficients is a function of the excluded category. For instance, the positive sign associated to the beta weight for the variable Hard Major indicates that students majoring in hard sciences are more likely to engage in research activities than those majoring in soft sciences, even after controlling for gender, pre-institutional commitments, interactions with faculty and peers, final commitments and academic growth measures.

The analogy between the logistic regression and OLS concerning beta weights stops at this point. In contrast to OLS, the interpretation of the coefficients is rather troublesome. Unlike OLS, the metric of the individual coefficients under logistic regression is expressed in terms of logits rather than in terms of the original scale of measurement. This problem is particularly accentuated for categorical variables since the corresponding beta weights represent contrasts among categories summarized in terms of differences in logits (see Hanusheck and Jackson, 1977; Freeman, 1987). In other words, Model 7 (see Table 7) predicts that students majoring in hard majors are .68 logit units more likely to conduct research than are students majoring in soft sciences. To overcome the problem of conveying the logistic regression results to the practitioner in a meaningful manner, several methods are available that illustrate the effects of the independent variables on outcome measures.

Interpretation of Results

Two methods are available to illustrate the effect that the predictor variables have on the outcome variable. One method attempts to estimate the overall change a given variable has on the outcome variable. The other attempts to illustrate how responsive the dependent variable is to changes in different independent variables. The former is usually accomplished by measures of overall change, while the latter is usually done via tables and graphs displaying estimated probabilities.

The Delta-p statistics. Petersen (1985) recommends the use of the Delta-*p* statistics[18] as the most suitable method to estimate the overall change in the dependent variable. The Delta-*p* can also be employed for assessing the relative size of these changes across variables provided that the variables are measured in a similar unit. St. John (1991) warns that such comparisons should be approached with care in those cases where the independent variables have a different metric.

[18]St. John and associates (St. John, 1990a, 1990b, St. John et al., 1989) pioneered the use of Delta-*p* to illustrate the results of logistic regression for higher educational outcomes.

Petersen's formula to compute changes in the probability resulting from a unit change in the dependent variable is as follows:

$$\text{Delta-}p = \exp(L_1) / [1 + \exp(L_1)] - P_o \quad (5)$$

where P_o is the sample mean of the dependent variable. As shown in Table 7, the probability of engaging in research for the whole sample is .494 ($P_o = 0.494$). The expression $L_1 = L_o + B$ represents the logit after the unit change in the variable under consideration. L_o represents the natural logarithm of $[P_o /(1-P_o)]$.

For Model 7 (see Table 7), the incremental effect attributable to Major can be illustrated as follows:

$$L_o = \ln [P_o / (1 - P_o)] = \ln [0.494 / (1-.494)] = -0.024$$

$$L_1 = L_o + B_{major} = -0.024 + .6834 = .6594$$

Thus, the logit before the change is $L_o = -0.024$, and after the unit change the logit is $L_1 = .6594$. Using the equation for Delta-*p* in (5) establishes the rate of change in terms of probabilities. In the present example, the Delta-*p* for Major yields the probability value of .1651. As in the case of beta weights for categorical variables, Delta-*p*s are to be interpreted in terms of the excluded category. **In other words, the model predicts that majoring in hard sciences increases the probability of engaging in research activities by 16.5 percentage points over majoring in soft sciences.** In the case of continuous variables, Delta-*p* represents the incremental effect on the outcome variable resulting from a unit change in the dependent variable. **For instance, the model predicts that a unit increase in interactions with faculty increases the probability of engaging in research activities by 8.4 percentage points while a unit increase in research skills increases such a probability by 17.3 percent points.** The last column in Table 7 displays estimates of changes in probabilities in the outcome variable for each independent variable under consideration. Echoing recommendations by St. John (1991), it would be advisable that the use of Delta-*p*s be constrained to those parameters found significant in the model since there are no known procedures to estimate the statistical significance of Delta-*p*s.

Tables and Graphs of Predicted Probabilities. As in OLS, one of the major applications of logistic regression is to predict how likely it is that subjects will engage in the particular outcome under consideration. Predicted probabilities provide the practitioner with basic information about how the variables under consideration can be used to forecast such a behavior and to illustrate how responsive such a behavior is to changes in a given variable. Predicted probabilities, then, can provide the practitioner with basic information when implementing and evaluating intervention strategies. In logistic regression, predicted probabilities for each subject can be obtained by multiplying the values in each

variable by its corresponding beta weights and adding the resulting products. The logistic equation in Table 7 can be displayed as follows:

$$\text{Logit}_{ith} = \frac{P}{1-P} = -2.965 + .3755\text{MALE} - .2918\text{PINT} - 0.0901\text{PPRF} + .3397\text{FACU} + .1560\text{PEER} + .0711\text{PROF} - .049\text{VOCA} - .1818\text{CSKI} + .7178\text{RSKL} + .6834\text{MJR} \quad [6]$$

where Logit_{ith} represents the predicted natural logarithm value or logit for the ith subject. The predicted logit value for the last subject in our example (case number 532 in Table 2), for instance, is .33408. By undoing the arithmetic of the logit transformation under the formula displayed below, this logit value corresponds to a predicted probability for engaging in research of .583[19].

$$P_{532} = \frac{\exp(.33408)}{1 + \exp(.33408)} = .583$$

It is also possible to estimate probabilities attributable to a particular variable that may be of interest to the educational researcher. Consider the example in which the researcher is interested in assessing the effects of interactions with faculty across type of major and gender. Such an assessment can be accomplished by estimating the corresponding logistic regression equations for different types of major and gender while holding the rest of the independent variables constant[20].

The next step is to choose those values for the independent variables that would make it possible to estimate the effects due to changes in interactions with faculty, multiplying this value by the corresponding beta weight, adding these products to obtain logits and estimating the corresponding predicted probability. Manski and Wise (1983), for instance, selected the mean as the value to hold constant the effect of the variables under consideration. Using the same criterion[21], Table 8 displays the estimated probabilities across major and gender. Figures 2 and 3 illustrate the predicted effects of interactions with faculty on the probabilities of engaging in research across males and females while holding constant the rest of the variables at their mean value.

As shown in Table 8 and in Figures 2 and 3, Model 7 predicts that increasing interactions with faculty also raises the likelihood of engaging in research independently of a student's major and gender. As shown in Table 8, a unit increase in the interactions with faculty scale, say from 2.2 to 3.2, increases the likelihood of conducting research between .073 to .085 across gender and major. The reader

[19]In GLIM, the predicted probabilities across all subjects are obtained by the following command: $DISPLAY R $ (see Table 3). In SPSS-X, the option PRED within the LOGISTIC REGRESSION command accomplishes the same purpose (see Table 4).

[20]The logistic regression equations are available upon request.

[21]The means are provided in Table 1.

TABLE 8. Estimated Probabilities of Research Engagement: Effects of Interactions with Faculty

Interactions with Faculty	Hard Major		Soft Major	
	Male	Female	Male	Female
1.00	0.4253	0.3370	0.2720	0.2042
1.20	0.4420	0.3524	0.2857	0.2155
1.40	0.4588	0.3680	0.2997	0.2272
1.60	0.4757	0.3840	0.3142	0.2394
1.80	0.4927	0.4002	0.3290	0.2519
2.00	0.5096	0.4166	0.3442	0.2650
2.20	0.5266	0.4332	0.3596	0.2784
2.40	0.5435	0.4499	0.3755	0.2923
2.60	0.5603	0.4668	0.3915	0.3065
2.80	0.5770	0.4837	0.4078	0.3211
3.00	0.5935	0.5007	0.4243	0.3361
3.20	0.6097	0.5177	0.4410	0.3515
3.40	0.6258	0.5346	0.4578	0.3671
3.47	0.6313	0.5405	0.4637	0.3726
3.60	0.6416	0.5515	0.4747	0.3830
3.80	0.6570	0.5682	0.4917	0.3992
4.00	0.6722	0.5848	0.5086	0.4156
4.20	0.6870	0.6012	0.5256	0.4322
4.40	0.7014	0.6174	0.5425	0.4489
4.60	0.7154	0.6333	0.5593	0.4658
4.80	0.7290	0.6489	0.5760	0.4827
5.00	0.7423	0.6642	0.5925	0.4997

will notice that this rate of change matches closely the one estimated by Delta-*p* (Delta-*p* = .0843 in Table 7).

Rows in Table 8 can be used to estimate the change in probabilities induced by changes in the independent variable, columns can be employed to study the moderating effects of gender and major. For each value of the interactions with faculty scale and within each major, males have higher chances of conducting research than females have; these differences range from .07 to .09. On the other hand, regardless of a student's gender, majoring in hard sciences yields higher probabilities of conducting research as compared with majoring in soft sciences; these probabilities range from .13 to .16. Again, the reader would notice that these differences match closely the estimates produced by Delta-*p*s for gender and major (Delta-*p*s = .093 and .165; respectively. See Table 7).

In sum, it is recommended that the reader adopt a comprehensive approach in judging a particular logistic model. A careful analysis of the statistical significance of the individual parameters coupled with a thorough analysis of the

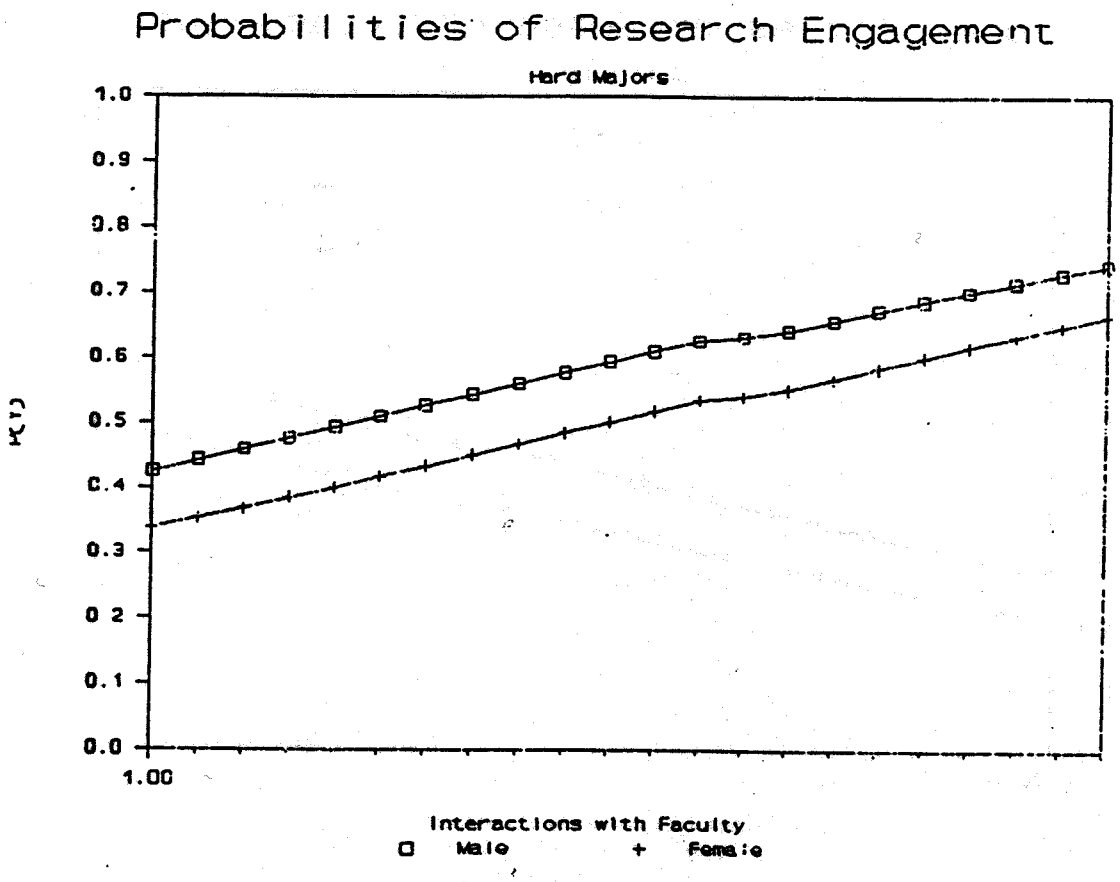

FIGURE 2.

indicators of fit along with an examination of standardized residuals produced by the model should assist the researcher or the practitioner in judging the validity of the model. Hierarchical analyses, individual parameters and the indicators of goodness of fit all suggest that Model 7 is a good representation for the data. This conclusion is further strengthened by the fact that Model 7 yielded only four out of 532 standardized residuals greater than two[22], a pretty low ratio (.75%).

III. CONCLUDING REMARKS

The use of logistic regression in higher education is not without controversy. The debate centers around the use of OLS over logistic regression. Advocates of OLS basically argue that linear regression models: a) are more commonly known, b) the results are easier to explain, c) the method can be used to analyze nonlinear relationships when the variables are properly transformed, and d) it is easier to implement given the increasing availability of statistical programs (Jackson 1980, 1988; Dey, 1991). These arguments are further strengthened by some

[22]These correspond to observations 74, 93, 311 and 386. The complete SPSS-X and GLIM outputs are available upon request.

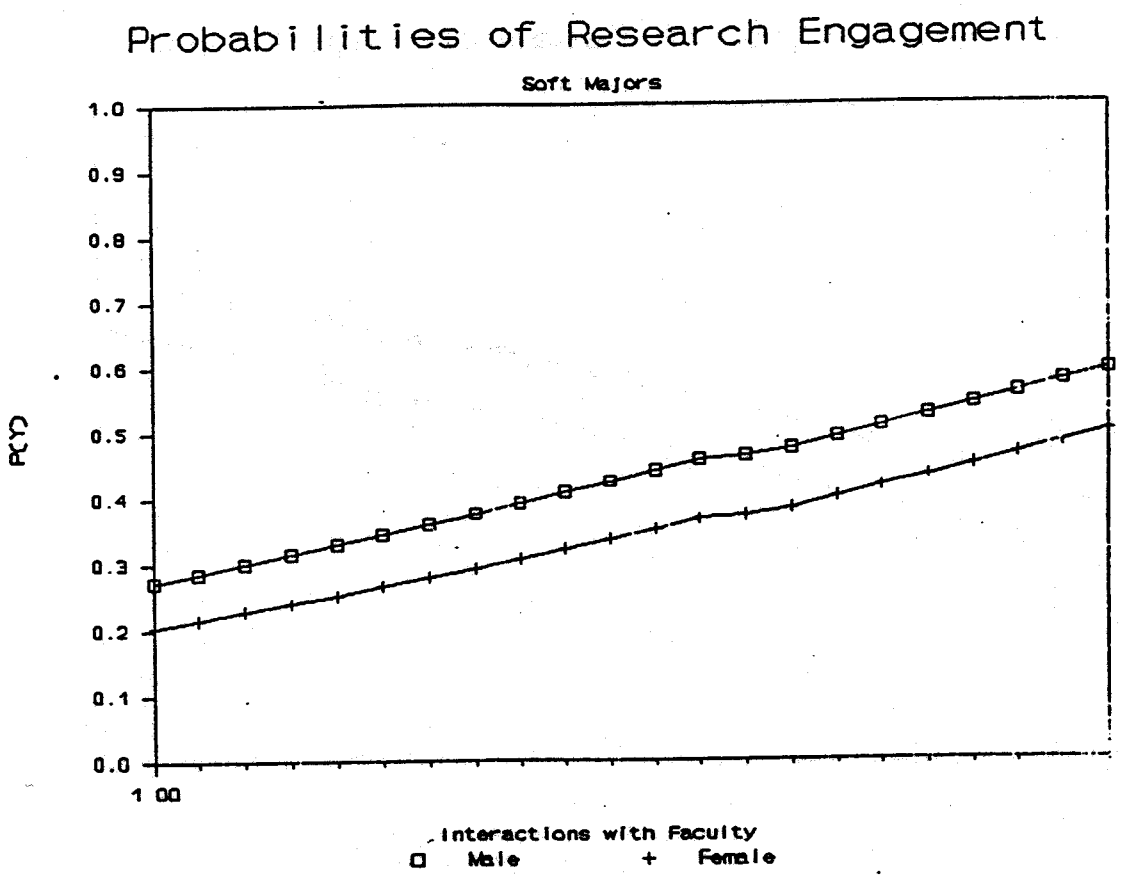

FIGURE 3.

studies finding the application of linear regression models to binary outcomes yielding comparable results to those produced under logistic regression strategies (Jackson, 1981). Dey (1991), for instance, reported that when the dependent variable was moderately distributed, with corresponding splits of .29/.71, .49/.51 and .54/.46, linear regression models replicated the logistic regression's findings concerning the direction of the effect, the ability to predict the statistical significance of the beta weights and the proportion of cases correctly identified as persisting upon graduation for a subsample of the CIRP data base.

On the other hand, advocates of the use of logistic regression argue that the straightforward application of OLS to binary outcomes essentially violates each of the assumptions upon which OLS rests (Hanusheck and Jackson, 1977; Aldrich and Nelson, 1986). Only under very unique circumstances, would the OLS approximation yield results equivalent to those produced under logistic regression (Aldrich and Nelson, 1986; Goodman, 1977; Hanusheck and Jackson, 1977).

At the core of this argument is the different way the two methods approach the functional form underlying the relationship between an outcome variable and an independent variable. Under OLS, the dependent variable is presumed to be

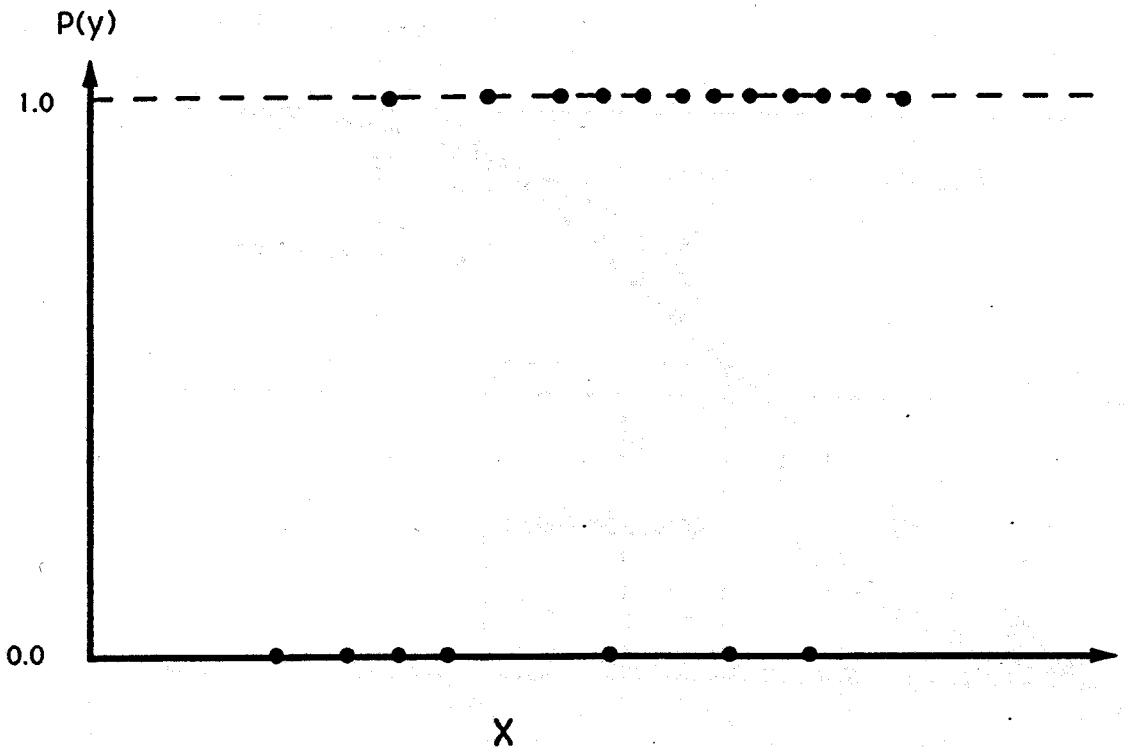

FIGURE 4. Observed distribution for a binary dependent variable.

continuous while the relationship between the outcome variable and, say, an independent variable is supposed to be expressed by a straight line. Neither condition is met when the dependent variable is binary. As shown in Figure 4, the dependent variable is not continuous. Observations lie flat on the X axis depending upon which value the variable assumes (0 or 1 with no other values in between).

Although several transformations can be attempted when the relationship between an independent variable and a dependent variable is not linear, ". . . the dichotomous nature of the dependent variable renders most of these ineffective" (Hanusheck and Jackson, 1977, p. 185). As noted by Hanusheck and Jackson (1977), the logistic approximation is the most appropriate transformation for dichotomous outcomes since empirical studies have shown that the relationship between binary dependent variables and continuous independent variables indeed resembles the already familiar *S* pattern (see Figure 1).

Figure 5 illustrates the effect of using the linear approximation to estimate a logistic distribution. Between x' and x'', the linear approximation would overestimate the true probabilities while the contrary is true for that region corresponding to x' and x'''. Moreover, the linear approximation is likely to yield such nonsensical estimates as negative probabilities when the equation is estimated below x', and probabilities greater than one when the equation is estimated

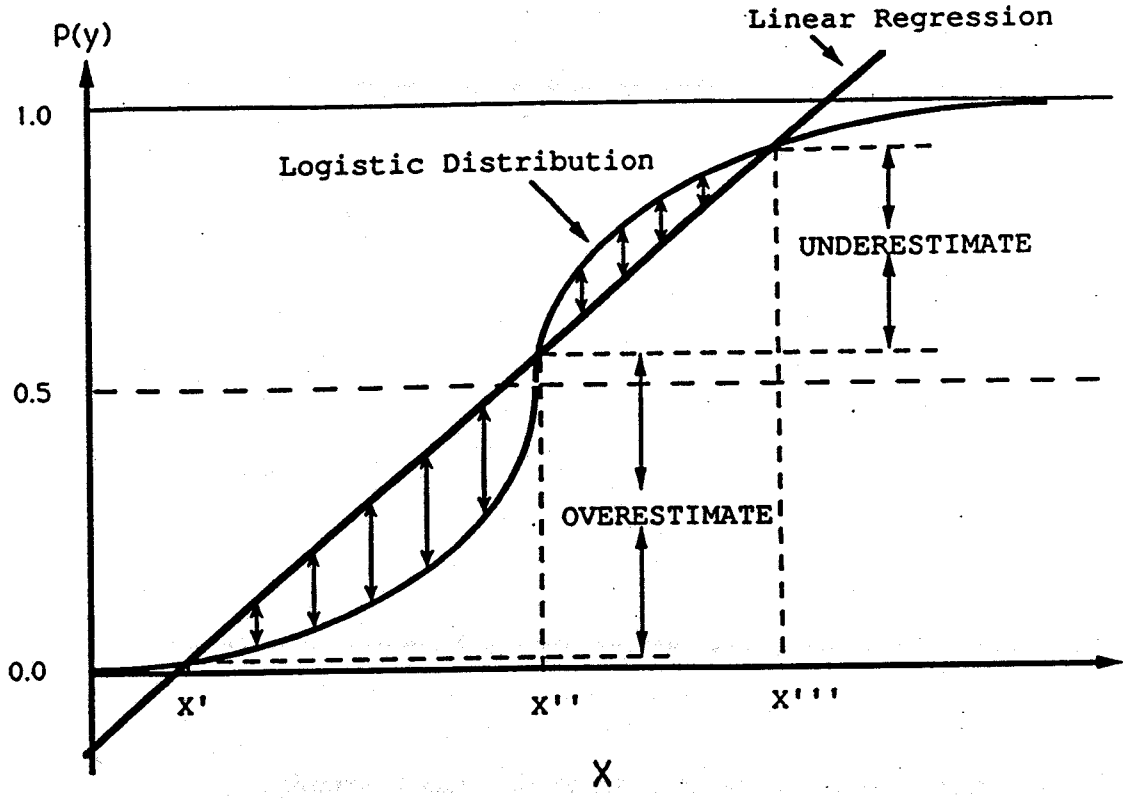

FIGURE 5. Linear approximation to a logistic distribution.

beyond x'''. Only on those points when the regression line intercepts the logistic function would the linear regression produce correct predictions.

Advocates of the logistic regression approach do acknowledge the fact that OLS can approximate the logistic function (Aldrich and Nelson, 1986; Hanusheck and Jackson, 1977). It is also noted that such an approximation may hold under two conditions; and both are sample driven. The first condition is met when the dependent variable is moderately distributed, with splits from as low as .25/.75 (see Goodman, 1977) to .50/.50. As shown in Figure 6, the linear solution would approximate the true relationship in the central part of the distribution. Accordingly, the linear regression approach would correctly identify significant effects while generating predicted probabilities closely resembling the true ones. Only in the extremes of the distribution, would the linear approximation yield incorrect estimates (Hanusheck and Jackson, 1977; Goodman, 1977).

The second condition is met when the sample fails to reproduce the potential range of values in the domain of the independent variable (Aldrich and Nelson, 1986). Figure 7 displays ranges for the independent variable corresponding to three hypothetical samples. As shown in Figure 7, the linear approximation would yield correct estimates in each of the three samples under consideration. This situation is reversed, however, when the linear approximation is applied to

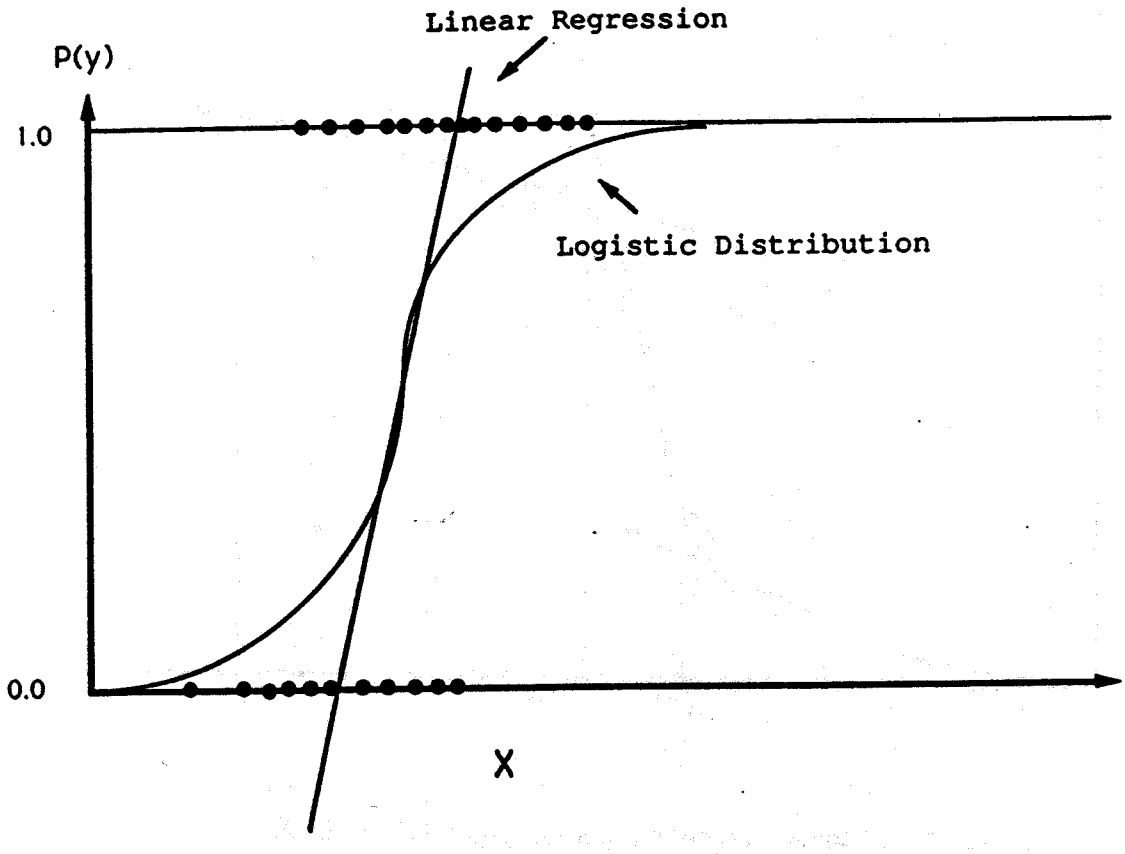

FIGURE 6. Linear approximation to a moderately distributed Y variable.

samples capturing the full range in the independent variable as illustrated in Figure 5.

OLS presumes that the variance around the straight line is constant across all observations in the independent variable (Marascuilo and Serlin, 1988). To the extent that this condition holds, OLS would generate sampling estimates for the beta weights that are both unbiased [in the sense that the sampling beta weights approach the true values in the population] and efficient [in the sense that the sampling beta weights have the smallest sampling variance] (see Pedhazur, 1982). In turn, these efficient sampling variances play a key role when testing hypotheses about the statistical significance of the estimated beta weights (t-tests). When estimating a logistic distribution the error term is all but constant. As noted above, the variance for each predicted Y value changes as a function of the sample size and the value that X happens to assume. Aldrich and Nelson (1986) have demonstrated that the OLS approximation to logistic distributions may produce unbiased sample beta weights whose sign resembles the ones estimated under the logistic regression approximation. Yet, the lack of homoscedasticity in the error term leads to the incorrect estimation of the sampling variances. Consequently, ". . . any hypothesis tests (e.g., the *t* and *F* tests) or

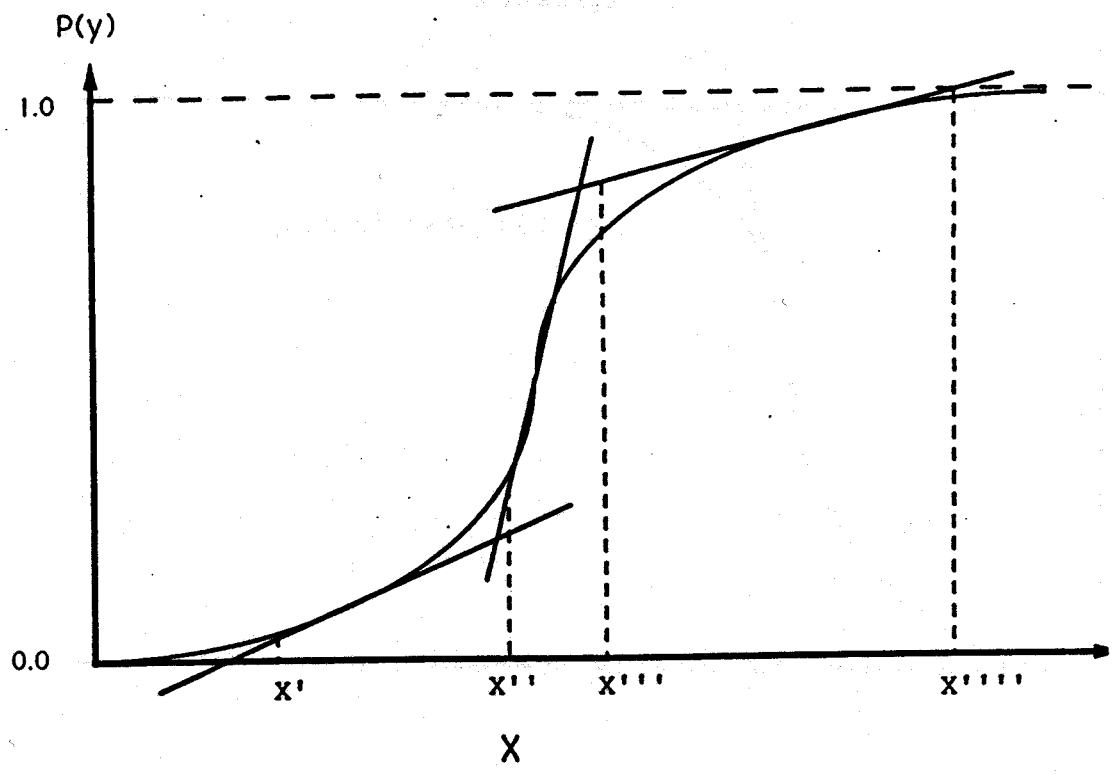

FIGURE 7. Linear approximation for given ranges in X.

confidence intervals based on these sampling variances will be invalid, even for very large samples" (Aldrich and Nelson, 1986, pp. 13–14).

Ordinary least square regression analysis remains a powerful technique. Nevertheless, it is a technique that operates under very stringent assumptions regarding the nature of the dependent variable and the relationship that such a variable has with a given set of independent variables. Unlike OLS, logistic regression conforms with the probability function underlying the relationship between a dichotomous outcome and corresponding independent variables. Factors such as familiarity with the technique and easiness in its use should not dictate the choice of the estimation procedure. Rather, the nature of the phenomenon under consideration should dictate such a choice.

REFERENCES

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1990). *Statistical Modeling in GLIM* (rev. ed). England. Oxford: Clarendon, Press.

Aldrich, J. H., and Nelson, F. D. (1986). *Linear Probability, Logit and Probit Models* (3rd edition). Beverly Hills, CA.: Sage Publications.

Backer, R. J., and Nedler, J. A. (1988). *The Generalised Linear Interactive Modeling* (release 3.77). England, Oxford: Numerical Algorithms Group.

Bentler, P. M. (1989). EQS: *Structural Equations Program Manual.* Los Angeles, CA.: BMDP Software.

Baird, L. L., and Smart, J. C. (1991, November). Graduate students and their academic and professional development: A study of interactions among personal characteristics, life circumstances and graduate school experiences. Paper presented at the annual meeting of the Association for the Study of Higher Education. Boston. Mass.

Biglan, A. (1973a). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology* 57(3): 195–203.

Biglan, A. (1973b). Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology* 57(3): 204–213.

Bishop, J. (1977). The effects of public policies on the demand for higher education. *Journal of Human Resources* 12: 285–307.

Christensen, R. (1990). *Log-linear Models*. New York, NY.: Springer-Verlag.

Cabrera, A. F., Stampen, J. O., and Hansen, W. L. (1990). Exploring the effects of ability to pay on persistence in college. *Review of Higher Education* 13(3): 303–336.

Collett, D. (1991). *Modelling Binary Data*. England, London: Chapman and Hall.

Dey, E. (1991, April). Statistical alternatives for studying student retention: a comparative analysis of logit, probit and linear regression. Paper presented before the 1991 annual meeting of the American Educational Research Association. Chicago, IL.

Ethington, C. A., and Bode, R. (1992, April). Differences in the graduate experience for males and females. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Fienberg, S. E. (1983). *The Analysis of Cross-Classified Categorical Data* (rev. ed.). Cambridge, MA: Massachusetts Institute of Technology.

Freeman, D. H. (1987). *Applied Categorical Data Analysis*. New York: Marcel Dekker.

Goodman, L. A. (1977). The relationship between modified and usual multiple-regression approaches to the analysis of dichotomous variables. In D. R. Heise (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

Hanusheck, E. K., and Jackson, J. E. (1977). *Statistical Methods for Social Scientists*. New York: Academic Press.

Hinkle, D. E., Austin, J. T., and McLaughlin, G. W. (1989). Log-linear models: Applications in higher education research. In J. C. Smart (ed.), *Higher Education: Handbook of Theory and Research*, Vol. V. New York: Agathon Press.

Hossler, D. (1991). *Evaluating Student Recruitment and Retention Programs*. (New Directions for Institutional Research, Vol. 70). San Francisco: Jossey-Bass.

Jackson, G. A. (1980). The case of the dependent dichotomy: practical approaches to college choice models. In *Proceedings of the American Statistical Association*. Washington, DC: American Statistical Association.

Jackson, G. A. (1981). Linear analysis of logistic choices, and vice versa. In *Proceedings of the American Statistical Association*. Washington, DC: American Statistical Association.

Jackson, G. A. (1988). Did college choice change during the seventies? *Economics of Education Review* 7: 15–27.

Jöreskog, K. G., and Sörbom, D. (1989). *LISREL 7*. Mooresville, IN: Scientific Software.

Maddala, G. S. (1987). *Limited-Dependent and Qualitative Variables in Econometrics* (rev. ed.). Cambridge, MA.: Cambridge University Press.

Mallette, B. I., and Cabrera, A. F. (1991). Determinants of withdrawal behavior: an exploratory study. *Research in Higher Education* 32(1): 179–194.

Manski, C. F., and Wise, D. A. (1983). *College Choice in America*. Cambridge, MA.: Harvard University Press.

Marascuilo, L. A., and Levin, J. R. (1983). *Multivariate Statistics in the Social Sciences: A Researcher's Guide*. Monterey, CA.: Brooks/Cole Publishing Co.

Marascuilo, L. A., and Serlin, R. C. (1988). *Statistical Methods for the Social and Behavioral Sciences*. New York: W. H. Freeman and Company.

Mare, R. D. (1980). Social background and school continuation decisions. *Journal of the American Statistical Association* 75(370): 295–305.

Muthén, B. O. (1988). *LISCOMP: Analysis of Linear Structural Equations with a Comprehensive Measurement Model*. Mooresville, IN.: Scientific Software Inc.

Nora, A., Cabrera, A. F., and Shiville, P. (1992). Graduate student involvement in scholarly behavior: a structural model. Paper presented before the 1992 AERA annual meeting. San Francisco.

Norusis, M. J. (1990). *SPSS Advanced Statistics*. Chicago, IL: SPSS Inc.

Pascarella, E. T.,and Terenzini, P. T. (1991). *How College Affects Students*. San Francisco: Jossey-Bass.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*. San Francisco: Holt, Rinehart and Winston.

Petersen, T. (1985). A comment on presenting results from logit and probit models. *American Sociological Review* 50(1): 130–131.

Plane, D. R.,and Oppermann (1977). *Statistics for Management Decisions*. Dallas, TX.: Irwin-Dorsey.

Smart, J. C., Baird, L. L., and Bode, R. (1991, November). Discipline differences in the learning demands of doctoral programs. Paper presented at the annual meeting of the Association for the Study of Higher Education. Boston, MA.

St. John, E. P. (1990a). Price response in enrollment decisions: An analysis of the High School and Beyond Senior Cohort. *Research in Higher Education* 31(2): 161–176.

St. John, E. P. (1990b). Price response in persistence decisions: An analysis of the High School and Beyond Senior Cohort. *Research in Higher Education* 31(4): 387–403.

St. John, E. P. (1991). What really influences minority attendance? Sequential analyses of the High School and Beyond Sophomore Cohort. *Research in Higher Education* 32(2): 141–158.

St. John, E. P., Kirshstein, J. R., and Noell, J. (1991). The effects of financial aid on persistence: A sequential analysis. *Review of Higher Education* 14(3): 383–406.

St. John, E.P., and Noell, J. (1989). The effects of student aid on access to higher education: An analysis of progress with special consideration of minority enrollment. *Research in Higher Education* 30(6): 563–581.

Stage, F. K. (1988). University attrition: Lisrel with logistic regression for the persistence criterion. *Research in Higher Education* 29: 343–357.

Stage, F. K. (1990). LISREL: An introduction and applications in higher education. In J. C. Smart (ed.), *Higher Education: Handbook of Theory and Research*, Vol. VI. New York: Agathon Press.

Stampen, J. O., and Cabrera, A. F. (1988). The targeting and packaging of student aid and its effect on attrition. *Economics of Education Review* 7(1): 29–46.

Stampen, J. O. and Cabrera, A. F. (1986). Exploring the effects of student aid on attrition. *Journal of Student Financial Aid* 16(2): 28–40.

Taylor, D. G. (1983). Analyzing qualitative data. In P. H. Rossi, J. D. Wright, and A. B. Anderson (eds.), *Handbook of Survey Research*. San Diego: Academic Press.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research* 45: 89–125.

Tinto, V. (1987). Leaving college: *Rethinking the Causes and* Cures of Student Attrition: Chicago: University of Chicago Press.

Weiler, W. C. (1987). An application of the nested multinomial logit model to enrollment choice behavior. *Research in Higher Education* 27(3): 273–282.