Chapter 7
Categorical and Limited Dependent Variable Modeling in Higher Education

Awilda Rodriguez, Fernando Furquim, and Stephen L. DesJardins

7.1 Introduction

Often the types of outcomes that higher education researchers examine are represented by categorical variables. These may include dichotomous or binary dependent variables, such as whether a student enrolls in college or not, whether they persist to their sophomore year (or not), or whether they graduate. In addition to studying binary representations of underlying constructs, we are often interested in studying outcomes that are multi-categorical, also referred to as polytomous. These might include outcomes that have some natural ordering (i.e., are ordinal) or those that are not ordered but have multiple nominal categories (i.e., are "multinomial"). Examples of ordered outcomes include survey questions evaluating teaching with response categories of excellent, good, fair, and poor or a Likert scale of agreement where the categories include strongly agree, agree, neutral, disagree, and strongly disagree. In terms of multinomial responses, where no order in the relationships among the categories is evident, examples include a person's college choice (e.g., no college, attend least selective, selective, or most selective college) or college major choice (e.g., liberal arts, engineering, science, business, other).

In addition to there being binary and multi-categorical outcomes, there are also other types of outcomes that require specialized estimation techniques. These include variables where the range of values for the outcome are restricted due to

A. Rodriguez (✉) · F. Furquim
Center for the Study of Higher and Postsecondary Education, School of Education, University of Michigan, Ann Arbor, MI, USA
e-mail: awilda@umich.edu; ffurquim@umich.edu

S. L. DesJardins
Center for the Study of Higher and Postsecondary Education, School of Education, Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI, USA
e-mail: sdesj@umich.edu

M. B. Paulsen (ed.), *Higher Education: Handbook of Theory and Research*,
Higher Education: Handbook of Theory and Research 33,
https://doi.org/10.1007/978-3-319-72490-4_7

censoring or truncation (known as limited dependent variables) and outcomes that are measured as counts or proportions. Examples of limited dependent variables include outcome values with ranges that are censored, such as income obtained from a survey with the top end of its distribution censored at some value (e.g., $150,000 and above). An example of truncation is when we examine the effects of a developmental program where students are placed into the program based on a placement test, but we only have observations for individuals whose test score was below some threshold. Examples of a count outcome are the number of applications a student sent to colleges, the number of AP classes a student took, or the number of students receiving Pell grants in a college. In terms of proportions, examples include the percentage of students in a college who are from underrepresented groups, the proportion persisting from the freshman to sophomore year, or what fraction in the public vs. for-profit sector default on their loans. The higher education literature is replete with studies examining binary and polytomous dependent variables, and to a lesser extent studies of limited, count, and dependent variables that are proportions.

When faced with estimating regression models with categorical or specialized dependent variables, researchers often simply employ linear (ordinary least squares, or OLS) regression. However, there are some well-known statistical and practical problems in doing so, including violations of important underlying assumptions when the dependent variable is not continuous, and problems (e.g., bias and/or inefficiency) with the estimates produced when using such an approach. Given these potential problems, knowing more about how to adequately model outcomes that are categorical or limited in some way is important. In an earlier edition of this Handbook, Cabrera (1994) provided a description of how to employ statistical models designed to deal with categorical dependent variable models, thereby providing higher education researchers with a grounding in these approaches. However, since Cabrera's (1994) chapter there have been important changes in the application of statistical methods to the study of categorical dependent variables. These include advances in the underlying statistical aspects of estimating such models, including an improved understanding about the strengths and weaknesses of some of the formal tests often used. There are also many new software packages available to estimate these models, with features that make estimation easier and improve our ability to interpret the results through tabular and graphical displays. Categorical dependent variable models are also widely used in software packages used to estimate some quasi-experimental models (e.g., propensity score matching; instrumental variable regression) now often employed for causal inferences. In addition, Cabrera's (1994) chapter focused almost exclusively on binary categorical dependent variable models. Given the ubiquity of the use of categorical dependent variables in higher education research, and advances in the application of these models, this chapter will build on Cabrera's (and others) work by (1) providing some of the conceptual and statistical underpinnings and rationale for the use of categorical and limited dependent variable regression models, (2) demonstrate how to estimate some of these models using a running example of a higher education issue, (3) provide examples of extensions of these models, and (4) to promote the use of the methods, point readers to additional literature and (in the appendix) provide the statistical code (in Stata) used to produce the results from our running example.

In the next section, we introduce the empirical example we will use for much of the chapter. We use the study of student college choice because (1) it is an important issue in postsecondary education; (2) the topic and underlying mechanisms should be well-known to many Handbook readers, permitting them to focus on the statistical content; and (3) we have access to very current, national data not yet extensively used to study student choice. After introducing the running example and data, we focus our discussion on binary outcome models, then move on to a discussion of estimating multi-categorical outcomes, and finish the chapter with other limited dependent variables, including an example of modeling count outcomes and brief discussions on modeling proportional, censored, and truncated outcomes. Throughout the chapter, we insert in the text the Stata commands we have used for analysis, highlighting them in a different font. We also include much of the statistical code used to conduct the analysis presented herein in the appendix.

7.1.1 Studying Categorical Outcomes in Higher Education

Student college choice is one of the most studied phenomena in higher education research. In the context of changing landscape in college preparation, increased competition for admission, and concerns about college affordability, college choice remains an active area of inquiry. Many scholars pay particular attention to the ways in which student characteristics (e.g., academic performance, family background, prior schooling) are associated with college application and enrollment behavior–especially in the context of enduring social stratification in postsecondary education. Previous quantitative research into college choice has studied whether students apply to or enroll in college (Bielby, Posselt, Jaquette, & Bastedo, 2014; Roderick, Coca, & Nagaoka, 2011; Kim, DesJardins, & McCall, 2009); where students enroll (e.g., by institutional sector or selectivity, Belasco, 2013; Chung, 2012; O’Connor, Hammack, & Scott, 2010; Perna & Titus, 2004; Posselt, Jaquette, Bielby, & Bastedo, 2012; Taggart & Crisp, 2011); how many college applications high school seniors submitted (Long, 2004); as well as the college-going rate of high schools (Engberg & Gilbert, 2014). All such outcomes are measured as categorical or limited dependent variables, and researchers frequently employ nonlinear regression techniques to study them. We therefore use various operationalizations of college choice outcomes throughout this chapter to illustrate regression techniques that are often employed to estimate models with these types of dependent variables.

7.1.2 Data and Sample

All analyses in this chapter make use of data from the High School Longitudinal Survey of 2009 (HSLS:2009). The National Center for Education Statistics (NCES) surveyed over 23,000 high school 9th grade students in 944 high schools in 2009, with follow-up surveys in 2012 as well as surveys of parents and school personnel. HSLS:2009 includes information about students' backgrounds, academic performance, course transcripts, college expectations, college applications, and high school environment.[1] We limited the data to high school graduates and excluded observations missing key measures, resulting in 10,940 students. Our choice to not account for missing data is based on our goal to focus on the modeling the various categorical and limited outcomes, and a concern about how much space it would take to include a detailed discussion of how to deal with missingness. A robust literature on missingness and imputation methods is available (Allison, 2002; Little & Rubin, 2014).

7.1.3 Variables

Dependent Variables In order to demonstrate the application of the methods used to study categorical and limited dependent variables, we used the HSLS data to construct three different outcome variables. To demonstrate how to model binary outcomes, we created a dichotomous variable measuring whether students enrolled in college after completing high school or not (discussed in Section II). To demonstrate the modeling of polytomous dependent variables, we created a multi-categorical measure that disaggregates whether the student enrolled in college into finer grains based on the selectivity of the institution attended. This dependent variable has four categories: no college, chose a less selective, selective, or most selective institution (see Sections III and IV). The third outcome we modeled is students' self-reported number of college application submitted, which we use to demonstrate the utility of count regression techniques (presented in Section V).

Independent Variables In the regressions estimated, we control for constructs thought to affect whether a student goes to college, and the type of institution they decide to attend. These constructs were chosen based on theories used to explain the college choice process and were operationalized using variables included in prior studies and available in the HSLS data set.

Academic Ability Given its strong sorting function in the provision of college opportunity and specifically in the college admissions process, academic ability is arguably the most important construct included in inferential studies of college

[1] Although we utilized a restricted version of the HSLS data, there is also a publicly available version (see https://nces.ed.gov/edat/).

choice (Clinedinst, Koranteng & Nicola, 2015). When modeling college choice, researchers often include prior academic achievement measures such as students' high school grade point average (e.g., GPA, a local measure of achievement); exam scores, which are often state- or nationally-normed measures (Engberg & Allen, 2011; Posselt et al., 2012); as well as academic *preparation* measures such as the highest-math course completed (Kim, Kim, DesJardins & McCall, 2015) or the number of college-preparatory courses taken in high school (Engberg & Wolniak, 2010). To operationalize prior academic achievement and preparation in the models estimated, we include 10th grade GPA; students' scores on the math exam administered by NCES; the highest level of math taken by 12th grade; and the total number of AP course credits students acquired during high school.

Demographic Characteristics Given the (1) historic exclusion of non-White, female, and low-income students from many forms of higher education; (2) persistent differences in high school resources across race and income (Office for Civil Rights [OCR], , 2016; Palardy, 2015); and (3) presence of Minority-Serving Institutions (MSIs) that shape choice (Freeman & Thomas, 2008; Teranishi & Briscoe, 2008), we follow most previous college choice studies and include race/ethnicity and income as explanatory variables in the models estimated. Gender is also another important characteristic to consider when studying college choice, as women are generally more likely to enroll in college but less likely to do so at selective institutions (Bielby et al., 2014). We also include a measure of parental education as parents who attended college are typically more able to assist their children with the college choice process, and because some scholars argue that students whose parents did not attend college rely on their high schools to help them navigate the complex college choice process (Ceja, 2001; Perna & Titus, 2005; Rowan-Kenyon, Bell & Perna, 2008).

College Expectations A number of college choice studies also control for students' stated college plans or aspirations (Gonzales, 2011; Posselt et al., 2012). In the student surveys, NCES asks students how much postsecondary education they intended to acquire, which we included and coded as: high school or less, some college, two-year degree, four-year degree or more. Another important measure that shapes college choice and is frequently included in choice models is peers' college enrollment plans (Engberg & Wolniak, 2010; Taggart & Crisp, 2011).

School Characteristics Many researchers include school-level measures in their college choice models to reflect that students are nested within schools, and that schools are an important context for students. High schools provide resources and present college-going norms that, in turn, shape individual student choice (McDonough, 1997; Perna, 2006; Roderick et al., 2011). But a school's college-going norms are challenging to measure. In this study, we use the share of students enrolled in two- and four-year colleges as proxies for college-going norms. Existing studies have also controlled for high school characteristics to acknowledge differences in demographic composition (representation by race or income); operating status such as charter and/or magnet schools; and urbanicity. We include these measures as well. Table 7.1 describes the dependent and independent measures discussed above and used in the applications of the modeling techniques demonstrated in this chapter.

Table 7.1 Description of variables

Dependent variables	Proportion/ Mean	S. D.	Description
$N = 10{,}940$			
College enrollment			Postsecondary institution attending as of Nov 1, 2013.
No college	33.7		
College enrollment	66.3		
Enrollment by selectivity			Enrolled college IPEDS selectivity code, as found in 2012 IPEDS institutional characteristics file
No college	34.0		
Less selective college	27.2		
Selective college	21.6		
Most selective college	17.2		
Number of applications	2.7	2.8	Self-reported.
Independent variables			
Demographics			
Race/ethnicity			Collected from the student questionnaire, school roster, or parent questionnaire, in order of preference.
Native American	1.0		
Asian	7.9		
Black	8.0		
Latino	14.2		
Multiracial	8.7		
White	60.2		
Income			Total family income from all sources 2008.
<35 K	23.5		
35–55 K	16.7		
55–75 K	13.9		
75–95 K	12.1		
95–115 K	9.2		
115 K and above	24.6		
Parental education			Highest level of education, taken from the base year parent questionnaire.
HS diploma or less	32.1		

(continued)

Table 7.1 (continued)

Dependent variables	Proportion/ Mean	S. D.	Description
Associate's or certificate	4.2		
Bachelor's or more	63.7		
Academics			
GPA, 10th grade	2.7	0.9	Ranges between 0 and 4.
Math test scores	42.2	11.6	Ranges between 16 and 70.
AP credits	1.3	2.2	Ranges between 0 and 16.
Highest math			Highest level mathematics course taken/pipeline in the 12th grade; drawn from transcript files.
Algebra I or below	3.4		
Algebra II/geometry	27.9		
Precalculus/ advanced	47.7		
Calculus or above	20.9		
Expectations			
Friends' PSE expectations	93.0		9th grader's closest friend plans to go to college.
Students' PSE expectations	91.4		Expect AA/BA as of senior year.
School controls			
Pct. 4-Yr college enrollment	54.7	26.4	Ranges between 0 to 100.
Pct. 2-Yr college enrollment	24.4	16.4	Ranges between 0 to 100.
Urbanicity			Characterizes the sample member's base year school from the common Core of data (CCD) 2005–06 and the private school survey (PSS) 2005–06.
Urban	28.0		
Suburban	35.4		
Town	12.8		
Rural	23.9		
School type			Drawn from school survey; special program school [or magnet school] includes a science or math school, performing arts school, talented or gifted school, or a foreign language immersion school.
Regular	93.2		
Charter	2.0		
Special program	2.9		
Career/ vocational	1.9		

Source: HSLS:2009

7.2 Binary Outcomes

There are three prevalent approaches to modeling binary outcomes–logistic, probit, and linear regression (i.e., linear probability models). Several texts discuss binary outcomes at length (e.g., Hosmer, Lemeshow, & Sturdivant, 2013; Long, 1997; Long & Freese, 2014; Menard, 2002; Pampel, 2000). Below, we situate binary outcomes in the higher education context and highlight post-estimation techniques that aid in the interpretation of the findings. We start with a discussion of some important statistical concepts—odds, odds ratios, probabilities, risk ratios, and relative risk ratios—as these measures serve as an important foundation for modeling binary and multinomial outcomes.

7.2.1 Odds, Odds Ratios, Probabilities, and Risk Ratios

Before moving into an explanation of binary regression techniques, first we formally define distinct ways of summarizing categorical outcomes that are, at times, conflated in common language usage— odds, odds ratios, and probabilities. We also formally present the risk ratio and relative risk ratios—measures that are essential in understating the estimation of multinomial models in Section IV.

The odds of an event occurring is the quotient of two probabilities: the probability the event will occur (*Pr(y = 1)*) divided by the probability that it will not occur (*Pr (y = 0)*), which takes the form:

$$Odds(y=1)=\frac{Pr(y=1)}{Pr(y=0)}=\frac{Pr(y=1)}{1-Pr(y=1)} \tag{7.1}$$

Odds have a lower bound of zero and upper bound of +∞. An event with a break-even probability of occurring (e.g., 0.50) has odds equal to 1. In our running example, the *probability* of college enrollment for the overall HSLS:09 sample is 0.52 (Table 7.2).[2] The *odds* of four-year enrollment in our sample of high school seniors in 2013 is therefore 1.08 (or, 0.52/[1–0.52]).

An odds *ratio* allows for comparisons of the odds of an event occurring between two groups as a quotient—the odds of the event ($y = 1$) given an additional condition ($x = 1$) divided by the odds of the event given another condition ($x = 0$). The odds ratio is defined below as:

$$Odds\ Ratio(y=1|x=1)=\frac{Odds(y=1|x=1)}{Odds(y=1|x=0)} \tag{7.2}$$

[2] Defined as two- or four-year college enrollment as of November of 2013.

Table 7.2 Comparison of probability, odds, and odds ratios for college enrollment[a] by gender, 2013

	(1)	(2)	(3)	(4)
	Probability	Odds	Odds ratio	Risk ratio
Total[b]	0.52	1.08	–	–
Gender				
Female	0.55	1.24	1.41	1.17
Male	0.47	0.88	Ref.	Ref.
Race				
Native American	0.39	0.65	0.51	0.70
Asian	0.63	1.70	1.32	1.12
Black	0.46	0.85	0.66	0.82
Latino	0.44	0.77	0.60	0.78
Multiracial	0.51	1.05	0.82	0.91
White	0.56	1.28	Ref.	Ref.

Sources: HSLS:2009

Notes: (a) College enrollment includes two- and four-year colleges; (b) sample includes all students with base year, follow-up data ($N = 25{,}210$)

Odds ratios have a lower bound of 0 and upper bound of $+\infty$. Using our HSLS sample, the odds ratio of four-year college enrollment for women relative to men equals the odds of four-year enrollment when female = 1, divided by the odds of enrollment for men (i.e., female = 0). In our sample, the odds of college enrollment for women is 1.24 and the odds for men is 0.88, yielding an odds ratio of 1.41 (1.24/0.88, see Table 7.2, column 3). In other words, the odds of women enrolling in college are 1.41 times those of men, or 41% greater odds (we subtracted 1 from the odds ratio to arrive at 41%).

With some algebraic rearranging of Eq. 7.1, the probability can be defined in terms of odds as:

$$Pr(y = 1) = \frac{Odds(y = 1)}{1 + Odds(y = 1)} \tag{7.3}$$

However, unlike odds and odds ratios, probabilities are bounded by zero and one. Continuing with our running example, the predicted probability of enrollment, given the student is female is [1.24/(1 + 1.24)] = 0.55 and for males the predicted probability is [0.88 / (1 + 0.88)] = 0.47.

The risk ratio (also sometimes called the relative risk) is the ratio of two probabilities—the probability of outcome y occurring under condition $x = 1$ divided by the probability of outcome y occurring under another (base) condition $x = 0$ (Eq. 7.4).

$$Risk\ Ratio\ (y = 1|x = 1) = \frac{Pr(y = 1|x = 1)}{Pr(y = 1|x = 0)} \tag{7.4}$$

For many, the term "risk" connotes a negative event, as its use is historically rooted in the health fields (e.g., a patient's "risk" of an adverse health event). In our example where y is college enrollment and x is gender, we divide the aforementioned probability of enrolling ($y = 1$) for females (where $x = 1$) of 0.55 by 0.47, which is the probability of enrolling for males (where $x = 0$) resulting in 1.17. This number represents the risk of women enrolling in college, relative to men. We can interpret this ratio as indicating that women's risk of enrolling in college is about 1.17 times that of men. Note that the calculation of the risk ratio is different than the odds ratio, with the former being the ratio of two probabilities (Eq. 7.4), and the latter being the ratio of two odds (Eq. 7.3). When the event occurrence (e.g., enrollment) is small ($<$10%), the odds- and risk-ratios will be similar. But these two measures diverge as the event becomes more common. Also, the relationship between ORs and RRs depends on the direction of the relationship between the outcome and regressor. When there is no association between the outcome and regressor OR = RR. When there is a negative (positive) relationship OR $<$ RR (OR $>$ RR). Thus, using these two terms interchangeably depends on the context.

The *relative* risk ratio relates the risk ratios for two possible outcome categories, for example, outcome m relative to a baseline outcome b out of J possible outcomes.

$$Relative\ Risk\ Ratio\ (y = m|x) = \frac{Pr(y = m|x)}{Pr(y = b|x)} \tag{7.5}$$

The interpretation of the relative risk ratio is always in relation to a base outcome, which is important to note when you have multiple outcome categories, so we will return to this topic in section IV.

Next, we discuss the three main regression-based approaches to estimate binary outcome models –the logit, probit, and the linear probability models. We begin with a formal presentation of the logit model and use it to frame our discussions of goodness of fit and interpretation of coefficients—much of which is applicable to the probit model. Throughout, we note the estimation and post-estimation commands available in the Stata software package that one can employ to estimate models and after the regressions are estimated, to facilitate the interpretation of results. Next, we turn to a discussion of the probit model, underscoring the points where it diverges from logit regression. The explanation of the probit model is followed by a presentation of the linear probability model, where we consider the conditions under which it might not be appropriate to use when modeling binary dependent variables. We close this section with a summary of the pros and cons of the three binary modeling techniques.

7.2.2 Logistic Regression

The logit model is commonly used in education studies to model the relationship between a set of predictors and a binary outcome. The outcome of interest takes on only two values, typically represented in the data by a 1, indicting the event of interest (e.g., enrollment in college), and 0, indicating the event did not happen (e.g., non-enrollment in college). For statistical reasons, and to ease in the estimation of such a model, we would like this binary dependent variable to be linear in the parameters. For this to be the case, the dependent variable is transformed into a continuous measure that ranges from –∞ to +∞. Conceptually, we can think of our observable binary outcome of interest (denoted by y) representing an unobserved latent construct (y* representing, for example, the underlying propensity to enroll in college), that ranges from –∞ to +∞. Higher values of y* are associated with the observable binary outcome $y = 1$, and lower values of y* are associated with $y = 0$. We can relate observed measures (x's) with the continuous latent y^* formally using:

$$y^* = X'\beta + \varepsilon \tag{7.6}$$

To illustrate, while we only actually observe whether students enroll in college (or not), individuals have some underlying unobserved probability (or propensity) to enroll. Some individuals are very likely to enroll in college (i.e., have higher values of y*) while others are very unlikely to enroll (have lower values of y*). Another set of individuals are somewhere in the middle, whereby they might enroll if the conditions are right (e.g., a conversation with a mentor, a subway ad, or a campus visit). There are a whole host of reasons why some students have high probabilities of enrolling in college and others do not. Potential x's for Eq. 7.6 may, for example, include a student's academic performance in high school, their family income, or peer influences. Given this unobserved probability to enroll, imagine there is also an unobserved threshold (τ) that separates those who attend from those who do not. Formally this can be represented as:

$$y = \begin{cases} 1 \; if \; y^* > \tau \\ 0 \; if \; y^* \leq \tau \end{cases} \tag{7.7}$$

where y is what we observe in the data. The task at hand, then, is to transform the observed binary y into a continuous measure that ranges from –∞ to +∞ in order to model the unobserved or latent tendency (y^*) to enroll.

Mathematically, we perform several steps to transform a binary measure into a continuous measure that ranges from –∞ to +∞. First, we transform the outcome into the probability of the event occurring because it allows us to conceptualize the outcome in a continuous form. We then take the (natural) log of the ratio of the probability of the event occurring or not. Known as the "logit," this variable is bounded by –∞ to +∞ allowing this outcome measure to be linearly related to the parameters. Taking the natural log of Eq. 7.1 above, and conditioning on a set of covariates $X^{'}$, the logit model can be formally defined as:

$$\ln\left(\frac{Pr(y=1|x)}{1-Pr(y=1|x)}\right)=X'\beta+\varepsilon \tag{7.8}$$

eliminating subscripts for ease of expression, the left-hand side of the equal sign is the natural log of the odds of an event occurring; with intercept α; a vector of covariates x with a corresponding vector of coefficients β; and errors ε. The logit model is typically estimated using maximum likelihood estimation (MLE), an iterative technique that estimates parameters sequentially until the likelihood that the estimates produced best fits the underlying data is maximized. This method is different from ordinary least squares (OLS) regression, which identifies parameters that minimize the sum of squared residuals. Several texts provide thorough overviews of maximum likelihood estimation (Eliason, 1993; Wooldridge, 2002). For our purposes, it is sufficient to keep in mind that estimates produced from the likelihood function are consistent, asymptotically normal, and asymptotically efficient (Long, 1997). However, given maximum likelihood's asymptotic properties, the logit model is not well-suited for small samples.[3] In fact, this caution holds for all of the regression techniques discussed in the chapter – when employed using small samples, their foundational assumptions may not hold, yielding potentially inconsistent estimates.

To identify the logit model, we need to make a number of assumptions. First, unlike a linear regression model—which assumes errors are normally distributed—the logit model assumes a distribution of errors that are logistically distributed ($\sigma = \pi^2/3$) with a mean of zero (the *zero conditional mean of* ε assumption). Since the error distributions from binary data are not directly observed, the variance is set to $\pi^2/3$ because the probability density and cumulative distribution functions are simpler to ascertain when using this value. When plotted, the probability density function for the logistic distribution has thicker tails than the normal distribution (see Fig. 7.1). As a result, the cumulative logistic distribution increases at a faster rate than the normal distribution. With a defined distribution for the errors, we can then estimate Pr(y = 1). Also, the right-hand side of Eq. 7.8 indicates that using the logit functional form forces a linear relationship between the outcome (the natural log of the odds) and the model parameters. Thus, the model is linear in the logit, or log-odds, but *not* linear in the probability. A third assumption of the logit model is that the included regressors cannot be a linear combination of each other (*no multicollinearity*, Menard, 2010).

Violations of these assumptions could lead to inefficient and biased estimates, making it difficult to establish the true effect of regressors on the dependent variable. Relatedly, the nature of the data and the covariates—particularly in small sample sizes with categorical predictors—can undermine model estimation due to separation

[3] For studies with few observations (e.g., fewer than 100), use exact logistic (Mehta & Patel, 1995).

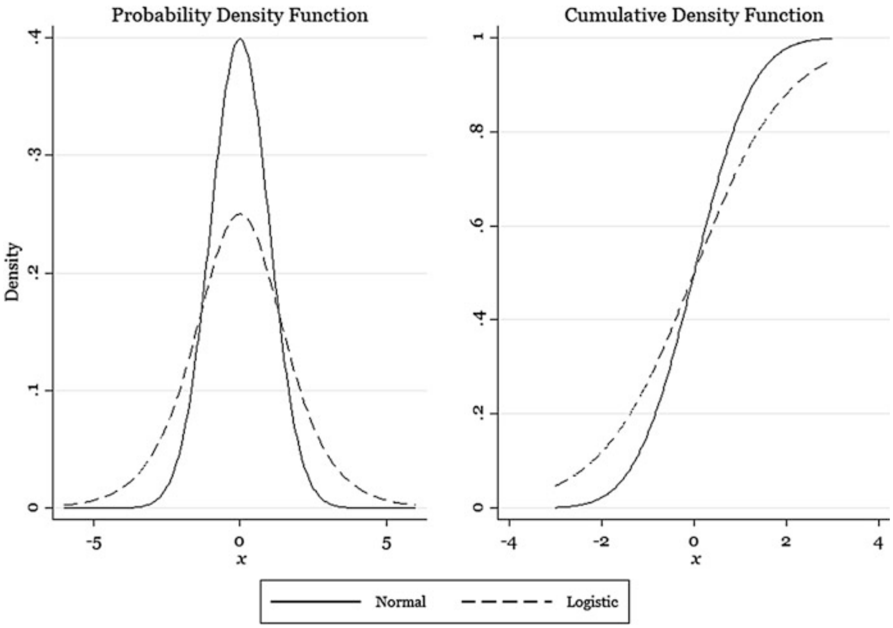

Fig. 7.1 Probability & cumulative density plots for normal and logit distributions

and empty cells. Perfect separation occurs when there is no variation for an independent variable across the dependent variable. For example, if every student who took calculus enrolled in college, highest math would perfectly (or near-perfectly) predict college enrollment. Maximum likelihood estimation procedures will tend not to work under such conditions. Another consideration is empty cells, which are a result of insufficient observations in a particular category of an independent categorical variable. This can be an issue for categories that are traditionally underpowered; such as the multi-racial category in race/ethnicity or for inferences into the intersection of categorical variables (e.g., low-income students who have taken calculus). See Menard (2010) for a detailed discussion of violations of these assumptions and how to address them.

Example: Modeling College Enrollment To demonstrate the use and interpretation of binary outcome models, we estimated college enrollment as our outcome of interest. We first estimated an unconditional (or restricted) model, that is, a regression with no covariates (an intercept only model) to compare with our manual calculations above. Using this model, we found the odds ratio for college enrollment is 1.04 (the same as the odds ratio we calculated by hand in Table 2). The full

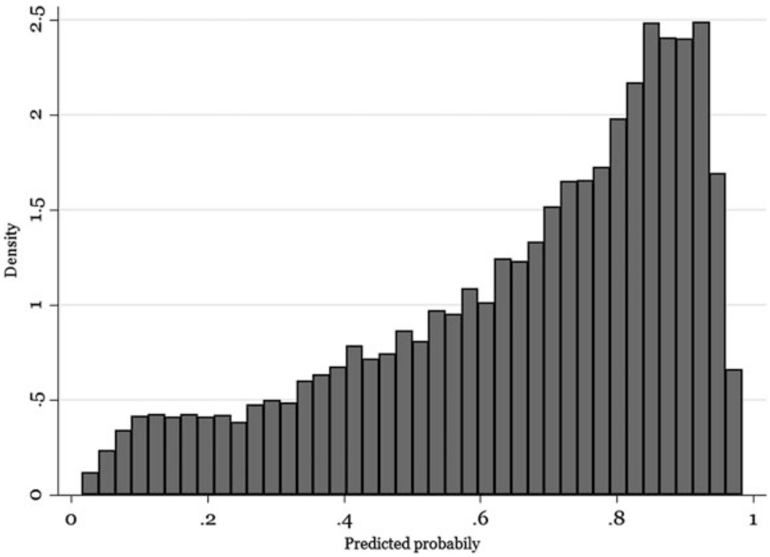

Fig. 7.2 Density plot of predicted probabilities of college enrollment, full model (Source: HSLS:2009)

(or unrestricted) model includes covariates that we hypothesized to explain four-year college enrollment:

$$Pr(Enroll = 1) = \beta_0 + \boldsymbol{\beta_1} DEMS + \boldsymbol{\beta_2} ACAD + \boldsymbol{\beta_3} EXPECT + \boldsymbol{\beta_4} SCHOOL \quad (7.9)$$

that includes student demographics (*DEMS*, gender, race/ethnicity, family income, parental education); academics (*ACAD,* high school GPA, test scores, number of rigorous courses in high school, highest math course completed in high school); the students and their friends' college-going expectations (*EXPECT*); and a number of high school controls (*SCHOOL*).[4] We visually checked the distribution of predicted probabilities using Stata's *predict* and *histogram* commands to get a general sense of the data (see the accompanying appendix for the presentation of the Stata code used in the chapter). Figure 7.2 indicates a left-skewed distribution—a sizeable share of the population has a greater than 50% predicted probability of college enrollment. A summary of our predicted probabilities indicates the mean predicted probability is around 0.66, with a range of 0.02 to 0.98.

[4] Our students were nested within high schools, which might suggest we adjust standard errors due to the heterogeneity found within high schools through the use of a vce(*cluster*) Stata option. However, there is a tradeoff here. As Long and Freese (2014) discuss, using robust standard errors no longer makes maximum likelihood an appropriate estimator. After comparing our model with and without school-level clustered errors, we confirmed little difference in our findings and decided to proceed without the robust errors. Models that include robust standard errors should rely on the Wald, rather than the likelihood test (Sribney, n.d.).

Goodness of Fit Next we take stock of how well our data fit the model by examining the goodness-of-fit measures. The likelihood ratio test is calculated as the difference between the logs of the likelihoods of the full (unrestricted) model and unconditional (restricted) model, multiplied by 2, whereby a worse fit is denoted by larger values:

$$Likelihood\ Ratio = 2lnL(Model_{Full}) - 2lnL(Model_{uncond}) \tag{7.10}$$

The likelihood ratio test statistic has a chi-squared distribution and we can therefore treat it as a chi-square statistic (Menard, 2002) to test the null hypothesis that all independent variables are simultaneously equal to zero (Long, 1997). Using Stata's *fitstat* post-estimation command provides a likelihood ratio test statistic of 2911,[5] allowing us to reject the null hypothesis because a chi-square of 2911 with 1 degree of freedom yields $p < 0.001$. The likelihood ratio test can also be used to compare goodness-of-fit across nested models. For example, perhaps theory or prior research indicates that English language learner (ELL) status would help improve the fit of the model. Adding a dichotomous variable that denotes whether students are classified as ELL (or not), the likelihood for the model drops slightly to −5531 (from −5533 in the previous model). A likelihood ratio test between the models with and without the ELL flag yields evidence of a modest improvement in model fit when adding the ELL measure ($\chi^2 = 3.53$, df = 1, $p < 0.10$). While there is strong conceptual justification for inclusion of this variable in prior literature (see Taggart & Crisp, 2011), we see that empirically doing so only marginally improves the fit of the model. Its inclusion is a matter of choice for the researcher. For the sake of consistency, we will use the model without the ELL covariate throughout the chapter.

As one of many post-estimation commands in Stata, *fitstat* displays a suite of summary diagnostic indicators.[6] For example, if we wanted to compare either non-nested models or the same model across different samples, we could use the Akaike Information Criterion (AIC) and/or the Bayesian Information Criterion statistics (BIC, Long & Freese, 2014). Both the AIC and BIC measures are calculated using the model's likelihood, the number of parameters P, and the size of the sample N:

$$AIC = -2lnL\left(Model_{full}\right) + 2P \tag{7.11}$$

$$BIC = -2lnL\left(Model_{full}\right) + Pln(N) \tag{7.12}$$

The models with the lower (rather than higher) AIC and BIC suggest a better fit.

As there is for linear regression models, there is no formal R^2 statistic to assess a logit model's goodness of fit. However, researchers have derived a number of

[5] 2*[(−6988)-(−5533)].

[6] If using survey data, Archer and Lemeshow (2006) argue one should account for survey sampling design to calculate goodness-of-fit using the Stata command *svylogitgof*.

Table 7.3 Comparison of estimates of college enrollment from the logit model[a]

	(1)	(2)	(3)	(4)	(5)	(6)
	Logit coefficients		Odds ratios		Marginal effects	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Female	0.130***	−0.048	1.138***	−0.055	0.022***	−0.008
Race						
Native American	−0.074	−0.221	0.929	−0.205	−0.012	−0.038
Asian	−0.084	−0.1	0.92	−0.092	−0.014	−0.017
Black	0.251***	−0.087	1.285***	−0.112	0.041***	−0.014
Latino	0.131*	−0.07	1.140*	−0.079	0.022*	−0.011
Multiracial	−0.1	−0.083	0.904	−0.075	−0.017	−0.014
White	–	–	–	–	–	–
Academic controls						
GPA, 10th grade	0.567***	−0.035	1.762***	−0.061	0.095***	−0.006
Math test score	0.005*	−0.003	1.005*	−0.003	0.001*	0.000
Number of AP credits	0.129***	−0.018	1.137***	−0.02	0.021***	−0.003
Other student-level controls[b]	x		x		x	
School-level Control[c]	x		x		x	
N	10,940		10,940		10,940	

Source: HSLS:2009

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ~ $p < 0.1$; (a) sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates; (b) other student controls includes parental education, income, highest math taken, whether friends plan to go to college; (c) school-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges

pseudo-R^2 measures that are available when using Stata (and other software packages) by invoking the *fitstat* command. As the default in Stata, McFadden's R^2 compares the log-likelihood of the full (unrestricted) model to an unconditional (restricted) model. Like the R^2 used in linear regression, this statistic is bounded by 0 and 1. The McFadden's R^2 for our unrestricted model is 0.204. An adjusted R^2 measure is also presented. Similar to its linear regression equivalent, this R^2 version accounts for the number of parameters included in the model. For a more detailed discussion about diagnostic statistics used for logistic regression see Long and Freese (2014).

7.2.3 Interpretation of Findings

Coefficients and Odds Ratios Now that we have a sense of model fit, we can turn to the model results reported in Table 7.3. This table includes a number of different point estimates for selected regressors included in the model. For example, given the functional form specified for the variance of the errors ($\pi^2/3$), the (raw) coefficients in column 1 are measured in log-odds or logit units, (Long, 1997). These coefficients are very difficult to interpret, as they lack any practical meaning. But for completeness, the 0.13 logit for the female variable indicates that the log-odds (logit) of enrollment for women is 0.13 higher than that of men.

To ease interpretation, one can transform the logit coefficients (β) into odds ratios (ORs) by exponentiating each raw (logit) coefficient using e^{β} = odds ratio (OR), where e is a mathematical constant that approximates to 2.718. To demonstrate, the logit coefficient for females can be changed to an odds ratio by taking $e^{0.13}$. Stata and other statistical packages will compute the OR for you automatically; or you could compute it using the exp. function either using a calculator or in Microsoft Excel, where $e^{(0.13)}$ produces an odds ratio of 1.138, or, about 1.14 when rounded the nearest hundredths (see the entry for "Female" in column 3 in Table 7.3).[7] Our unconditional (no regressors included) odds ratio for women presented in Table 7.2 was about 1.41, indicating that when we do not control for any other variables, women have about a 41 percent [(OR – 1) × 100 = (1.41–1) × 100 = 0.41] higher odds of enrolling in college than their male counterparts.[8] However, when we control for a set of variables that may confound this relationship, women have about 14 percent greater odds of enrolling in college than men (OR = 1.138, $p < 0.01$), which is statistically significant (Table 7.3, column 3). Additionally, when compared to the unconditional odds ratios of Black and Latino students' (OR = 0.66, $p < 0.001$ and OR = 0.60, $p < 0.001$, respectively), odds flip signs when we control for other factors, with the conditional model indicating higher odds of college enrollment for Blacks and Hispanics versus conditional (OR = 1.285, $p < 0.01$ and OR = 1.140, $p < 0.10$) compared to their White peers. Examining continuous academic measures, we find that for every AP course credit received, the average increase in the odds of college enrollment increases by about 76% (OR = 1.762, $p < 0.001$). Although odds ratios are easier to interpret than the estimated logit coefficients, it is important to note that odds ratios and probabilities are not on the same scale (see Eq. 7.3). Therefore, a doubling of odds is not equivalent to a doubling of the probability (see Long & Freese, 2014, for details).[9]

Marginal Effects In Column 5 of Table 7.3 we present the estimates as marginal effects, or the change in the probability of the outcome given a unit increase in an independent variable. Marginal effects have the desirable property of being measured as percentage point changes in the probability of the outcome, which is likely of substantive interest to researchers and their audience, and makes for a more direct interpretation of coefficients in nonlinear models such as logit and probit models. For

[7] Stata will automatically output odds ratios instead of raw coefficients by using the *logistic* command, or one can obtain odds ratios by invoking the option when using the *logit* command.

[8] A logistic regression model with college enrollment as the outcome and gender as the only covariate will confirm that the odds ratio is indeed 1.41 ($p < 0.001$).

[9] In other words, there is a built-in nonlinearity to the relationship between each covariate and the outcome. However, even with this nonlinearity imposed by the functional form, researchers still need to consider whether any higher order (i.e., polynomials) of covariates are appropriate to account for nonlinear relationships in the logit (or log-odds).

indicator (dummy) or categorical variables, the marginal effect represents the contrast between the reference (or omitted) category and the level of interest. From Table 7.3 we observe that the marginal effect for females (female = 1) is significant but very small—women (female = 1) have predicted probabilities of enrolling in college that are 2.2 *percentage points* (0.022 x 100 = 2.2) higher than men (female = 0), and this effect is significant at the $p < 0.01$ level. Calculated as a partial derivative of a covariate (x) with respect to the outcome (y), the marginal effect for *continuous* independent variables is the change in probability associated with an instantaneous change in the given explanatory variable, holding all other covariates constant. We see (Column 5 in Table 7.3) that the effect of a marginal increase of one Advanced Placement credit is associated with an increase in the probability of college enrollment of 2.1 percentage points ($p < 0.001$). In Stata, marginal effects for the covariates can be obtained using the *margins* post-estimation command invoked after estimating any regressions.

There are a number of ways that marginal effects can be computed, and there is robust discussion about the pros and cons of each (Long & Freese, 2014). As noted above, the marginal effect of any given independent variable depends on the values of all other covariates (i.e., the values at which we hold them constant). A marginal effect at the means (MEM) uses the mean values for each independent variable to calculate the marginal effect.[10] Therefore, the marginal effect is calculated for someone who is average on all of the independent variables included in the model. Though familiar and computationally less intensive than most alternatives, one drawback of the MEM approach is that it raises the question of who, exactly, is "average." This is particularly salient for covariates measured categorically or as integers. Does it make sense to hold the value of AP courses constant at 3.4, even though taking fractional courses is impossible? Or, when controlling for gender using a variable where female = 1 and the proportion of women in the sample is 0.52, does it make sense that the average "gender" in the sample is held constant at this value?

The average marginal effect (AME) approach is a more computationally intensive alternative to MEM that bypasses the concerns mentioned earlier. To calculate AMEs, we first compute the probability of the outcome (in our case, enrollment) for each observation (person) using their actual values for the explanatory variables included in the model. Then one variable is changed by some amount, often 1 unit for categorical variables (i.e., the "delta" method), or a very small amount for continuous measures, (but any interval *could* be used depending on the context), and the outcome probability is recalculated for each person. The difference between these two calculated probabilities is calculated for each observation (person) and then averaged over the entire sample, leading to the AME. As such, the AME is interpreted as the average change in the probability of the outcome resulting from changing the independent variable by some amount. Because of advances in

[10] Computed by adding the *atmeans* option when using the Stata *margins* command.

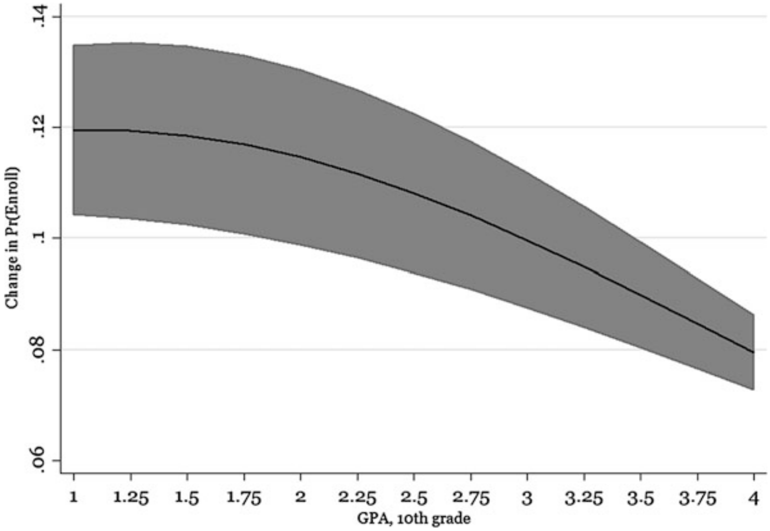

Fig. 7.3 Average marginal effects on college enrollment by student GPA (Source: HSLS:2009)

statistical software, average marginal effects have become more prevalent in the literature (Long & Freese, 2014). Note, however, that both approaches yield a single point estimate for the marginal effect, but this effect may vary depending at what point on the independent variable's distribution the value is chosen. In both approaches, a change in an independent variable in the tails of the S-shaped logit (or probit) curve would yield different changes in probability than a one-unit change near the center (at the mean) of this distribution due to the nonlinear nature of these functions. Therefore, before choosing one of these approaches to interpretation of the results, it is important to consider the pros and cons of AMEs and MEMs and which one seems most appropriate given the objectives of the study.

A third approach is to calculate marginal effects at representative values (MER), where marginal effects are calculated while holding the explanatory variables at user-specified values. MERs allow for the computation of marginal effects along different points of the distribution of independent variables (and not just the mean). For example, we examined the marginal effect of high school GPA on college enrollment across a wide range of plausible GPA values (1.0 to 4.0), and these marginal effects are plotted in Fig. 7.3. While the marginal effect of GPA on enrollment (the solid black line) remains positive (>0) across the different values of GPA, the marginal effect decreases with increases in GPA, and the precision of the marginal effect (as indicated by the confidence interval) increases with GPA. That the marginal effect declines with GPA is unsurprising because college enrollment for high-achieving students is quite high and distinctions between a 3.75 and a 4.0, for example, are challenging to isolate. Regardless of the way the marginal effects are calculated, they are now quite easily available using statistical software packages, and are often reported in tabular format or—for ease of interpretation—plotted graphically.

Predicted Probabilities An alternative to coefficients or marginal effects is to directly compute $\widehat{p}$, or the predicted probabilities of the outcome of interest. Predicted probabilities are particularly useful for the interpretation of interaction terms. Remember that marginal effects compute partial derivatives, allowing one variable to change while holding all others constant. For interactions, however, such a calculation is impossible – to vary the interaction term, $x_1 * x_2$, we cannot hold either variable constant. Further, interaction coefficients can be difficult to interpret in logistic regressions because of the log-odds transformation, which leads to frequent misinterpretation in the literature (see Norton, Wang, & Ai, 2004, for a detailed exposition). In the computation of marginal effects and predicted probabilities, some software packages (such as Stata) automatically aggregate the effect of interacted and polynomial terms.

To illustrate the use and interpretation, we added to the full model (Eq. 7.9) a vector (*INTERACT*) of interaction terms of race, gender, and GPA (female*race, GPA*race, female*GPA), as well as squared and cubed terms of GPA (GPA^2 and GPA^3):

$$\begin{aligned} P(Enroll = 1) = \beta_0 &+ \boldsymbol{\beta_1} DEMS + \boldsymbol{\beta_2} ACAD + \boldsymbol{\beta_3} EXPECT + \boldsymbol{\beta_4} SCHOOL \\ &+ \boldsymbol{\beta_5} INTERACT \end{aligned} \tag{7.13}$$

The raw coefficients table is different than in main effects models (models without interactions) in two important ways. First, we can no longer interpret the estimated coefficients for GPA, race/ethnicity, and gender as main effects, but rather simple effects for White males with average GPAs (the reference group). Second, coefficients that include gender, race, and GPA appear multiple times in the regression output (not shown here) –3, 3, and 5 times, respectively—rendering the net relationships between these measures and college enrollment challenging to interpret. Most of the interaction terms as well as the cubed GPA term were statistically significant. A table of predicted probabilities may be helpful in interpreting differences across categorical groups (e.g., race and gender), which can be attained using the *mtable* post-estimation command with the at() option in Stata to specify values for the covariates (table not shown here, see Appendix for relevant Stata code). Creating an *mtable* for Black, Latino, and White and by gender revealed that, net of student- and school-level variables, the probability of college enrollment is quite similar across groups (the probabilities range from 0.649 for White men and 0.704 for Black women).[11]

For continuous variables (e.g., GPA), researchers may also want to examine predicted probabilities over the plausible values—perhaps through graphical

[11] A formal statistical test can be applied to test the difference between two probabilities using *mtable* and *mlincom*. For more, see Long and Freese (2014).

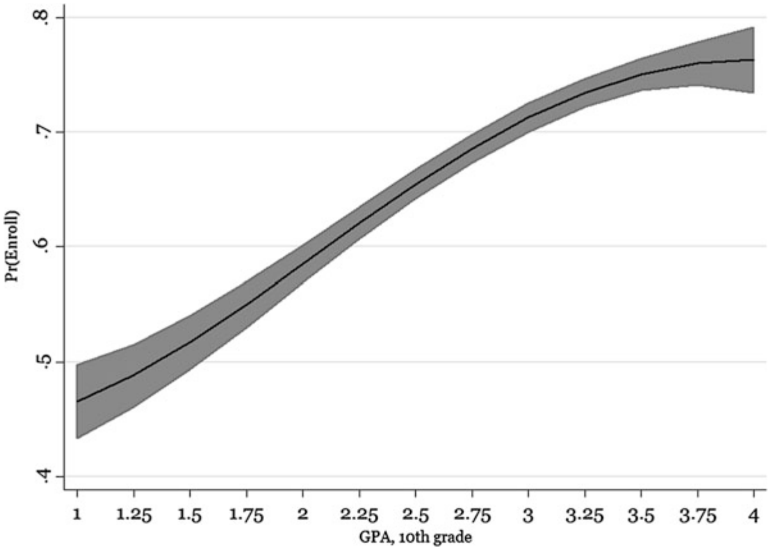

Fig. 7.4 Predicted probability of college enrollment by GPA (Source: HSLS:2009)

analysis. We plotted the probability of college enrollment across the range of high school GPAs (Fig. 7.4). As expected, students with higher GPAs have higher probabilities of college enrollment. There is an inflection point at a GPA at about 2.5, where the slope of the curve begins to flatten out – a function of both the logit functional form and of the polynomial terms of GPA included in the regression. This shape of the curve indicates there is less differentiation in probability at the upper end of the GPA curve, as A and B students are going to college at similar rates. Continuing with our interrogation of college enrollment by gender and race, we plotted the predicted probabilities of college enrollment for Black and White students including an interaction of race and gender. Figure 7.5 shows that the probability of enrolling in college increases for both women and men as GPA increases, net of other variables. Although they are largely parallel, the gender gap increases slightly at the upper end of the GPA distribution. On the other hand, while Black students are more likely to enroll in college, students with approximately a GPA of 3.0 enroll in college at similar rates, irrespective of race or gender. Plots of predicted probabilities can illustrate the nuances that exist across the range of values as well as interactions.

Subgroups of Interest Another advantage of using margins to explain the results of binary regression models is the ability to estimate predicted probabilities for specific groups within one's sample. For example, if you want to produce marginal effects for student profiles of interest, you can use the Stata *mtable* command. Using our running example, if we want to examine the probability of college enrollment for female students by parental education and income, a table of the marginal effects for each of these contrasts can easily be produced (see Table 7.4). First, we employ a

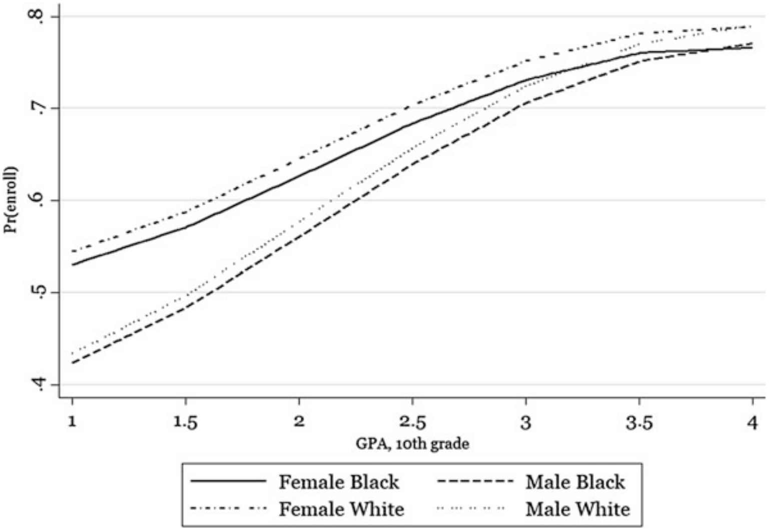

Fig. 7.5 Probability of college enrollment by race, gender, and GPA (Source: HSLS:2009)

Table 7.4 Comparison of the probability of female college enrollment by select income and parental education levels

	Pr(college enrollment)	Lower CI	Upper CI
Panel A: at the means			
Low-income student whose parent(s) has no more than high school degree	0.637	0.604	0.670
Middle-income student whose parent (s) enrolled in but did not attain a college degree	0.650	0.592	0.708
High-income student whose parent(s) earned a college degree	0.808	0.787	0.829
Panel B: at local means			
Low-income student whose parent(s) has no more than high school degree	0.554	0.520	0.587
Middle-income student whose parent(s) enrolled in but did not attain a college degree	0.618	0.557	0.679
High-income student whose parent(s) earned a college degree	0.902	0.890	0.914

Source: HSLS:2009

Notes: Student-level controls include: gender, race, parental education, income, highest math taken, whether friends plan to go to college; School-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges. Sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates (N = 10,940) crosstab of parental education and income (using tab in Stata) to estimate clustering of observations to build our profiles. We find there is a cluster of low-income students whose parents have no more than a high school degree (low-income first-generation college goers). There is also a cluster of high income students whose parents have bachelor's degrees (high-income non-first-generation college graduates). There is an additional set of students we can identify as being first-generation four-year college-goers (their parents have not completed a four-year degree) and are middle-income. We then examined the probability of college enrollment for these three groups using *mtable* to set the values for gender, parental education, and income associated with these three profiles and calculate their probabilities while holding all other independent variables at their means (Panel A in Table 7.4).

In terms of their enrollment probabilities, there is about a 17 (probability) point difference between the least (0.637) and most advantaged (0.808) students. These findings may be limited because parental income and education is often related to other measures, for example the availability of AP courses to students, with low-income students being less likely to gain access to such advanced courses. Therefore, plugging in mean values of the overall sample to calculate predicted probabilities (as in Panel A) may not be as meaningful as plugging in local means for covariates that are more representative of each group. This is an important difference when you have covariates that are markedly different across groups (e.g., the mean GPA for the least advantaged group is 2.59 versus 3.24 for the most advantaged group). To recalculate the predicted probabilities using the local means, we first created variables that identify each group of interest (e.g., low-income students whose parents did not go to college). Then we used these three identifiers to construct three separate *mtables* each producing sets of probabilities where the covariates are held constant at their local means. These results allow us to observe how students with high-income and bachelor's degree holding parents are advantaged. Their probability of enrolling in college is much higher (a 45-point difference) relative to their low-income, first-generation peers (see Panel B in Table 7.4). Long and Freese (2014) discuss how to formally test for differences in these probabilities among subgroups. In general, predicted probabilities are useful for interpreting differences in outcomes across subgroups, and computing predicted probabilities ($\widehat{p}$) using local means can adjust for differences in covariates by subgroups.

Classification or Predictive Accuracy Binary regression models are typically used to predict the probability of outcomes for individual observations and they can also be used to classify individuals (using these predicted probabilities) into categories. For example, in higher education research, previous studies have predicted individual student enrollment propensities (DesJardins, 2002) and others have used predicted probabilities to classify students into groups based on their chances of gaining admission into selective colleges (2013). A simple way to examine how well your model predicts the outcome of interest, another measure of goodness of fit, is to extract the classification diagnostic information produced by logit regression techniques. In Stata, this classification information is available by invoking the *estat classification* command. The classification rate is a calculation of how often a model correctly classifies observations into either y = 1 or y = 0; in our running example, how well our logit model classifies college enrollment or not. A correct classification rate (CCR) of 0.5 means the model correctly predicts outcomes 50% of the time. Such a model would not outperform a random classifying scheme (e.g., flipping a coin to categorize). In addition to the overall classification rate, there are two measures that researchers also often examine. One is sensitivity, or the rate at which the model will correctly classify those experiencing the event or outcome.[12] The unrestricted logit model we estimated correctly identified college-goers 89% of the time (sensitivity of 0.89). Specificity is the rate at which the model correctly classifies those who do not experience the outcome. Our model was not very accurate in classifying non-college-goers (specificity of 0.50). Overall, our model correctly classified college-going across the entire sample 76% of the time.

Hosmer et al. (2013) note that sensitivity and specificity are calculated based on a single threshold value used to classify observations. Statistical programs such as Stata, SPSS, and SAS all use a default predicted probability threshold of 0.50, but researchers may want to specify a different cutoff probability for events with particularly high or low probabilities of occurrence. One way to try to assess whether a 0.50 cut point is optimal is by using the *lsens* command in Stata. This command produces a plot of the sensitivity and specificity across the entire range of possible threshold cut points that could be used. Ideally, we would select a cut point that maximized both sensitivity and specificity measures – at their intersection. In the left side panel in Fig. 7.6, we find the ideal probability cut point for classification is about 0.68. While the *lsens* graph can provide some indication of alternatives to the 0.50 default cut point, it does not tell us how well our model can discriminate college goers from non-college goers in our data. To do so, we need a measure that captures our ability to identify 100% of college-goers (sensitivity) and misidentify non-college-goers 0% of the time (1-specificity) over all possible cut points. This plot is known as the receiver operator characteristic (ROC) curve, whereby the area under the curve is used to determine model fit—the closer to 1, the better the fit of the model. The right-hand side panel of Fig. 7.6 illustrates the ROC curve, plotted with the *lroc* command. The diagonal line represents random assignment to 0 or 1. Therefore the area above that line represents a net increase in sensitivity and reduction in specificity. In our example, the area under the curve is 0.795, which is on the margin of being considered a "very good" fit.[13] Hosmer et al. (2013) warn that the extent to which a model can discriminate between outcomes is not only dependent on the fit of

[12] The default classification threshold in Stata is a probability of 0.5 – observations with probabilities above 0.5 are classified as 1; 0 otherwise.

[13] To be clear, there are no absolute definition of an area under the curve measure that is a "good fit," but rather rules of thumb ranges: 0.5 is no discrimination (or no better than chance); 0.5 to 0.7 is considered poor; 0.7 to 0.8 is acceptable; 0.8 to 0.9 is excellent; and greater than 0.9 is outstanding (Hosmer et al., 2013).

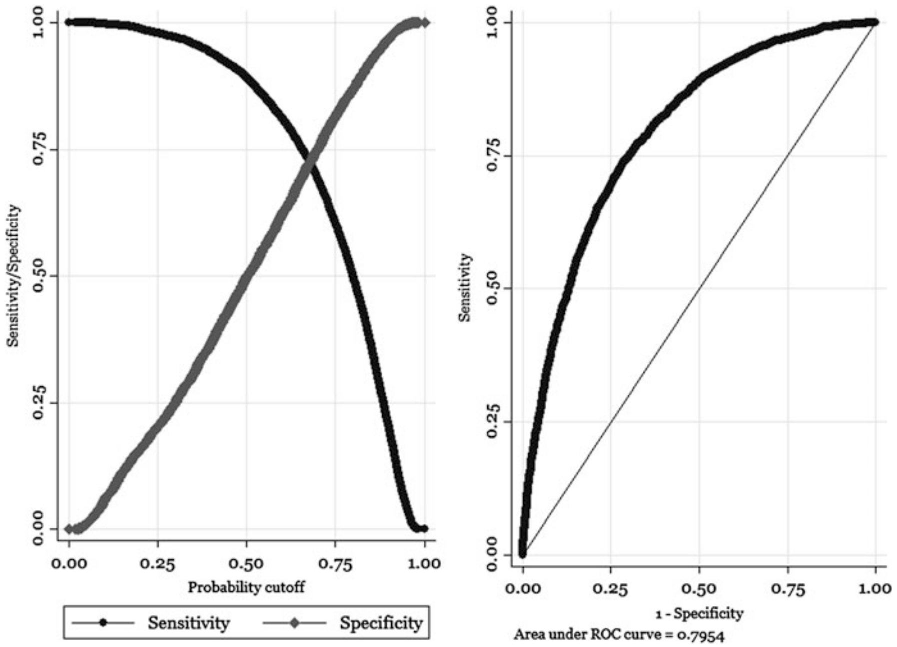

Fig. 7.6 Sensitivity and specificity versus probability/receiver operator characteristic curve (Source: HSLS:2009)

the model, but on the nature of the outcome and differences between the two groups: "we can have a well fitting models that discriminate poorly, just as we could have models with poor fit that discriminate well" (Hosmer et al., 2013, p. 174).

Some scholars note that the aforementioned measures that assess predictive accuracy actually overestimate the precision of these models (DesJardins, 2002; Hosmer et al., 2013). If the researcher's intention is to use data outside of the sample used to derive the model (e.g., in predicting admission or enrollment behaviors using historical data), they should not assume that their models will have similar predictive accuracy. Therefore, in order to better make the case for a model's accuracy in classifying observations or predicting outcomes, researchers should first estimate the model with a random subsample of observations, and then test their predictive accuracy on the reserve (or validation) sample using these tests. See DesJardins (2002) and Chapter 5 of Hosmer et al. (2013) for more on the out-of-sample validation approach.

7.2.4 Probit Regression

We now turn to another technique often used to model binary outcomes, the probit model, and juxtapose it to the logit model discussed above. To transform probabilities into a continuous variable that ranges from $-\infty$ to $+\infty$, the probit approach relies on the inverse cumulative distribution function based on a normal distribution, called the probit link. The cumulative distribution function can transform any value into a value between 0 and 1. Therefore, its inverse can transform the probabilities that range from 0 to 1 into $\pm\infty$. The probit function is formally defined by:

$$\Phi^{-1}[Pr(y = 1|x)] = X'\beta \tag{7.14}$$

$$Pr(y = 1|x) = \Phi(x\beta) \tag{7.15}$$

Where Φ is the cumulative normal distribution function, the -1 takes its inverse, and $X'\beta$ results in a z-score for the probability of the outcome occurring for each record. As such, the coefficient of a probit regression is interpreted as the change in the z-score of the probability of the event occurring. As with logit, probit is typically estimated using maximum likelihood estimation. One assumption of the probit that is distinct from logit is that the errors are assumed to be normally distributed, with a mean of zero and a variance of 1. Recalling Fig. 7.1, the distribution of errors follows the normal curve for both the probability and cumulative density functions, with thinner tails for the logit than for the probit. Approaches to ascertaining goodness-of-fit are similar to those discussed for logit regression.

Interpretation We estimated the same unrestricted model used for the discussion of the logit model. In Table 7.5, the probit coefficients are presented as well as their accompanying marginal effects. The coefficients estimated using the probit model are interpreted in the following way: for each one unit change in the regressor of interest, the z-score of enrollment changes by $\widehat{\beta}$, with larger z-scores being associated with higher probabilities for the outcome of interest. In our running example, we find women's probabilities of enrolling in college are 0.075 standard deviations higher than that of their male counterparts ($p < 0.01$). Interpreting these results in a slightly different way, each one-point change in high school GPA (measured from 0.0 to 4.0) increases the probit index by about one-third of a z-score ($\widehat{\beta} = 0.337, p < 0.01$).

When compared to the logit coefficients in Table 7.3, the magnitude of the probit coefficients are smaller by roughly $\sqrt{3}/\pi$, the conditional variance of the errors assumed for the logit. (Equivalently, the logit coefficients are larger than the probit coefficients by a factor of about 1.7). This difference in the magnitude of the point estimates reflects the assumptions made about the distribution of the (conditional) error variances in the logit and probit models.

Some people find it difficult to interpret the z-score coefficients from probit regressions directly, as they are not expressed in readily understood units. Researchers often revert to presenting probit regression results using predicted probabilities and marginal effects, and we include the latter from our estimated

Table 7.5 Comparison of estimates of the probability of college enrollment, probit and linear probability models[a]

	(1)		(2)		(3)	
	Probit coefficients		Probit MEs		LPM coefficients	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Female	0.075***	–0.028	0.022***	–0.008	0.020**	–0.008
Race						
Native American	–0.055	–0.129	–0.016	–0.038	–0.013	–0.039
Asian	–0.064	–0.056	–0.019	–0.016	–0.013	–0.016
Black	0.147***	–0.051	0.041***	–0.014	0.045***	–0.015
Latino	0.075*	–0.041	0.021*	–0.011	0.022*	–0.012
Multiracial	–0.06	–0.048	–0.017	–0.014	–0.018	–0.014
White	–	–	–	–	–	–
Academic controls						
GPA, 10th grade	0.337***	–0.02	0.096***	–0.006	0.111***	–0.006
Math test score	0.003**	–0.002	0.001**	0	0.001**	0
Number of AP credits	0.066***	–0.01	0.019***	–0.003	0.013***	–0.002
Other student-level controls[b]	x		x		x	
School-level control[c]	x		x		x	
N	10,940		10,940		10,940	

Source: HSLS:2009

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ~$p < 0.1$; (a) sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates; (b) other student controls includes parental education, income, highest math taken, whether friends plan to go to college; (c) school-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges

model results displayed in the third column of Table 7.5. Not surprisingly, the marginal effects derived from the probit are quite similar to those produced by the logit model presented in Table 7.3, and are interpreted in an equivalent manner.

7.2.5 Linear Probability Model

It is not uncommon for researchers to use the linear probability model (LPM) to estimate models where the outcome is binary (e.g., Dynarski, 2004; Hurwitz, 2012). The appeal of the LPM stems from the straightforward interpretation of its coefficients because the coefficients are, simply, marginal effects (i.e., changes in probabilities), holding all other variables constant. A dichotomous dependent variable takes on only two values (e.g., enrollment in college = 1; non-enrollment = 0), thus, OLS regression estimates the mean of that dichotomous outcome – i.e., its expected frequency, and the predicted dependent variable from an LPM, $\widehat{y}$, is the (conditional) predicted probability of enrollment.

Formally the LPM model can be defined as:

$$Pr(y_i = 1|x) = X'\beta + \varepsilon_i \tag{7.16}$$

where y is a categorical outcome for student i who enrolls in college ($y = 1$) or not ($y = 0$); *X'* is a vector of explanatory variables (e.g., academic ability, demographic characteristics, college-promoting networks, and school measures) thought to be related to one's enrollment probability; β is a corresponding vector of parameters to be estimated, and ε represents the error term, which is assumed to be normally distributed.

The LPM has the same set of assumptions as an OLS regression using a continuous dependent variable, and interrogating these assumptions is essential to understanding whether the model is appropriate to the estimation task at-hand (Long, 1997). One assumption is *linearity*, where the dependent variable (y) and the independent variables (*x's*) are assumed to be linearly related through the parameters in vector β. A second assumption is *collinearity*, where the x's are assumed to be independent, that is, none of the regressors (x's) are a linear combination of the other covariates. Next, the error term (ε) is expected to be normally distributed (*normality*) with a mean of zero given a set of x's (the *zero conditional mean of* ε assumption). Additionally, the errors are assumed to be uncorrelated (*uncorrelated errors*) and to have a constant variance across observations, the latter being known as *homoscedasticity*. Intuitively, these last two assumptions suggest that the values observed for one student should not depend on the observed values of another student, and the distribution of the errors should be similar across each covariate (x). A common way to estimate the LPM is using ordinary least squares (OLS), where the objective is to minimize the sum of the squared errors (Long, 1997).

Example: Modeling College Enrollment As with our examples discussed above, we estimated the probability of college enrollment using the following model:

$$\begin{aligned} Pr(Enroll = 1) = \beta_0 + \beta_1 DEMS + \beta_2 ACAD + \beta_3 EXPECT \\ + \beta_4 SCHOOL \end{aligned} \tag{7.17}$$

where *Enroll* is 1 if a student enrolled in college, and 0 if they did not; *DEMS*, *ACAD*, *EXPECT*, and *SCHOOL* are vectors of independent variables (described previously) and their corresponding parameters β's that are to be estimated, and ε is a randomly distributed error term accounting for mis- and unmeasured explanatory variables related to college enrollment. The LPM relies on the same measures of goodness of fit, such as the R^2, as when using OLS to estimate a continuous dependent variable. Our results indicate that the R^2 for this model is 0.24, which is a measure that is not (technically) comparable to McFadden's R^2 often used for the logit and probit models.

Interpretation of Findings The interpretation of the coefficients $(\widehat{\beta})$ is similar to that of a standard linear regression model with a continuous outcome–a one unit change in an explanatory variable x (e.g., one's high school GPA), results in a $\widehat{\beta}$

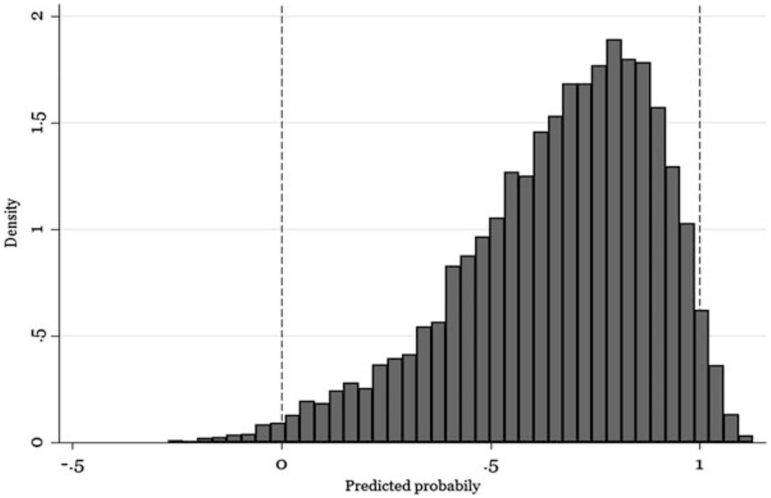

Fig. 7.7 Distribution of predicted probabilities of enrollment: Linear probability model (Source: HSLS:2009)

change in the *probability* of the outcome, in this case, college enrollment (ceteris paribus). The fifth column in Table 7.5 displays the coefficient estimates produced by the LPM (as well as the associated standard errors). On average and net of other variables, women have probabilities of college enrollment that are about two percentage points higher than men ($\widehat{\beta} = 0.020$, $p < 0.05$). When examining race, Black (Latino) students' probabilities of enrolling in college are 4.5 (2.2) percentage points higher than White students. In terms of high school GPA, each one-unit increase results in an 11.1 percentage point increase in the probability of college enrollment ($p < 0.001$). All of these estimates are similar to the marginal effects produced by the logit and probit regressions (see Table 7.3). However, an important distinction is that the LPM imposes a linear constraint such that the effect $\left(\widehat{\beta}\right)$ for each variable (x) is the same (constant) no matter the value of x (i.e., plotting the OLS estimate in Fig. 7.4 would produce a horizontal line at about 0.11).

As is true for the non-linear logit and probit regression models, we can predict the probability of college enrollment for each individual using the LPM results, and these results are presented in Fig. 7.7. It may be troubling that some predictions (about 4%) fall outside of the [0,1] probability interval, thereby providing clearly nonsensical predictions.

We then examined how to use the OLS results to classify students. Although we are unable to use post-estimation commands for classification as we did for logit and probit, we classified students into college enrollment using a threshold of 0.5 and compared it to the observed outcome. We found that similar to the logit model, the LPM's sensitivity (the percent of observations it correctly classified as college-going) was 90.0 percent and the specificity (the percent of non-college-goers it correctly classified) was slightly lower at 47.8 percent.

Drawbacks of the LPM Although LPM is appealing due to its familiarity and intuitive coefficient estimates, the out-of-range predictions in Fig. 7.7 suggest there are limitations to this model. Indeed, many of the assumptions used in the linear regression framework are violated when using a dependent binary outcome. Long (1997) points to four issues with the LPM that we illustrate with our data, below.

Functional Form A fundamental assumption about the linear model is that a given variable (x) will have the same relationship with the outcome (y) across all values of x. In our example from Table 7.5, a one-unit increase in GPA results in a constant change a student's probability of college enrollment across all values of GPA, holding all other variables constant. This implies that the difference in the probability of college enrollment between students with GPAs of 4.0 and 3.0 to be 11 points—the same as the difference between students with 1.0 and 2.0 GPAs (net of other variables). However, we know that college enrollment is quite high among B students; and the differences in college-going may be greater between C and D students than A and B students (as suggested in Fig. 7.3). Therefore, a linear relationship may not best describe how changes in GPA influence changes in college enrollment. One potential way to address such nonlinearity in the relationship between y and a given x is to include nonlinear terms or other transformed versions of x (e.g., polynomials or logged terms).

Heteroscedasticity The assumption that there is constant variance in the x's across the ranges of values is categorically (no pun intended) violated. Mathematically, the variance of a binary outcome y is μ(1-μ), given mean μ. When conditioning on variables $\boldsymbol{x}$, then:

$$Var(y|x) = Pr(y = 1|x) * [1 - Pr(y = 1|x) = x\beta * (1 - x\beta) \tag{7.18}$$

meaning that the conditional variance of y, conditional on x, varies with x. Thus, as Long (1997) notes, the variance of the errors for a binary outcome is not constant, nor are the values of the x's independent. We plot the residuals from the LPM model against its predicted values (using the *rvfplot* command in Stata) in Fig. 7.8, which demonstrates significant heteroscedasticity in the observations in our sample. If the variance was constant, we would expect to see a random pattern of observations around the length of the horizontal line located at y = 0. Although such graphical approaches are useful, to formally test whether the variance is constant we use the *estat imtest* command, which tests the null hypotheses that the variance of the errors is constant and normally distributed. The results of this test (not shown) indicates that the residual variance of the errors is heteroskedastic, thus "the OLS estimator of β is inefficient and the standard errors are biased" (Long, 1997, p. 38).

Non-Normality of Errors The errors of a binary outcome are not normally distributed around the x's. Residuals, you may recall, are calculated as the difference between the observed and estimated (or fitted) values. Because binary outcomes can

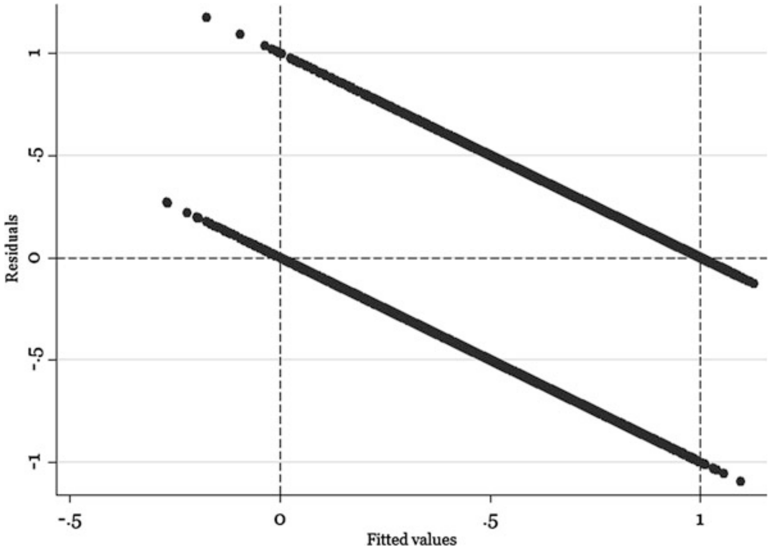

Fig. 7.8 Results of residual-versus-fitted plot for the linear probability model (Source: HSLS:2009)

only take on the values of 0 or 1, residuals can take on only one of two values (Fig. 7.8). For example, for all students who a have an estimated probability of enrollment of 0.80, they have one of two residual values: +0.80 if they actually did not enroll in college or −0.20 if they did. Therefore, structurally, the distribution of errors cannot be normal. You can also examine the normality of the distribution in Stata (Chen, Ender, Mitchell, & Wells, 2003). We first stored the errors using the *predict* command and then compared the density plot of the errors to the normal distribution using the *kdensity* command (Fig. 7.9), which shows a skewed distribution of errors. In addition to a visual inspection of the errors we can also employ one of a number of statistical tests of normality in finite samples. The skewness and kurtosis test for normality (*sktest* in Stata) assesses the symmetry and tail thickness of a distribution, which indeed confirms our visual inspection of Fig. 7.9 ($p < 0.001$ for both skewness and kurtosis).

Out-of-Range Predictions As we illustrated in Fig. 7.7, the LPM can produce probability estimates that are out of the range of plausible values. Indeed, 4% of our sample had predicted probabilities that were either less than zero or greater than 1. However, the college enrollment rate for this sample is somewhat balanced (66%), but when modeling rare or very common events—where the majority of the probabilities are in the tails of the distribution, a LPM will likely produce a larger share of out-of-range predictions. To demonstrate this issue we produce an example where we modeled student *expectations* to enroll in college –which is known to be universally high (91% of our sample expects to go to college)—we find that the

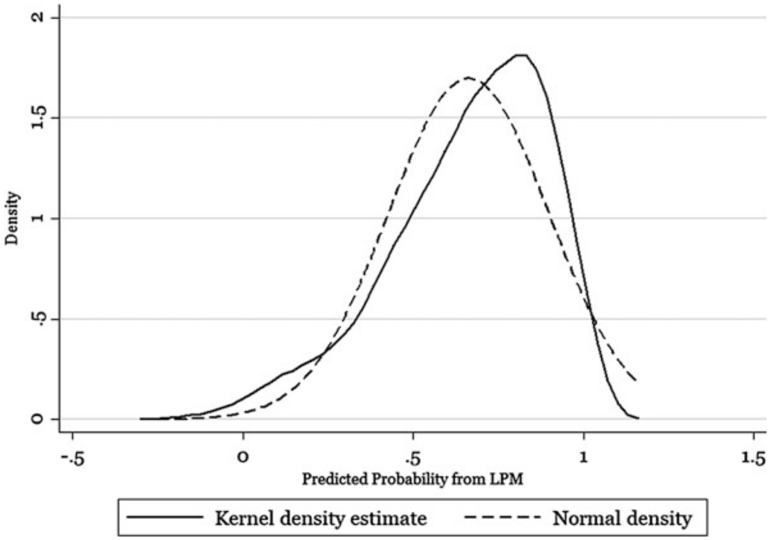

Fig. 7.9 Comparison of kernel density plots, linear probability model estimates and normal distribution (Source: HSLS:2009)

LPM produces predicted probabilities greater than one for almost one-quarter (22%) of the sample, but none less than zero. These findings are, however, sample dependent, as indicated by no predicted probabilities less than one which is due to the very high percentage (91%) of students in the sample who have expectations for going to college. To further illustrate these differences, the boxplot in Fig. 7.10 compares the range of predicted probability estimates for college expectations for the logit, probit, and LPM college expectations model. The LPM has a slightly lower mean predicted probability of expecting to go to college, a larger range of predicted values than the logit and probit, and the upper whisker extended beyond the upper limit of 1, whereas the predicted probabilities for the logit and probit models are bounded by 0 and 1 by construction.

A final illustration details the differences in predicted probabilities in the tails of distribution. Because the probabilities of college enrollment would largely lie in the linear portion of the probit's s-curve, we would expect the college enrollment probability estimates derived from the probit model not to deviate as much from the LPM (save for the tails). However, because the range of probabilities for the college expectations model are generally at the upper end of the distribution, we would expect the linear probability model to diverge for many of the aforementioned reasons. We therefore plotted the predicted probabilities produced by the LPM against those produced by the probit (the logit exhibits a similar result) in a scatter plot for both the college enrollment (Panel A) and college expectations (Panel B) models to illustrate the differences in results (Fig. 7.11). A perfect alignment

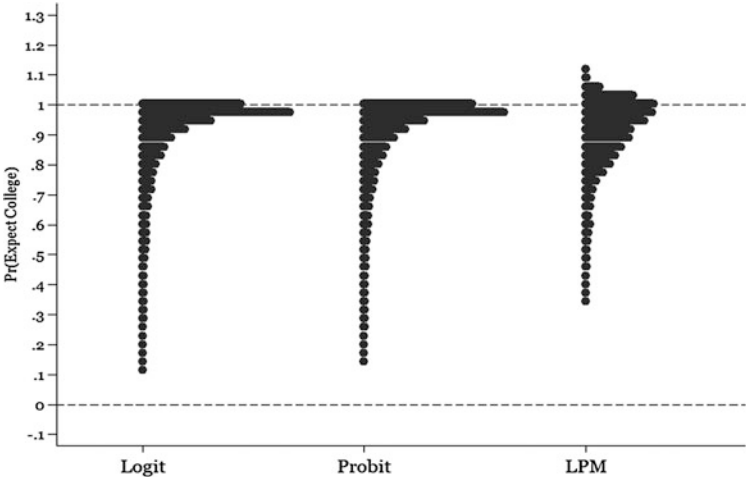

Fig. 7.10 Comparison of predicted probabilities for college expectations (Source: HSLS:2009)

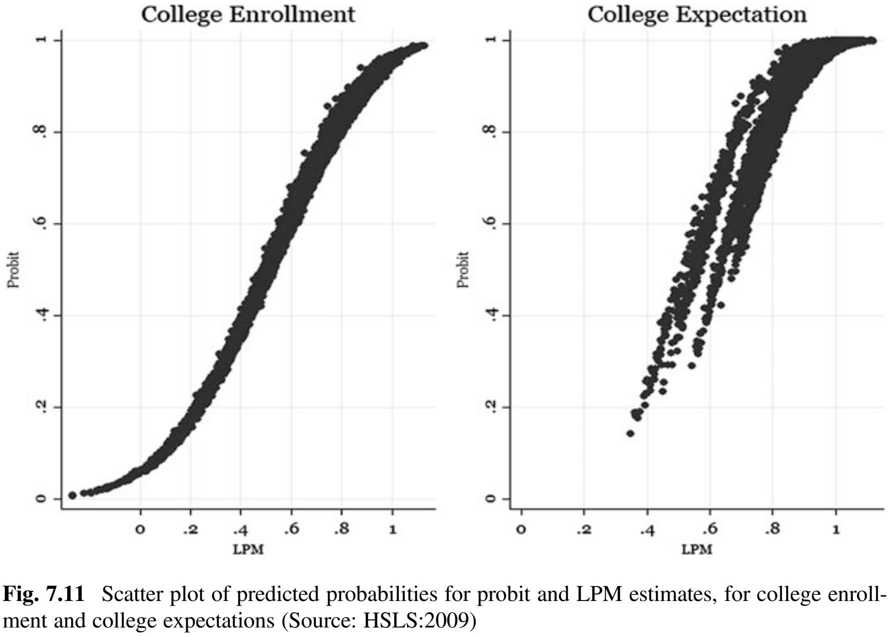

Fig. 7.11 Scatter plot of predicted probabilities for probit and LPM estimates, for college enrollment and college expectations (Source: HSLS:2009)

between the probit and LPM predictions would produce a diagonal line from (0, 0) to (1,1). In Panel A, we observe that the deviations between the linear and probit models are largely in the tails of the sigmoid (S-shaped) curve, which is consistent with their underlying assumptions about the distribution of errors. In Panel B, we find there are large differences in the probit and LPM estimates when modeling an event at the top end of the probability distribution (e.g., where over 90 percent of events occur). This provides further evidence that the LPM may be an inappropriate approach when modeling rare or common.

Moreover, some measures in an LPM may require transformations that are not necessary in logit and probit models. Consider the measure number of AP credits, which the LPM exhibits a small negative (but insignificant) relationship with college expectations of ($\widehat{\beta} =$-0.001, $p > 0.10$) yet the probit estimates a positive and significant relationship ($\widehat{\beta} =$0.011, $p < 0.001$, Table 7.6). A plot of the predicted probabilities indicates that the LPM estimates diverge from the probit and logit as the number of AP courses increases, and the LPM estimates also become much less precise with increases in the number of AP courses completed (Fig. 7.12). A closer look at AP credits reveals that it is heavily skewed right, as many students take none or only one AP course. Due to the linear relationship assumed in the functional form when using the LPM, modeling of nonlinear outcomes with highly skewed distributions while using highly skewed covariates may yield unexpected results. Without transformations of covariates into nonlinear terms, the LPM may not properly account for the clustering of observations at the extremes of the variable distributions. Researchers should consider the prevalence of their outcome and distribution of their covariates before employing this approach.

7.2.6 Conclusion

The ubiquity of binary outcomes in education research has necessitated the use of nonlinear estimation approaches such as the logit and probit. To be sure, linear probability models remain quite popular for estimating dichotomous dependent variables. Notwithstanding the problems noted, some of the reasons the LPM remains popular is its familiarity and the simplicity of the interpretation of the point estimates. Also, there are adjustments that can be made that will remedy some of the assumption violations, such as employing the use of robust standard errors and transforming independent variables that one may think are non-linearly related to the outcome. Nonetheless, the decision of whether to use LPM, logit, or probit when estimating binary outcomes remains a topic of active discussion. Some scholars contend that researchers who employ the LPM should do so with caution because of functional form violations (Long, 1997) and/or the production of

Table 7.6 Comparison of marginal effects on college expectation from logit, probit, and linear probability models[a]

	Logit MEs		Probit MEs		LPM Coefficients	
	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Female	0.021***	−0.005	0.019***	−0.005	0.018***	−0.005
Race						
Native American	0.031	−0.019	0.027	−0.02	0.041*	−0.025
Asian	−0.006	−0.014	−0.013	−0.013	−0.001	−0.010
Black	0.031***	−0.007	0.032***	−0.007	0.044***	−0.010
Latino	0.003	−0.007	0.002	−0.007	0.005	−0.008
Multiracial	0.002	−0.009	−0.001	−0.009	0.007	−0.009
White	–	–	–	–	–	–
Academic controls						
GPA, 10th grade	0.040***	−0.003	0.041***	−0.003	0.056***	−0.004
Math test score	0.002***	0.000	0.002***	0.000	0.002***	0.000
Number of AP credits	0.018***	−0.004	0.011***	−0.003	−0.001	−0.001
Other student-level controls[b]	x		x		x	
School-level controls[c]	x		x		x	
N	10,940		10,940		10,940	

Source: HSLS:2009

Notes: ****p* < 0.001, ***p* < 0.01, **p* < 0.05, ~*p* < 0.1; (a) sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates; (b) other student controls includes parental education, income, highest math taken, whether friends plan to go to college; (c) school-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges

inaccurate estimates (Horace & Oaxaca, 2006). In contrast, other scholars argue that the LPM is a parsimonious estimation approach that yields similar results to logit or probit modeling under a variety of common conditions (Angrist & Pishke, 2009).

There are, of course, tradeoffs to using each approach, and understanding one's data, the conceptual foundations of the issues being examined, and underlying statistical assumptions and how robust the method is to violations of these assumptions are important considerations when choosing an estimator for binary outcomes. The choice between logit or probit is largely dependent on researcher preference and disciplinary norms, though under some circumstances, the probit will have a marginally better fit than the logit (see Hahn & Soyer, 2005, for details). However, given the assumption used for the error distribution, the logit performs well with explanatory variables containing extreme values concentrated in the tails of the distribution. In addition, the logit link function allows for the calculation of the odds ratio, which may be useful in interpretation of one's findings. Moreover, there are some statistical applications that use a specific link function, such as the two-step Heckman selection model which relies on the probit link function for the first step/first-stage equation because the technique assumes bivariate normal errors (Greene, 2002), so understanding it is necessary when employing these techniques.

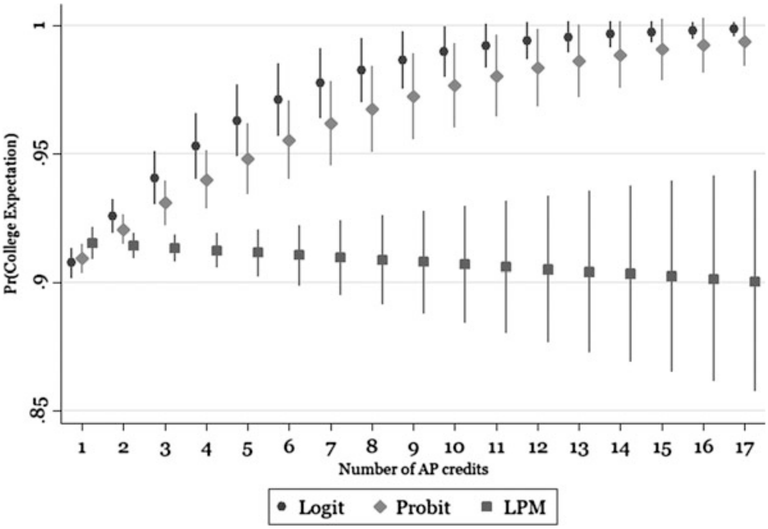

Fig. 7.12 Comparison of predicted probabilities of college expectations by AP credits and modeling approach (Source: HSLS:2009)

7.3 Ordinal Outcomes

In higher education research, there are a number of commonly studied categorical outcomes that take on more than two values and have values that can be *ranked* or set in some *hierarchy*. For example, researchers may want to analyze higher education public opinion data using Likert scales (e.g., ranked from "strongly disagree" to "strongly agree"); one could estimate the probability of students enrolling in colleges according to hierarchical categories of institutional selectivity (e.g., from "least" to "most" selective institutions); we could estimate the probability of earning a particular grade in a college course where grades are ranked from "A" to "F"); or we might estimate high school students' postsecondary expectations from "no college" to "doctoral degree." To model the probability of the event when the outcome measure is ordinal, scholars have employed the ordered logit or ordered probit models (Brasfield, Harrison, & McCoy, 1993; Cheng & Starks, 2002; Doyle, 2007; Morrison, Rudd, Picciano, & Nerad, 2011; Myers & Myers, 2012). For example, in their study of prestige and job satisfaction, Morrison et al. (2011) used ordered logistic regression to examine responses from survey data of faculty perceptions of institutional prestige.[14]

[14] An additional approach that is not discussed here but may of use to higher education researchers is the sequential logit, which models events that individual experience in sequence—for example course-taking (Algebra I, Algebra II, Pre-Calculus); admission stages (application, admission, enrollment); tenure-track faculty positions (Assistant, Associate, Full).

Ordinal outcome variables may represent an underlying continuous latent construct. Drawing from the aforementioned examples—faculty's perceptions of job satisfaction, institutional selectivity, learning (as captured through course grades), etc.—are all complex constructs that may have underlying but unobserved values that are actually continuous. However, we only observe the realizations of this underlying continuous construct. As an extension of binary regression, ordinal regression is similar to logit or probit modeling except there are several (rather than one) cut points along the distribution of the latent dependent variable that cut this distribution into categories that can be observed. To illustrate this latent variable concept, the structural model can be defined as:

$$y^* = X'\beta + \varepsilon \tag{7.19}$$

where $y^*$is a latent continuous outcome that is unobserved and ranges from $\pm\infty$; X' is a vector of regressors; β is a set of corresponding parameters; and ε is a vector of error terms. The categories of outcomes are then defined by thresholds (τ) using the following measurement model:

$$y = c \textit{ if } \tau_{c-1} \le y^* \le \tau_c, \textit{for } c = 1 \textit{ to } J \tag{7.20}$$

where the observed outcome (y) provides "incomplete information about an underlying y^*" (Long, 1997, p. 116) but these thresholds assign the c^{th} outcome category of J possible categories depending on whether the latent measure y^* falls between a lower bound $\tau_{c\,-\,1}$ and upper bound τ_c (Long, 1997).[15] To illustrate, our latent construct we use the selectivity of an institution that a student might choose to attend and we want to model this as a four-category ordinal outcome. These categories are defined as follows:

$$y = \begin{cases} 1, & \text{no college} & \textit{if } \tau_0 = -\infty \le y^* < \tau_1 \\ 2, & \text{less selective college} & \textit{if } \tau_1 \le y^* < \tau_2 \\ 3, & \text{selective college} & \textit{if } \tau_2 \le y^* < \tau_3 \\ 4, & \text{most selective college} & \textit{if } \tau_3 \le y^* < \tau_4 = \infty \end{cases}$$

Using some algebraic manipulation and making some assumptions about error distributions allows us to provide estimates of the *probabilities* that a student will be in each of the categories noted in Eq. 7.20. Formally,

$$Pr(y = c|x) = F(\tau_c - X'\beta) - F(\tau_{c-1} - X'\beta) \tag{7.21}$$

where F is the cumulative distribution function (for either the logit or probit), x is a vector of covariates; and β is a corresponding vector of parameters. The probability of observing outcome c is equivalent to the difference between the probabilities of

[15] For a graphical representation of the cut points, see Long (1997).

being bounded by two thresholds along the cumulative distribution function. Similar to the binary models, this model is estimated using maximum likelihood estimation techniques.

7.3.1 Assumptions

In order to estimate the ordinal regression model, a number of assumptions need to be made about the distribution of errors that are similar to those noted in the binary outcomes section. The ordinal logit has a logistic error distribution with a mean of zero and a variance of ($\pi^2/3$), and the errors are assumed to be normally distributed with a mean a zero and variance of 1 for the ordinal probit model. One additional assumption for ordinal regression is that the slopes of the included regressors are constant across all the outcome categories, which is known as the *parallel slopes (or proportional odds) assumption*.[16] To illustrate, if we modeled enrollment across institutional selectivity categories and included gender as a covariate, women would have the same slope coefficient ($\widehat{\beta}$) for enrollment at a less selective college as they would for a most selective college. This is a very stringent assumption that commonly fails formals tests. There are formal tests of this assumption that are commonly used (discussed below), one comparing the fit of the model using its log likelihood to a model with relaxed assumptions, known as the generalized ordered logit model[17] (likelihood ratio and score tests) or a test whether the $\widehat{\beta}$s are significantly different across categories (Wald or Brant tests).

There are a number of approaches one can take when the parallel slopes (or parallel regression) assumption fails. First, one can identify and remove variables thought to differ across outcome categories, but this strategy may not appeal to researchers if the variable(s) in question are conceptually important. Second, one can fit the model using multinomial regression (discussed in the following section). When employing a multinomial regression (*mlogit*), the assumption that the categories are ordered is relaxed and thus the parallel slopes assumption no longer applies. Imposing a rank-order of outcomes that are not ordinal (and thereby the parallel slopes assumption) will bias your estimates (Borooah, 2002). However, if the dependent variable is truly ordinal and we treat it is as nominal, we may be faced with a loss of efficiency because we have “fail[ed] to impose a legitimate ranking on the outcomes” (Borooah, 2002, p. 3). In choosing between the tradeoff of model efficiency and estimate bias, the former is usually favored, and therefore applying a multionomial regression would be an appropriate course of action..

In some cases, researchers have used OLS regression to model ordinal outcome variables (e.g., modeling Likert-scale responses as continuous). Conceptually, such

[16] See Long (1997) for the derivation of the parallel regression assumption.

[17] The generalized ordered logit model does not assume that the $\widehat{\beta}$'s are equal. See Long and Freese (2014).

an approach assumes that the categories are spaced equidistantly. However, this equidistance assumption may not be true for ordinal data because the magnitude of the differences between categories can vary in ways that are unknown. For example, when employing OLS to estimate our college selectivity dependent variable, the magnitude in the underlying latent construct between not going to college and choosing a less selective college is assumed to be the same as the distance between choosing a very selective and most selective college. However, this may not be case, as there is a lot of heterogeneity within institutional selectivity categories, particularly among less-selective institutions (Bastedo & Flaster, 2014). As a result, the use of OLS to model ordinal outcomes may violate a number of assumptions – particularly the normality and heteroscedasticity assumptions. Winship and Mare (1984) discuss how the ordinal probit and OLS models can produce disparate estimates, some of which parallels our own discussion of the use of linear probability models in Section III above.

7.3.2 *Our Example: College Enrollment by Institutional Selectivity*

In this section we build on our example from the binary outcomes section where the dependent variable is a dichotomous outcome of enrollment/not in college. In this section the dependent variable is one containing four college choice categories: did not attend college, attended a less selective, selective, or most selective college, which is (a priori) assumed to be ordinal. In terms of how students are distributed across these four categories, 49% of students did not enroll in college; 23% enrolled in a less selective institution; 15% enrolled in a selective college; and 12% enrolled in one of the most selective colleges. We included as covariates the same variables used in the binary outcome model discussed above (demographic, academic, expectations, student networks, and school controls), and estimated the ordinal model using the *ologit* command in Stata.

Before interpreting the point estimates produced by the ordinal regression,[18] we check whether the parallel regression assumption is satisfied using a likelihood ratio test (*oparallel*) and Brant test (*brant*). The likelihood ratio test compares the overall model fit of a ordinal logit with a generalized ordered logit model that does not impose the parallel regression assumption.[19] Here, the null hypothesis is that the two models fit the data similarly. For our running example, the likelihood ratio test is statistically significant ($p < 0.001$), meaning we can reject the null hypothesis because the generalized model is a better fit. You can also test the extent to which individual covariates violate the parallel regression assumption using the *brant* command. Brant test results are displayed in the first column of Table 7.7 and

[18] For brevity, we do not include hypothesis tests of the ordered logit or probit, but refer readers to Long and Freese's (2014) overview.

[19] For more on generalized ordered logit models, see Long and Freese (2014).

Table 7.7 Comparison of marginal effects for ordinal logit and multinomial logit models of enrollment by selectivity

	Ordinal logit estimates					Multinomial logit estimates			
	Brant test	No college	Less Sel.	Sel.	Most Sel.	No college	Less Sel.	Sel.	Most Sel.
Female	FAIL	−0.010	−0.001	0.004	0.007	−0.024**	0.024**	0.011	−0.011
		(0.006)	(0.000)	(0.002)	(0.004)	(0.008)	(0.008)	(0.008)	(0.006)
Race									
Native American		0.034	0.001***	−0.013	−0.022	0.011	0.047	0.019	−0.077**
		(0.028)	(0.000)	(0.011)	(0.017)	(0.039)	(0.041)	(0.040)	(0.028)
Asian	FAIL	−0.030**	−0.003*	0.011**	0.022**	0.022	−0.024	−0.027	0.029**
		(0.011)	(0.001)	(0.004)	(0.008)	(0.018)	(0.018)	(0.014)	(0.010)
Black		−0.027**	−0.003*	0.010**	0.020*	−0.039**	0.021	0.007	0.010
		(0.010)	(0.001)	(0.004)	(0.008)	(0.014)	(0.015)	(0.016)	(0.013)
Latino	FAIL	0.004	0.000	−0.001	−0.003	−0.015	0.061***	−0.039***	−0.006
		(0.009)	(0.001)	(0.003)	(0.006)	(0.012)	(0.013)	(0.012)	(0.010)
Multiracial		0.029**	0.001***	−0.011**	−0.019**	0.021	0.013	−0.009	−0.025*
		(0.011)	(0.000)	(0.004)	(0.007)	(0.014)	(0.015)	(0.014)	(0.011)
White	–	–	–	–	–	–	–	–	–
Academic controls									
GPA, 10th grade	FAIL	−0.096***	−0.007***	0.035***	0.068***	−0.106***	−0.021***	0.047***	0.081***
		(0.004)	(0.001)	(0.002)	(0.003)	(0.006)	(0.006)	(0.007)	(0.006)
Math test score	FAIL	−0.002***	−0.000***	0.001***	0.002***	−0.001	−0.002***	−0.001	0.003***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Number of AP credits	FAIL	−0.031***	−0.002***	0.011***	0.022***	−0.007*	−0.019***	0.010***	0.016***
		(0.002)	(0.000)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)	(0.001)
Other student-level controls[b]	FAIL	x				x			
School-level control[c]	FAIL	x				x			
N	10,940	10,940				10,940			

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ~ $p < 0.1$; (a) sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates ($N = 10{,}940$); (b) other student controls includes parental education, income, highest math taken, whether friends plan to go to college; (c) school-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges indicate that the test fails for about half of our $\widehat{\beta}$ s, which is not wholly uncommon. Parallel assumption tests are highly sensitive and can fail due to factors unrelated to the parallel regression assumption (Long & Freese, 2014). A less formal way to test whether the parallel regression assumption is violated is to compare the estimates produced by the ordinal model to estimates produced by a multinomial regression model. We did so (see Table 7.7) and when comparing the marginal effects[20] between the two models we find that some of the estimates are substantively different across the ordinal and multinomial models (e.g., gender, Asian students). Taken together, we have evidence that the ordinal regression approach is not appropriate for examining the probability of enrollment across institutional selectivity in this sample, thus, in the next section we demonstrate how to employ a multinomial regression as an alternative.

Although the parallel slopes violation indicates that the ordinal model is not appropriate in this context, for illustrative purposes we discuss and interpret the ordinal regression model estimates (not shown here) to serve as a reference for researchers employing ordinal regression approaches. Similar to coefficients resulting from binary regression, ordinal logit and ordinal probit coefficients differ by a factor of 1.7, given underlying assumptions of the distribution of errors (as discussed in Section III). Moreover, a two-category ordinal regression yields the same coefficients as a binary regression. It is important to note that regression output from common statistical packages also includes estimates for the J-1 cut points. If you recall from Eq. 7.20, the ordinal outcome is conceptualized as a latent continuous measure (y*) which is carved up into J categories by the J-1 cut points. When $J = 2$ (i.e., when there are two outcome categories, as in a binary regression), the cut point is (basically) equivalent to the constant or intercept (α) in a binary model. But in our running example we have four categories ($J = 4$) which results in 3 (J-1) cut points being estimated by the model. The estimated values produced by for each of the cut points are 3.1, 4.7, and 6.5. Thus, students with an estimated y^* < 3.1 are categorized as not enrolling in college; students with an estimated y^* between 3.1 and 4.7 are categorized as enrolling in less selective colleges; those between 4.7 and 6.5 are in selective colleges; those with a y^* greater than 6.5 are in the most selective colleges group. These cut points ($\widehat{\tau}$'s) are estimated but are generally not of substantive interest and are therefore not often interpreted. However, they can provide some valuable information. If the difference between the cut points is about the same it suggests that the dependent variable is not ordinal but rather on an interval scale. Recall that OLS assumes that the outcome is interval scale, suggesting that using a linear regression may be appropriate.

The raw coefficients produced by the ordered regressions are, as is true for the binary logit case, not intuitive but they can be transformed into odds ratios (if using ordered logit) or predicted probabilities. In our example, the raw coefficient for the Female variable is 0.0675 ($p < 0.10$) which can be transformed into an odds ratio.

[20] Marginal effects are useful here in comparing across models.

The Stata output produces these raw coefficients by default, and they represent cumulative odds of belonging to a category or higher versus belonging to the lower categories.[21] In our example the odd of females belonging to the no college group vs. all the other categories (less/selective/most) are about 1.07 ($\exp^{0.0675} = 1.0698$) times that of males. Equivalently, the odds of females belonging to the no college/less selective groups vs. the selective/most selective groups are also about 1.07 times that of males. This demonstrates how the effect of being female on the different contrasts does not vary, which will not be the case for the multinomial models discussed later in the chapter (see Long & Freese, 2014, for a further discussion of interpretation issues).

Although there are numerous outcomes that interest higher education researchers that are ordinal in nature, ordinal regression analyses is less often used because many times the parallel regression assumption tests fail. Scholars (Borooah, 2002; Long, 1997) also caution about the use of ordinal measures when categories can take on multiple meanings and ordering. For example, if we are considering earning potential, we might order a category of institutional levels as no college, two-year college, and four-year college. However, if we are considering time to earn a degree from shortest to longest, we might reorder the institutional levels as two-year, four-year, and no college, whereby those who are not yet enrolled in college are considered to (theoretically) have the longest time to earn a degree. The main point here is that different conceptualizations of the latent construct, and the context in which the categorical ordering is being used, can lead to different conclusions (Long & Freese, 2014). When the parallel regression assumption fails or the ordering of categories is not certain, researchers may want to consider the use of one of the multinomial regression models available, discussed in the next section.

7.4 Nominal Outcomes

Some categorical outcomes are not rank ordered but are rather measured on a nominal scale. In higher education research there are many nominal outcomes of interest: choices among college majors (e.g., liberal arts, pre-professional, STEM, other); reasons for selecting or leaving a college (e.g., availability of financial aid, familial obligations, academic rigor); college-going outcomes (e.g., graduated, still enrolled, transferred, no longer enrolled); or the types of jobs PhD students select upon graduation (e.g., private industry, faculty, public service, non-research). Regression-based models used to estimate multiple nominal outcomes are known as *multinomial models* and these have been used to study many different issues in higher education (e.g, Bahr, 2008; Belasco, 2013; Eagan et al., 2013; Porter & Umbach, 2006; Wells, Lynch, & Seifert, 2011). To illustrate, Bahr used a multinomial probit model to examine the relationship between math course-taking and the

[21] This is why this model is often called the "cumulative logit model."

long-term degree outcomes of community college students, where the outcome categories of interest were transfer with credential, transfer without credential, degree with or without certificate, certificate only, or no credential). Researchers also employ multinomial regression when they are uncertain about the ordinal nature of their data or are not able to employ ordinal regression. In their study of the representation of women at selective institutions, Bielby et al. (2014) estimated multinomial logit models of college application behaviors by institutional selectivity—arguably ordinal and consistent with our analysis above—because the ordinal model they initially estimated failed the parallel regression assumption test.

When analyzing multi-categorical outcomes, one might be tempted to run separate binary regressions to estimate each pairwise contrast of the categories. For example, we might examine enrollment by institutional selectivity by modeling separate binary regressions for: no college enrollment versus less selective enrollment; college enrollment at less selective versus selective colleges; and so forth. However, this approach results in several regression results that have different sample sizes, leading to a loss of efficiency of the estimates. Furthermore, this approach is deficient because it does "enforce the logical relationship among the parameters" for each of the categories (Long, 1997, p.151). In contrast, multinomial regression simultaneously estimates all of the possible outcome category relationships, and does so making full use of all the available data, thereby remedying the problems noted above when employing binary regression to estimate multinomial outcomes.

To demonstrate the utility of the more popular multinomial regression techniques available, below we formally present these models, how they are identified, discuss their underlying assumptions and model fit tests, and provide examples of how to interpret the results. We do so using our running example of the study of college enrollment (by institutional selectivity).

When estimating a multinomial environment, the probability of observing outcome category m among J possible categories can be modeled as:

$$Pr(y_i = m|x) = \frac{exp\left(X'\beta_{m|b}\right)}{1 + \sum_{j=2}^{J} exp\left(X'\beta_{j|b}\right)} \tag{7.22}$$

where b is the base outcome, x is a vector of covariates and $\beta_{m\ |\ b}$ is a corresponding vector of coefficients relating outcome category m with respect to the base outcome. The reader may notice that this equation is an extension of Eq. 7.3, which formally describes the binary outcome model. One difference is that the denominator in Eq. (7.22) is modified to accommodate more than two outcomes categories. Equation 7.4 (the binary logit representation) can also be modified to account for any number of (J) outcome categories:

$$\ln \frac{Pr(y=m|x)}{Pr(y=b|x)} = X'\beta_{m|b} \, for \; m = 1 \; to \; J \tag{7.23}$$

where b is the base outcome, X' is a vector of covariates and $\beta_{m \mid b}$ is a corresponding vector of coefficients relating outcome category m to the base or reference category. In Eq. 7.23, known as the multinomial logit, we take the natural log of the relative risk ratio, $\frac{Pr(y=m|x)}{Pr(y=b|x)}$. As noted in section III above, the relative risk ratio is not to be confused with the odds ratio – the ratio of two odds (Menard, 2010).[22] Like the binary logit model, the multinomial logit is linear in the parameters (the logits), making the underlying statistical calculation easier to perform. Also of note, in Eqs. 7.22 and 7.23 is the inclusion of a base or reference category. The parameters are estimated using maximum likelihood. Multinomial logit regression output typically only includes estimates for J-1 of the outcome contrasts, with the base or reference category estimates being omitted. Given that the pairwise comparisons of the estimates for coefficients produced by the model will be relative to the reference category, scholars are encouraged to carefully consider the choice of the base category.

7.4.1 Assumptions

One of the assumptions underpinning the multinomial logit is the independence of irrelevant alternatives assumption (IIA), whereby the odds of observing an outcome do not depend on the other available alternatives.[23] In words, this means that the addition or elimination of outcome categories (i.e., alternatives) will not change the odds of observing the outcome. For example, suppose students have three college options available to them – let’s call them colleges A, B, and C—and the odds of a student choosing between College A and B are evenly split. Under the IIA, the presence (or elimination) of the third College C (the alternative) should have no bearing on the students’ odds between the other two choices (A and B), essentially making College C an “irrelevant alternative.” However, in practice, this assumption will not make sense from a conceptual point of view. The elimination of College C might mean that more students seek out College A, if for example it had very similar program offerings as College C, thereby fundamentally changing the relative odds between students choosing between Colleges A and B. The main argument being, if there are enough similarities between the added alternative and one of the already available options, then IIA will not hold. Empirically, there are formal tests—the

[22] Researchers should be careful to distinguish between the risk ratio and odds ratio, as they are not interchangeable terms. In particular, odds ratios and risk ratios are most dissimilar in the middle of a distribution (Menard, 2010). Only when J = 2 are the relative risk ratio and odds ratio equal. For a clear explanation, see https://www.stata.com/statalist/archive/2005-04/msg00678.html

[23] The IIA also applies to the conditional logit (not discussed here).

Hausman-McFadden and Small-Hsiao tests—available to test whether the IIA assumption holds. However, these tests have been shown to be inconsistent in identifying violations of the IIA that are related to the size and structure of the data. Long and Freese (2014) question the relevance of these tests and argue that one should select outcomes categories that appear to be theoretically distinct, in order to argue that the multinomial categories are valid. When there is insufficient theoretical guidance and/or strong empirical evidence that the IIA assumption is violated, one can also employ *multinomial probit regression*, which does not rely on the IIA assumption (e.g., Titus, 2007). As is often the case, there are tradeoffs to consider when choosing to use the multinomial probit rather than the multinomial logit. The former does not produce risk ratios, which may ease interpretation (Long & Freese, 2014). The multinomial probit is more computationally intensive, but advances in computing power make differences in estimation time negligible for moderately sized datasets (Greene, 2002; Long & Freese, 2014). However, as was the case for logit and probit models, researchers can now easily produce predicted probabilities and marginal effects for both the logit and probit multinomial models, and there are many possibilities for displaying these results in graphical format.

7.4.2 Estimating College Enrollment by Institutional Selectivity

We revisit estimating enrollment by institutional selectivity, which failed the parallel regression assumption test for the ordinal regression analysis in the previous section. As you may recall, we are interested in understanding the relationships between college enrollment where the outcome categories are: no college, less selective, selective, or most selective colleges. We regress this dependent variable on a number of variables thought to explain this choice (e.g., student demographic characteristics; academic achievement, etc.). Given there was no evidence that the outcome needed to be estimated using ordinal regression we will now employ an alternative technique, multinomial logistic regression. Using enrollment at a less selective institution—where the majority of students enroll in college—as the base outcome, we estimate the following multinomial logit model:

$$\ln \Omega_{NC|LS} = \beta_{0,NC|LS} + \beta_{1,NC|LS} DEMS + \beta_{2,NC|LS} ACAD + \beta_{3,NC|LS} EXPECT + \beta_{4,NC|LS} SCH \tag{7.24}$$

$$\ln \Omega_{S|LS} = \beta_{0,S|LS} + \beta_{1,S|LS} DEMS + \beta_{2,S|LS} ACAD + \beta_{3,S|LS} EXPECT + \beta_{4,S|LS} SCH \tag{7.25}$$

$$\ln \Omega_{MS|LS} = \beta_{0,MS|LS} + \beta_{1,MS|LS} DEMS + \beta_{2,MS|LS} ACAD + \beta_{3,MS|LS} EXPECT + \beta_{4,MS|LS} SCH \tag{7.26}$$

where $\Omega = \frac{Pr(y=m|x)}{Pr(y=b|x)}$ and the numerator is the probability of observing the m^{th} outcome category [e.g., no college (*NC*), selective (*S*), or most selective (*MS*) institution) relative to the probability of being in the base outcome (the denominator) *b,* whether the student chose a less selective (*LS*) college. This ratio of two probabilities (risk ratio) is, thus, a relative measure, leading to it being dubbed the relative risk ratio.[24] Included as regressors are *DEMS*, *ACAD*, *EXPECT*, *SCH*, vectors of demographic, academic, college expectation, and school characteristics, respectively, described in Table 7.1. As was true for the binary and ordinal regressions, the β's are parameters to be estimated. Recall that the ordinal regression model produced only one set of parameter estimates for the regressors included, whereas the multinomial model produces such estimates for each covariate for each of the outcome categories.

Goodness of Fit and Combining Outcomes As with the binary logit or probit, we can use Stata's *fitstat* command to examine how well the model fits the data. This command produces a number different measures of the model's goodness of fit (see help files for details). Additionally, the likelihood ratio and Wald tests can be invoked to test the null hypothesis (H_0) that all of the coefficients are simultaneously equal to zero. These tests can be conducted using the *mlogtest* post-estimation command and the *wald* and lr options, respectively.[25] Relatedly, if the coefficients ($\widehat{\beta}$'s) are not significantly different across outcome categories, then there is evidence that these non-distinct categories can be combined, which would improve the efficiency of the model and ease interpretation as there will not be as many pairwise contrasts to explain. One way to test if any of the outcome categories can be combined is by using the *mlogtest* command with the *combine* or *lrcombine* options in Stata. The former option uses the Wald test, the latter a likelihood ratio test. We employed these tests and found that, in our sample, the null hypothesis that the any of the outcome categories could be combined was rejected, providing no evidence for combining any of the four categories.

Interpretation Output from a multinomial logistic regression (MNL) can easily overwhelm because there are *J*-1 panels of estimates presented as regression output (as noted earlier, the base outcome results are not presented) and estimated coefficients for each of the regressors included. In our case, we have four panels of regression output, one for each of the outcome categories. Although the output produced by Stata (and other statistical packages) typically includes only the statistics for the non-base outcome category, here we use Stata's *listcoef* post-estimation

[24] For an explanation of odds and risk ratios/relative risks see: http://www.theanalysisfactor.com/the-difference-between-relative-risk-and-odds-ratios/

[25] Only the Wald test works when using robust standard errors or survey commands, See Long and Freese (2014) for a discussion of tradeoffs between the Wald and likelihood ratio tests.

command (discussed below) to present the statistics for each of the four outcome categories (all pairwise comparisons; see Table 7.7). One should approach the interpretation of multinomial regression results with a targeted analysis plan a priori (e.g., focusing variables of interest), so that the interpretation of the results does not overwhelm the reader. Below we briefly discuss three ways to examine and present findings: relative risk ratios, marginal effects, and predicted probabilities.

Relative Risk Ratios As noted above, we have to select a base outcome in order to fit the multinomial logit (remember in our example, the base outcome is less selective institutions). However, researchers may have an interest in making contrasts to pairwise categories that do not include the base outcome. For example, we may want to contrast selective and most selective institutions, which is not available in the default output produced by Stata. Given the *somewhat* ordered nature of our outcomes by increasing levels of selectivity (i.e., no college < less selective < selective < most selective), examining contrasts of adjacent categories is one way to interpret the results. Stata's *listcoef* post-estimation command can help in presenting the results by providing results about any pairwise contrasts the analyst might be interested in examining. We used this option to produce such results, and Table 7.8 displays the relative risk ratios (the exponentiated $\widehat{\beta}$'s) for our covariates of interest: gender, race, and student GPA, for each of the outcome categories. The results provide evidence of the female advantage that was observed for the binary logit results, but this relationship is more complex than initially thought. The differences being that the gender differences are concentrated on the no college/less- selective college margins. Relative to men, women had about a 19% higher risk (probability) of enrolling in a less-selective college compared to not attending college (the base category). But no statistically significant gender differences were evident for the less selective to selective institution contrast, but between selective and most selective institutions, women had a 13% *lower* risk of enrolling at the most selective institutions, consistent with previous research (Posselt et al., 2012). These gender differences were masked when using a binary representation of the outcome of interest, demonstrating the utility of using the multinomial representation of the dependent variable and modeling approach that permits a more detailed examination of the relationships among the outcome categories and explanatory variables.

To demonstrate the interpretation of a non-categorical regressor, we present the results for AP credits. There is no evidence of a statistically significant relationship of AP credits with enrollment in less selective colleges, relative to not enrolling in any college. However, a one-credit increase in AP credits is associated, on average, with an 18% increase in the relative risk (probability) of enrollment at a selective institution, relative to a less selective institutions, and a 12% increase in the relative risk of enrollment at a most selective institution, relative to selective institution.[26]

[26] Note these are increases in *probability*, rather than *odds*.

Table 7.8 Comparison of relative risk ratios by college selectivity[a]

	No college-less selective	Less selective-selective	Selective-most selective
Female	1.185***	0.961	0.872~
Race			
Native American	1.115	0.79	0.392*
Asian	0.844	0.989	1.428**
Black	1.257*	0.997	1.071
Latino	1.283***	0.634***	1.105
Multiracial	0.966	0.852	0.805
White	–	–	–
Academic controls			
GPA, 10th grade	1.407***	1.750***	1.801***
Math test score	0.998	1.009*	1.034***
Number of AP credits	0.964	1.181***	1.123***
Other student-level controls[b]	x	x	x
School-level control[c]	x	x	x

Source: HSLS:2009

Notes: ***p < 0.001, **p < 0.01, *p < 0.05, ~p < 0.1; (a) sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates ($N = 10{,}940$); (b) other student controls includes parental education, income, highest math taken, whether friends plan to go to college; (c) school-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges

Marginal Effects In multinomial regression, marginal effects are one useful way to check on the relationships between the outcome categories and the included covariates. As noted above, marginal effects help researchers understand the average change in probability associated with a change in the given covariates. One advantage of using marginal effects is the ability to compare results across models. Using marginal effects, the researcher can consider the different ways in which one might contrast the outcomes—at adjacent margins (as we did in Table 7.8), against one base outcome (the default), or some other configuration that makes most sense for your analysis. Given the myriad of contrasts available to the researcher for J-1 outcomes and k covariates, a full table of output may not be an effective way of ultimately presenting findings. Long & Freese, 2014 further caution that as marginal effects are computed using partial derivatives, they are highly dependent on the shape of the probability curve and on the levels of all variables in the model—potentially leading to large changes in sign and magnitude, depending on the place on the probability curve where relationships are being examined. Therefore, researchers are encouraged to examine marginal effects along the various points on the probability curve, similar to the presentation in Fig. 7.3. Many of the issues we covered in the Binary Outcomes section related to marginal effects apply to multinomial regression, and will not be discussed further in this section.

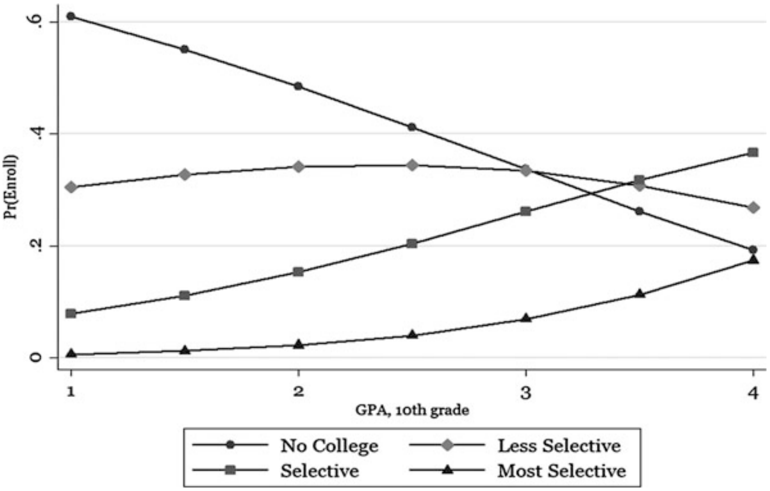

Fig. 7.13 Predicted probabilities of enrollment by college selectivity and 10th grade GPA (Source: HSLS:2009)

Predicted Probabilities Predicted probabilities also allow researchers to evaluate the relationship between covariates and outcomes at different points in the probability distribution in a readily understood metric. Predicted probabilities can be presented in tables (typically useful for categorical variables); graphical plots (for continuous variables); and for specific subgroups. In our specific example, interpreting predicted probabilities across the range of high school GPA values and for specific populations of interest, may help us better understand the relationship between gender, GPA, and institutional selectivity. In Fig. 7.13, we plotted the effect of GPA across the four enrollment outcome categories. Students with higher GPAs are less likely to opt-out of college immediately after high school than students with lower GPAs. Interestingly, the enrollment effects are relatively flat across the range of GPAs for students who are likely to enroll in less selective colleges. The probability of enrollment for students choosing the most selective colleges is relatively flat for students with average and below high school GPAs (<= 2.0), but then rises to about 20% for the students with GPA's of 4.0.

Analysis of specific subgroups is also a useful approach for understanding the results produced by such models. Expanding on our example of female college enrollment from Table 7.4, for the multinomial model we find that the probability of a low-income female student whose parents had not attended college have a probability of 0.54 of not attending any college, and less than a 1 percent chance of attending a most selective institution (see Table 7.9). A middle-income woman whose parents attended college but did not obtain a degree also had high probabilities of either not going to college (0.47) or attending a less selective institution

Table 7.9 Probability of college enrollment for female students by select income and parental education levels

	No college	Less Selective	Selective	Most Selective
Low-income student whose parent(s) has no more than high school degree	0.542	0.365	0.088	0.005
Middle-income student whose parent(s) enrolled in but did not attain a college degree	0.472	0.328	0.182	0.017
High-income student whose parent(s) earned a college degree	0.113	0.174	0.371	0.342

Source: HSLS:2009

Notes: Student-level controls include: gender, race, parental education, income, highest math taken, whether friends plan to go to college; School-level controls includes urbanicity, school type, and share of students enrolled in 2-year and 4-year colleges. Sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates ($N = 10{,}940$)

(0.33), and very low chances of choosing to attend a highly selective college (0.02). In contrast, women from families who had high income and college-degreed parents had relatively high probabilities of attending a selective (0.37) and most selective college (0.34), net of academic, school, and other measures. These results suggest that the observed female advantage is in many ways driven by the choice of institution type as well as one's family background.

The multinomial regression models allow for the estimation of unordered categorical outcomes by relaxing some of the assumptions imposed when employing ordinal regression. While multinomial logit is commonly used to model nominal dependent variables, when there are conceptual grounds and/or empirical evidence that the IIA assumption is violated the multinomial probit is an alternative. Regardless of the link function that is chosen (logit or probit), the amount of output produced by multinomial regression models is oftentimes described as "overwhelming" (Long & Freese, 2014, p. 411). Hosmer et al. (2013) note that although the complexity of the multinomial model produces considerable output to interpret (especially when there are numerous outcome categories), the researcher has multiple estimates for each covariate, thereby providing "a complete description of the process being studied" (p.289). To ease the interpretation burden, in the section above we presented a number of different approaches to present the findings in a digestible way.

7.5 Limited Dependent Variable Models

So far, we have focused our attention on categorical dependent variables, whether they have two (binary), ordered (ordinal), or multiple (nominal) categories. We now turn our attention to limited dependent variables that may seem at first glance to resemble continuous measures but "whose range of values is substantially restricted" (Wooldridge, 2008, p. 529). These outcome variables may be restricted to integer values, as in the case of count variables; they may be restricted to observing values only over specific ranges, such as proportions that lie between zero and one; or these variables may be censored or truncated in various ways, either by definition (e.g., variables that cannot take on negative values) or because of data generating processes (e.g., top-coded variables used in surveys, sample selection). In the next section we begin the discussion of these limited dependent variables by illustrating Poisson and negative binomial regression techniques for count variables. We then discuss analytical approaches for other forms of limited dependent variables, such as fractional logistic, Tobit, and double-hurdle regression models.

7.5.1 Count Outcomes

We begin our discussion of limited dependent variable models by discussing count outcomes. Count outcomes are those that enumerate the number of occurrences of particular events, and such outcome variables abound in higher education. Scott-Clayton (2011), for example, used a count of total semesters enrolled over 4 years as an outcome in her evaluation of West Virginia's PROMISE scholarship program. Researchers may also be interested in the number of courses students take, as enrollment intensity is associated with several educational outcomes such as time to degree and persistence (Stratton, O'Toole, & Wetzel, 2007). As demonstrated by Goldrick-Rab (2006), students' transfer behavior is also of substantive interest, as the frequency and timing of transfers vary across a number of student and institutional characteristics, with important consequences for completion and time to degree. In many cases, researchers study count outcomes using OLS regression techniques (e.g., Scott-Clayton, 2011). It is also possible to create discrete categories of count outcomes instead of using the count as the dependent variable. For example, Goldrick-Rab's (2006) research on "swirling" students defined multinomial outcome (e.g., did not transfer; stopped out and returned; transferred without interruption) from the underlying frequency and direction of student transfers. Such transformations may be appropriate for some research questions, but in other contexts may result in loss of information that is of conceptual or empirical importance.

In cases where outcomes are measured as counts of events directly, the use of OLS regression "for count outcomes can result in inefficient, inconsistent, and biased outcomes" (Long, 1997, p. 217). Count outcomes take only integer values, may have a relatively high preponderance of zeros, and may take on many small values – suggesting that alternative regression techniques that account for these characteristics may improve on OLS estimates (Greene, 2002). In this section, we use students' college applications to introduce count regression models. Today's high school graduate is more likely than ever to apply to multiple postsecondary institutions (Pryor, Hurtado, Saenz, Santos, & Korn, 2007). The increase in college applications submitted by students is the result of numerous co-occurring trends: increased competition in college admissions (Bastedo & Jaquette, 2011; Eagan, Lozano, Hurtado, & Case, 2013); simplification of the college application process

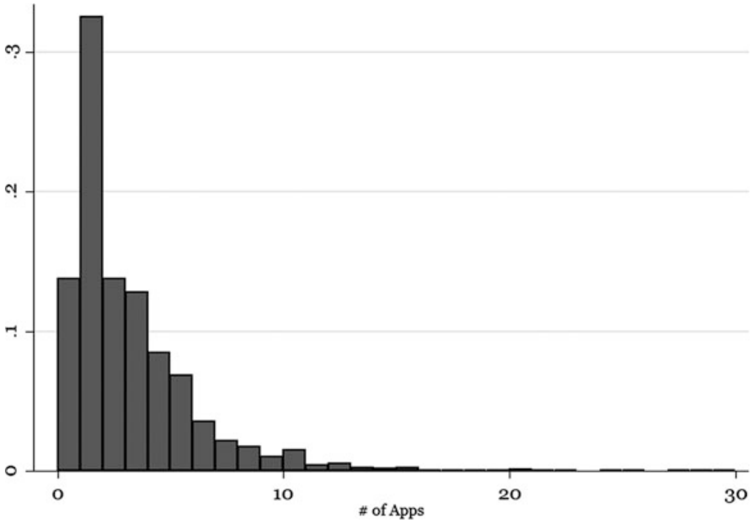

Fig. 7.14 Histogram of number of college applications (Notes: Histogram includes nonmissing values of S3CLGAPPNUM, winsorized at 30. Source: HSLS:09)

(Pryor et al., 2007); and proactive marketing and outreach by colleges and universities (McDonough, 1994). Applying to college is an important step in the college choice process, as the application set defines and constrains the choices eventually available to students and reflects students' preferences, constraints, and the appeal of institutions to individuals. The number of applications students submit to college display many of the properties that count regression techniques are intended to address. In our sample, we observe a nontrivial number (density) of students who apply to zero colleges; the modal number of applications is one; and the mean number of applications is quite low (2.7). As shown in Fig. 7.14, the distribution is skewed to the right indicating some students apply to many institutions (the maximum is 30).

An extensive literature addresses students' college application behavior, ranging from research into applications to single institutions (e.g., Gonzales & DesJardins, 2002); research into "undermatch" (e.g., Smit et al., 2013); and studies of the characteristics of students' college application sets (e.g., Arcidiacono, 2005; Blume, 2016; Niu & Tienda, 2008). A few authors have studied the number of applications students submit as an outcome. Hurtado, Inkelas, Briggs, and Rhee (1997) used standard OLS regression to explore differences in the number (count) of college applications across students' race and ethnicity, finding significant disparities particularly for traditionally underserved student populations. Howell (2010), Pallais (2015), and Smith, (2014) analyzed how various changes to application or admissions policies affected the number of applications students submit. Howell (2010) used ordinal probit regression, with the outcome specified as zero, one, two to four, or more than four applications, to investigate the effect of affirmative action on college applications and enrollment. Pallais (2015) estimated a difference-in-difference model using OLS to identify the effect of changes to the cost of ACT score sending on the number of applications that students submitted to colleges. Finally, Smith (2014) investigated how expansion of the Common Application influenced the number of colleges to which students applied, also using OLS.

To our knowledge, no paper has made use of regression techniques specifically designed for count data to study the number of college applications students submit. There are a number of regression-based methods that can be used to study outcomes that are counts, such as when studying college application submissions, and herein we explore how to do so. The prevalent count regression techniques include Poisson and negative binomial regression, as well as the zero-inflated variants of each. Choosing among these four options depends on two characteristics of the outcome variable:

1. The dispersion of the outcome (its conditional mean relative to the conditional variance).
2. The nature of zero counts in the data: whether they are "excessive" and whether they are the product of mechanisms distinct from those governing positive count values.

7.5.2 Regression Techniques for Counts

To begin this discussion, we employ Poisson regression as the starting point in modeling count outcomes. If we are interested in a variable y that measures the number of applications students submit, Poisson regression treats the count of y as though it is drawn from a Poisson distribution with parameter μ (Long, 1997). This distribution can be related to covariates of interest through a log-linear model (Greene, 2002). The log transformation ensures that the regression model cannot result in negative values (Atkins & Gallop, 2007). Thus, we can think of the outcome as measuring the probability of student i applying to y colleges as:

$$\Pr\left(Y_i = y_i | X_i'\right) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{7.27}$$

Using the log-linear model, we can associate the (natural log of the) μ parameter to our explanatory variables of interest (X') as:

$$\ln\left(\mu_i\right) = X_i'\beta \tag{7.28}$$

where X' is a vector of relevant student characteristics such as collegiate expectations, demographics, and academic achievement. This model can also be used to estimate the expected (average) count or number of applications each student submitted using:

$$E\left(y_i|X_i'\right) = e^{X_i'\beta} \tag{7.29}$$

The model is estimated by maximum likelihood.

The Poisson distribution has a few properties that may be of interest to researchers. Its shape is largely governed by the mean rate – as the mean count approaches zero, the distribution grows more right-skewed and at a sufficiently high mean count, the distribution approaches normality. In addition, as the mean increases, we would expect to observe fewer counts equaling zero (Atkins & Gallop, 2007; Cameron & Trivetti, 1998; Long, 1997). One of its most limiting characteristics is that the Poisson distribution assumes equidispersion: the variance and mean of y are assumed to be equal. In other words, the variance of y is equal to mean value, shown in Eq. 7.29 above. However, count data are frequently overdispersed, with (conditional) variance exceeding the (conditional) mean. Descriptively, we observe strong suggestive evidence of overdispersion in college application counts. The histogram of the distribution of application counts in Fig. 7.14 indicates that these counts are right-skewed, with an unconditional mean of 2.7 that is exceeded by its unconditional variance of 7.6. Table 7.10 also provides evidence that the conditional (on the variables shown) variance exceeds the conditional mean for some important variables used as regressors to explain counts in college applications (more on this below).

In the presence of overdispersion, Poisson regression yields consistent but inefficient estimates, and can understate standard errors (Long, 1997). There are several formal tests for overdispersion following estimation of a Poisson regression (Cameron & Trivetti, 1998; Greene, 2002). In Stata, we can calculate a goodness of fit Chi-squared statistic with the gof command – a large Chi-squared value indicates a poor fit that may be the result of overdispersion. When there is evidence of overdispersion, negative binomial regression is often employed. The negative binomial regression combines the Poisson distribution for the mean of the outcome variable with a gamma-distributed parameter that adjusts its variance, with the result that (conditional) variance exceeds the (conditional mean), thereby accounting for the overdispersion (Long, 1997). If we treat r as the shape parameter of the gamma function, the following describes the mean (the same as Eq. 7.29):

$$E(Y|X') = e^{X'\beta} \tag{7.30}$$

Whereas the variance for negative binomial is:

$$var(Y|X') = \mu + \frac{1}{r}\mu^2 \tag{7.31}$$

which means that the variance is larger relative to the mean for small values of r; with the negative binomial distribution converging to Poisson as r approaches infinity (Cameron & Trivedi, 1998). There are several formal tests for overdispersion, many of which are built into statistical software. Stata's *nbreg* command for negative binomial regression estimates an overdispersion parameter termed alpha, which is defined as $\alpha = r$ – in other words, the inverse of ${}^{1}/_{r}$ in Eq. 7.31.

Table 7.10 Mean and variance for college applications by student characteristics

Variable	Number of applications	
	Mean	Variance
Race		
Native Americans	2.3	6.2
Asian/Pacific islanders	4.0	12.2
Black	2.9	8.9
Hispanic	2.4	6.9
Multiracial	2.5	7.0
White	2.5	6.7
Gender		
Male	2.4	6.9
Female	2.9	8.3

Source: HSLS: 2009
Notes: Summary statistics for S3CLGAPPNUM variable

An α parameter that is statistically different from zero indicates overdispersion. The *nbreg* command performs a likelihood ratio test of the significance of α by comparing a model that constrains α to equal zero and a model where α is empirically estimated.

A final consideration when studying counts in a regression framework is how to treat values of zero. In our college application example, it is possible for students to report (in a survey) that they did not apply to any college. About 14% of respondents in the HSLS sample report applying to zero colleges, a non-trivial number of non-applicants. There are two general approaches to modeling count outcomes with zero counts. One could treat zero counts as any other positive integer, thereby not differentiating them from other values. Alternatively, if the proportion of zeros relative to positive counts is sufficiently large, *or* if the occurrence of a zero count is of substantive interest as its own phenomenon, one could employ the zero-inflated variants of count models. Zero-inflated Poisson or zero-inflated negative binomial regressions are similar in spirit to other mixture or two-part models, which we discuss in greater detail later in this chapter. Two-part models allow for the zero counts to be estimated separately from the rest of the distribution, with each model having its own set of covariates. In the case of college applications, such a model would allow us to model the decision not to apply to college at all as its own outcome. For zero-inflated count regressions, we first fit a logistic model of the probability of observing a count equal to zero, and then fit a Poisson or negative binomial regression on the positive integers (Greene, 2002).[27] In other words, we would model zero counts that occur with probability π using logistic regression, and would model positive integer counts that occur with probability 1-π using Poisson or negative binomial regression, as outlined in Eq. 7.32:

[27] If overdispersion is the result of excess zeros, a zero-inflated Poisson model may be preferable over negative binomial regression (Long, 1997).

$$\Pr\left(Y_i = y_i|X_i'\right) \sim \begin{cases} \pi_i + (1-\pi_i) * g\left(Y_i = 0|X_i'\right) \; \mathit{if} \; y_i = 0 \\ (1-\pi_i) * g\left(Y_i|X_i'\right) \; \mathit{if} \; y_i > 0 \end{cases} \tag{7.32}$$

Using a zero-inflated model, we would model $\pi_i = 0$ with logistic regression – in the example we use below, this would mean estimating the probability of (not) applying to college. We would then model $g\left(y_i|X_i'\right)$ with Poisson or negative binomial regression for positive integer values of y_i (i.e., conditional on applying to college at all). This approach may be valuable in instances where the underlying mechanisms governing zero and positive counts differ. For example, a study of drinking behavior on college campuses may include in its sample subgroups of students who are not at risk for drinking at all (say, for example, due to religion). Stata's zip and *zinb* commands estimate these models.

7.5.3 Applying Count Regression to College Applications

We can model the count of applications submitted by student i as a function of students' and families' characteristics, measured academic achievement, extracurricular involvement, postsecondary intentions, and characteristics of their high school:

$$\begin{aligned} \Pr\left(Y_i = y_i|X_i'\right) = \beta_0 &+ \boldsymbol{\beta_1} race_i + \boldsymbol{\beta_2} gender_i + \boldsymbol{\beta_3} family_i + \boldsymbol{\beta_4} acad_i \\ &+ \boldsymbol{\beta_5} extra_i + \boldsymbol{\beta_6} pse\ plans_i + \boldsymbol{\beta_7} HS\ context_i + \varepsilon_i \end{aligned} \tag{7.33}$$

We estimate Poisson and negative binomial models of Eq. 7.33.[28] Recall that the distribution of college application counts suggested possible overdispersion. We observe that the α parameter for overdispersion is statistically significant (χ^2 =2053.3, $p < 0.001$) in the negative binomial regression. We also find that the Chi-square measure of goodness of fit for the Poisson regression is also highly significant (χ^2=15,203.4, $p < 0.001$), again suggesting the presence of overdispersion. As discussed above, overdispersion results in underestimated standard errors for Poisson regression. We observe that in the results included in Table 7.11, that this is indeed the case – the coefficients of both models are quite similar, but the standard errors of the negative binomial model are larger than for the Poisson regression because they are inflated by the overdispersion parameter. We can also look to the AIC and BIC statistics of the two models to further assess fit. All three statistics suggest that the negative binomial is preferable to the Poisson. Yet another way to compare these models is to test how well they fit the underlying distribution of college applications. In Figure 7.15, we plot the residuals of $\Pr\left(Y_i = y_i|X_i'\right)$ for the Poisson and negative binomial regressions at each value of

[28] Recall that if we were concerned about the 14% of students that do not apply to college (and thus have a count of zero), or if we wanted to understand the decision not to apply to college separately, we could estimate a zero-inflated count model.

Table 7.11 Comparison of estimates for count models

	Poisson	Neg Bin	Poisson IRR	Neg Bin IRR	AMEs (Neg Bin)
Gender (ref. male)					
Female	0.115***	0.116***	1.121***	1.123***	0.327
	(0.013)	(0.017)	(0.014)	(0.019)	
GPA, 10th grade	0.285***	0.297***	1.330***	1.345***	0.842
	(0.011)	(0.014)	(0.015)	(0.018)	

Source: HSLS:2009

Notes: ***p < 0.001, **p < 0.01, *p < 0.05, ~p < 0.1. Models include additional controls for race, parental education, family income, AP/IB credits, extracurricular activities, hours worked, and high school characteristics. Standard errors in parenthesis. Sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates ($N = 9740$)

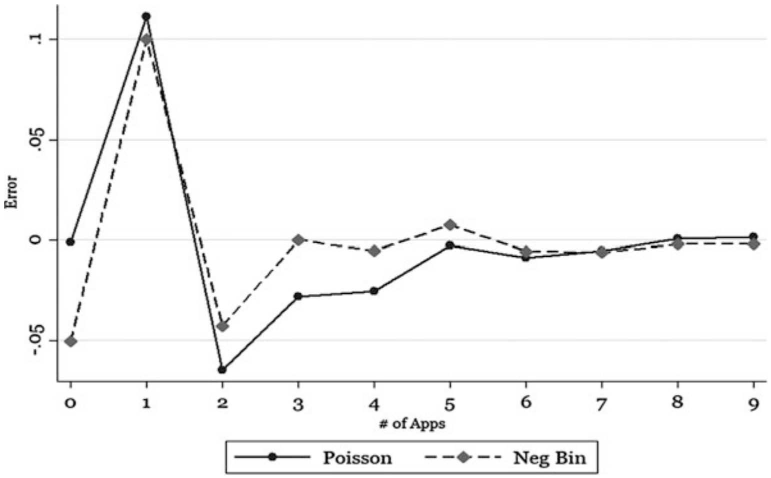

Fig. 7.15 Comparison of residuals for poisson and negative binomial regressions of count of college applications (Notes: Positive residuals indicate underpredictions. Source: HSLS:2009)

y_i. In other words, the residuals show the degree to which the models under- or overpredict the probability of each count value. The graph indicates that the negative binomial model does a slightly better job of fitting the observed distribution of applications.[29]

We have addressed goodness of fit and, given evidence of overdispersion, have chosen negative binomial as our preferred modeling approach. Turning to the coefficients, as with other nonlinear regression techniques, interpreting regression

[29] This may seem like a large number of goodness of fit tests to run, but in Stata the user-written command *countfit* provides all of these results simultaneously.

results can be tricky. The coefficients of a negative binomial regression are linear and additive with respect to the logged values of the expected count, as seen in Eqs. 7.28 and 7.29 - which are not easily interpretable. Table 7.11 reports coefficients for a few select explanatory variables. An increase of 0.1 in 10th grade GPA, for example, is associated with a 0.029 increase in the expected number of applications (0.1 multiplied by the coefficient of 0.297). Similarly, the coefficient for female implies that female students have an expected number of college applications that is about 11% higher than for male students. Exponentiating these coefficients yields incidence rate ratios (IRR), which may be more readily interpretable, and are reported in Table 7.11. These IRRs represent *factor changes* in the expected count $E\left(y_i|X'_i\right)$, and so can be interpreted as multiplicative like an odds ratio. That is, a one-unit increase in a covariate is associated with an increase of e^{β} in the expected outcome, all else held constant (Long, 1997). Keeping with the same two variables as examples, a one-unit (or one point) increase in high school GPA increases the expected *number* of applications by a factor of 1.35 – a 35% increase. Similarly, female students have 1.12 times the expected number of college application of male students, or 12% higher applications.

An alternative to IRR is to compute marginal effects. As we discussed in section about binary outcomes, marginal effects provide a useful way to summarize associations at mean, observed, or representative values of interest. They also help us translate coefficients from percent or factor changes in the expected number of applications to a more intuitive unit of measure; i.e., the actual count of applications. The marginal effect for high school GPA tells us that a one-point increase in GPA is associated with 0.84 additional college applications, while being female is associated with 0.33 additional applications.

Marginal effects also help us make sense of interaction terms. To demonstrate this, we estimate an additional model that includes an interaction of gender and high school GPA (as reported in Table 7.12). The interaction term in this model allows for the relationship between high school GPA and college applications to vary by gender. When we add this interaction to the previously estimated model, we find a larger main effect of gender (1.32 vs. 1.12) than before, though we must also consider the interaction effect. Interestingly, the interaction effect of GPA and gender is *negative* (or, in IRR terms, less than 1). This suggests that as GPA increases, the difference in the number of applications submitted by men and women declines.

To ease the interpretation of main and interaction effects, we graph (see Figure 7.16) the relationship between high school GPA and the number of college applications separately by gender. The graph indicates a slow convergence of the two groups, especially at higher values of GPA.

There are numerous variables in higher education that enumerate phenomena of interest. Researchers always have the option to take such outcomes and transform them into dichotomous or categorical measures, or to treat them as continuous. However, count regression techniques are simple and have desirable robustness

Table 7.12 Interaction terms in count regression for college applications

	Negative binomial (IRR)
Gender (ref. male)	
Female	1.320***
	(0.099)
GPA, 10th grade	1.381***
	(0.025)
GPA, 10th grade*female	0.949*
	(0.022)

Source: HSLS:2009

Notes: ***p < 0.001, **p < 0.01, *p < 0.05, ~p < 0.1. Models include additional controls for race, parental education, family income, AP/IB credits, extracurricular activities, hours worked, and high school characteristics. Standard errors in parenthesis. Sample includes all students with base year, follow-up, and transcript data that are not missing data on covariates ($N = 9740$)

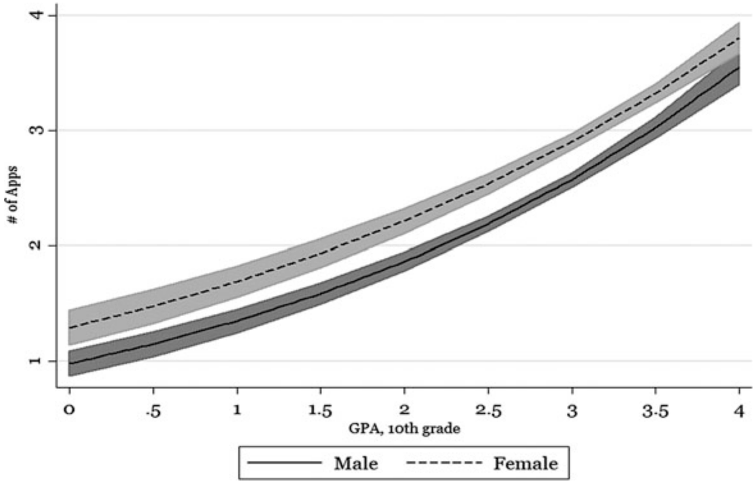

Fig. 7.16 Predicted count of college applications by gender and GPA (Notes: Shaded region indicates 95% confidence interval. Model includes controls for student characteristics; measured academic achievement; extracurricular involvement; postsecondary intentions; high school characteristics; and an interaction of gender and academic achievement. Source: HSLS:2009)

properties (Wooldridge, 2008) that enable us to study these variables with few compromises.

7.5.4 Proportional/Fractional Outcomes

Researchers in higher education frequently encounter outcomes of interest that are measured at the institutional level. Such data are widely available from sources such as IPEDS, and capture several measures relevant to scholars, prospective students, and regulatory agencies. These variables include the composition of the student body (e.g., proportion from underrepresented groups, percentage of students that are Pell-eligible) and institutional outcomes such as persistence, graduation, or student loan default rates. All such variables are proportions or fractions, with a range from zero to one.

With few exceptions, much of the research using proportions or rates as outcomes uses ordinary least squares regression. For example, scholars studying stratification of enrollments in higher education often attempt to explain such stratification across race and income, noting that there is a dearth of low-income or marginalized students at the nation's most prestigious universities (Bastedo & Gumport, 2003; Carnevale & Strohl, 2013). They often do so by employing OLS regression to examine outcome variables that are proportions. Belasco, Rosinger, and Hearn (2015) used linear regression to study the effect of test-optional admissions on the share of Black and Hispanic enrollments. Hillman, 2013 used much the same approach to study changes in the proportion of Pell students at institutions adopting no-loan financial aid policies. Institutional student loan default rates are another widely studied topic, with recent studies by Hillman, 2014, Ishitani & McKitrick, 2016, and Kelchen & Li, 2017 about individual and institutional characteristics associated with default.

In each of these studies, OLS regression provided readily interpretable coefficients and insights into policies or institutional characteristics of interest. Earlier, we highlighted the ways in which OLS regression with binary outcomes (LPM) may violate assumptions necessary for efficient, unbiased estimates. As was the case with linear probability models, "the drawbacks of linear models for fractional data are analogous to the drawbacks of the linear probability model for binary data" (Papke & Wooldridge, 1996, p. 620). Proportions or rates are bounded by zero and one, whereas fitted values from OLS regressions are not. Second, the relationship of any independent variable to a fractional outcome cannot be linear through the full range of outcome values (Papke & Wooldridge, 1996). Finally, the residuals from OLS regression of fractional outcomes are likely heteroskedastic, with greater variation at middle values and smaller variation near the lower and upper bounds (Cribari-Neto & Zeileis, 2010). As such, researchers should approach linear regression of fractional dependent variables with caution.

To be sure, under many conditions OLS regression may prove to be a reasonable enough approximation of a fractional outcome. For example, when the proportional outcome is largely distributed within the linear portion of the logistic curve, the estimates produced using a linear specification may be a reasonable approximation (Cribari-Neto & Zeileis, 2010). Institutional graduation rates are a good candidate for such an approach because graduation rates in many colleges and universities are near the middle of the distribution (the average being about 60%). However, other outcomes may not lend themselves well to OLS regression. At selective institutions, for example, we know that the fraction of enrolled students receiving Pell grants is quite low, with many such colleges having fewer than 20% of students as Pell recipients (Carnevale & Van der Werf, 2017). Similarly, though there is much justified media and scholarly attention paid to student loan default, institutional student default rates averaged 11.3% for the 2013 cohort (Federal Student Aid, 2016). In these instances, researchers may be well served by exploring alternatives to OLS regression, just as we have done with binary dependent variables. We discuss some approaches below.

7.5.5 Alternatives to OLS for Fractional Outcomes

There are several alternatives to linear regression for the modeling of proportions. One common approach is to use a transformation of the dependent variable, such as a log transformation in Eq. 7.34 (Baum, 2008; Papke & Wooldridge, 1996). If p is the measure of the relevant fractional outcome; by log transforming the fraction the model becomes linear in the parameters, similar to what we saw in our discussion of logistic regression:

$$E\left(\ln\left(\frac{p}{p-1}\right)|x\right) = X'\beta \tag{7.34}$$

One example of using such a transformation is Scott, Bailey, and Kienzl (2006), who studied the relationship between graduation rates and the characteristics of institutions and their student bodies. However, this transformation has an important limitation: it excludes fractions at the endpoints of the [0,1] interval, as the term $\left(\ln\left(\frac{p}{p-1}\right)\right)$ is undefined for p equal to either zero or one. This transformation is also of limited interpretability: the coefficients of the regression measure changes in the log-transformed outcome, not in the actual fractional outcome of interest. Finally, this transformation does not address the heteroskedastic nature of rate or proportion data (Ferrari & Cribari-Neto, 2004).

One alternative is to use beta regression, which improves on the log transformation in two ways. First, beta regression can accommodate outcome variables in the (0,1) interval that are left- or right-skewed, or that are flatly distributed over the full range. This is because beta regression treats the dependent variable as following a beta distribution, which is highly flexible, with the beta density taking a variety of shapes. This also means that beta regression can accommodate the heteroskedacity inherent to fractional outcomes. Second, the coefficients of a beta regression are directly interpretable as changes to the mean expected value of the outcome. Thus, they require no additional effort for calculation of marginal effects or use of graphs. However, beta regression does share one important limitation with log transformation, which is that it is only defined for fractions in the (0,1) interval – dependent variables equaling precisely zero or one are excluded (Ferrari & Cribari-Neto, 2004).

In cases where proportions at the extreme values of 0 or 1 exist and researchers want to retain such observations, neither a log-transformation nor a beta regression may be appropriate. Fractions at either extreme may seem rare or unlikely to be observed – this is certainly the case for metrics like graduation rates, for example. However, these values do occur for measures that tend to be concentrated at the tails of the distribution or among select subsamples of postsecondary institutions (e.g., retention rates at highly selective institutions, which are extremely high and could reach 100%; the proportion of low-income students at small colleges with high net prices, which are very low and may be 0%). The zero-inflated (and one-inflated) variants of beta regression can retain such observations. Similar to our discussion of zero-inflated count models, these variations on beta regressions are mixture models. These models first estimate the probability of observing a fraction equal to zero or one using logistic regression, and then estimate a beta regression model for outcomes within the (0,1) range (Ospina & Ferrari, 2012). One advantage of this approach is that it allows us to specify different covariates for each of the models estimated, which may be particularly valuable for researchers that posit different underlying mechanisms for observations at the extremes of [0,1].

Yet another alternative is the use of fractional response models, as outlined in Papke and Wooldridge (1996). Fractional response models allow for modeling proportions in the [0,1] interval, using a generalized linear model with a link function:

$$E(y|x) = G(X'\beta) \tag{7.35}$$

Where $G(\cdot)$ is a link function, typically the logistic or standard normal (probit) cumulative density functions, and $0 \leq y \leq 1$. The model is estimated using quasi-maximum likelihood (which does not require knowledge of the full distribution of outcomes), with the link function indicating the distribution of mean values for the outcome variable.[30] The coefficients from a fractional logistic regression are not easily interpreted; their sign indicates the direction of marginal effects but otherwise convey little readily usable information. As with probit or logit models, marginal effects, graphical representations of the relationships between covariates and the outcome of interest, and predicted values provide a more easily interpretable way to communicate results (see Furquim & Glasener, 2016 for an application of fractional logistic regression to the proportion of Pell eligible students at highly selective institutions).

Though these alternative approaches to modeling proportions require transformations of the dependent variable or the use of link functions, statistical software such as R, SAS, or Stata can estimate any of them using their respective GLM

[30] One could also use heteroskedastic probit to model the variance rather than the mean of a proportional outcome.

regression commands. Stata version 14 and higher also includes specific commands for beta regression (*betareg*) and for fractional response models (*fracreg*). These regression techniques are readily available to researchers, and may be of use to higher education scholars studying the fractional or proportional outcomes so commonly of interest to policymakers and prospective students.

7.5.6 Censoring and Truncation

Many of the dependent variables of interest to higher education scholars can be censored or truncated in ways that warrant special consideration. For example, researchers interested into students' decision to work while in school run into a censoring issue, as students cannot work fewer than zero hours. The same form of censoring affects studies of student indebtedness – students cannot borrow amounts below $0. In many instances, this censoring is overlooked and researchers rely on OLS regression. For example, Addo, Houle, and Simon (2016) studied the relationship between parental wealth and student debt using OLS regression, and excluded nonborrowers from their analyses. If we are interested in the censored observations (non-borrowers), however, OLS estimates of censored variables can be inconsistent, as OLS fails to "account for the qualitative difference between limit (zero) observations and nonlimit (continuous) observations (Greene, 2002, p. 762).

Tobit regression provides a workaround for censored variables. As in our discussion of categorical outcomes, Tobit regression is also a latent variable technique. In the case of student loans, we can think of Tobit regression as modeling a latent demand for student loans of the form:

$$y^* = X'\beta + \varepsilon \tag{7.36}$$

And

$$y = \begin{cases} 0 \; if \; y^* \leq 0 \\ y_i^* \; if \; y^* > 0 \end{cases} \tag{7.37}$$

Where y^* is a latent construct capturing the true demand for loans; the observed outcome y is the measure of student loans that is censored at zero for negative values of y^*.[31] Taken together, Eqs. 7.1 and 7.2 tell us that in a Tobit regression a change to any element of X' affects both the probability of y_i being greater than zero (in our case, of taking on student loans) as well as the conditional mean of y^* for $y^* > 0$ (Greene, 2002; Long, 1997). See Hart and Mustafa (2008) for an application of Tobit regression to study the effect of increased access to subsidized loans on student debt.

[31] Tobit models also work for censoring from above, such as when data from surveys top-code variables like income for privacy reasons.

One limiting assumption of Tobit regression, however, is that $X^{'}\beta$ is assumed to equally affect both the likelihood of borrowing *and* the mean amount borrowed (Lin & Schmidt, 1984). This may not be a desirable or sensible assumption in some cases. If individuals face a participating decision, such as a decision of whether to borrow at all, the double-hurdle model introduced by Cragg (1971) may be preferable. The double-hurdle model allows for the specification of a decision to participate (borrow, in our example) and then separately to model the amount borrowed. These two regressions can take different functional forms and include distinct covariates, allowing researchers to better consider the mechanisms underlying the two distinct decisions of *whether* to borrow and then *how much* to borrow (e.g., Cha & Weagley, 2002; Cha, Weagley, & Reynolds, 2005; Furquim, Glasener, Oster, McCall, & DesJardins, 2017). Double-hurdle regression is of the form:

Decision equation:

$$\Pr(Participate = 1|X') = \Phi\left(X_1'\beta_1\right) \tag{7.38}$$

Equation 7.38 is estimated via probit regression with a normally distributed error term. Then, the level equation is:

$$\begin{cases} y^* = X_2'\beta_2 \\ y = y^* \; if \; Participate = 1 \\ \varnothing if \; Participate = 0 \end{cases} \tag{7.39}$$

Equation 7.39 is estimated using truncated regression, because observations where *Participate* $= 0$ are excluded. The unknown parameters to be estimated, β_2, can differ from those in the decision equation (β_1), as can the included covariates. One can then analyze several outcomes: the probability of participation (Pr(*Participate*)); the conditional expected outcome (y^*); and the unconditional mean outcome ($y^* \ast$ Pr (*Participate*)).

Truncated regression can more generally be used to deal with truncation of data. Truncation occurs when the data generating process excludes "observations based on the characteristics of the dependent variable" (Long, 1997, p. 187). So while in the case of censored data we observe censored values of the dependent variable for some observations, truncated data reduces the analytic sample based on the dependent variable. Truncation may be a byproduct of sample selection (e.g., a study of family income for Pell eligible students) or other analytical choices. For example, in their study of student debt, Addo et al., (2016) excluded non-borrowers, thus truncating the dependent variable at some value greater than zero. The result of truncation is that the mean of the dependent variable is higher (in case of truncation from below) or lower (for truncation from above) than the "true" mean, and the variance of the truncated variable is smaller than that of the untruncated. Ordinary least squares regression can yield biased coefficients in the presence of truncation (Long, 1997). In these cases, researchers can use truncated regression, which is easily estimated in most statistical software (in Stata, the *truncreg* command). Truncated regression yields coefficients that can be directly interpreted as partial changes to y_i that is truncated at some value τ, just as in OLS regression, as seen in Eq. 7.40 (Long, 1997):

$$y_i = X_i'\beta + \varepsilon_i \, \textit{for all i such that } y_i > \tau \tag{7.40}$$

Truncation that results from sample selection can also be addressed by sample selection corrections, such as Heckman type sample selection correction, that use probit regression to model the likelihood of being in sample and incorporate the inverse Mills ratio into estimates of the observed data (Wooldridge, 2008). Instrumental variable techniques may also be brought to bear in such cases (see Bielby, House, Flaster, & DesJardins, 2013, for an overview of instrumental variable techniques applied to higher education).

7.6 Conclusion

Linear regression models have long been an essential part of an education researcher's statistical toolkit. Although the statistical foundations underlying the use of categorical dependent variable regression models have been around for many decades (see Cramer, 2003, for a history of the logit model, and Dey & Astin, 1993 for early work comparing and contrasting these models in higher education), they really became an important addition to the statistical tools used by higher education researchers in the middle to late 1980s. This is probably a result of many converging trends, such as the availability of these techniques in then available statistical software packages; the teaching of these methods in programs training education researchers; discussion of the use of the methods in higher education publications, including this Handbook (Austin, Yaffee, & Hinkle, 1992; Cabrera, 1994); and the increase of publications using these techniques in main higher education journals (Peng, So, Stage, & St. John, 2002). Given the ubiquity of these methods in higher education these days, having a solid understanding of their foundational statistical concepts and of their application is essential for conducting research into many important issues facing postsecondary education. More recently, limited dependent variable regression models, which have been employed successfully in other disciplines, are also increasingly being utilized in higher education research.

As demonstrated herein, these categorical and limited dependent variable models often remedy some of the statistical problems that arise when using traditional regression methods, such as linear regression, to study binary, multi-categorical, and limited outcome variables. But the application of these non-linear methods often come with a price, including complex estimation routines that are computer-memory intensive, and, importantly, additional complexities in the interpretation of results produced by such techniques. The former problem is of less concern with the advent of computers with multiple processors and high-capacity memory, lowering the time and memory resources needed to estimate such models. But interpreting the results of non-linear regression models remains a vexing problem for some, one that can be resolved by employing a variety of measures and using the graphical displays now available in many statistical software packages.

Our intention in producing this chapter was to update the published resources already available, to provide details about recent advances in the models used to study categorical and limited dependent variables, and to provide our colleagues with examples of how to use alternative ways to present and discuss the results produced by these regression methods. Our intention was not to provide a comprehensive treatment of the literature about these methods; to that end, we provide references to additional articles and books that can assist researchers in learning more about the underlying concepts and application of these methods.

To facilitate the educative goal of the chapter, we provided a running example of an important higher education issue that many readers should be familiar with: research on student college choice. Although the results produced by the applications of the various modeling techniques may inform the literature on college student choice, this empirical application was really designed for expository purposes. We used college choice as the exemplar because many treatments explaining non-linear models use examples that are not familiar to those in our field, such as applying the methods to medical research (for example, the work of Hosmer and colleagues).

We hope our efforts provide researchers with additional information about the application of the categorical methods described herein. In addition, we hope that our (brief) discussion of limited dependent variable models will encourage others to learn more about these methods and construct novel ways to apply them. We believe that using categorical and limited dependent models has, can, and will improve our collective understanding of many of the important issues facing higher education.

Appendix

```
/
*****************************************************************_
*********************************
These are examples of commands used to estimates the models in the
chapter.
The full code is not contained here for space constraints.
*****************************************************************_
*********************************/
**Set directories, open data, start log as needed.
*set macro vars
global $iv = " "

*enrl_college is the outcome variable we created.
```

```
**** Goodness of Fit ******
* unconditional model
logit enrl_college, or

*full model
logit enrl_college $iv
estimates store loges
predict loges, pr

*describing the pred probs
predict pprob5
set scheme s2mono
histogram pprob5, title("", color(black) margin(zero) size
(small)) ///
  xti("Predicted probabily", size(small)) graphregion(color
(white)) /// plotregion(color(white)) yti("Density", size(small))

summarize pprob5

*examining LR
fitstat

*examining classification
estat classification
lsens, title("", color(black) margin(zero) size(small)) ///
   graphregion(color(white)) plotregion(color(white)) xti(,size
(small)) yti(,size(small))
lroc, title("", color(black) margin(zero) size(small)) ///
   graphregion(color(white)) plotregion(color(white)) xti(,size
(small)) yti(,size(small))

/******LOGIT*****/
estimates restore loges
margins, dydx(*) post
estimates store loges_me

*graphing
estimates restore loges
margins , dydx(gpa) asobserved at(gpa=(1 (.25) 4))
set scheme s2mono
marginsplot, recastci(rarea) recast(line) ciopts(color(*.7)) ///
  graphregion(color(white)) plotregion(color(white)) ti("") yti
("Change in Pr(Enroll)", size(small)) xti("GPA, 10th grade",
size(small))
```

```
*look at a few populations of interest
mtable, rowname(1 Female first-gen low-inc ) ci clear at
(student_gender==2 parental_ed==1 family_income==(1 2) ) atmeans

/******PROBIT*****/
probit enrl_college $iv
estimates store probes
predict probes, pr
margins, dydx(*) post
estimates store probes_me

/******LPM *****/
regress enrl_college i.student_gender $dems $acad $expct $netwk
$sch
estimates store lpm
predict lpm, xb

*diagnostic of lpm
histogram lpm
set scheme s2mono
histogram lpm, title("", color(black) margin(zero) size(small))
///
xti("Predicted probabily", size(small)) graphregion(color(white))
plotregion(color(white)) yti("Density", size(small)) xline(0 1,
lstyle(foreground) lpattern("--"))

*plot residual v fitted
set scheme s2mono
rvfplot, yline(0, lstyle(foreground) lpattern("--")) graphregion
(color(white)) plotregion(color(white)) xline(0 1, lstyle(fore-
ground) lpattern("--")) xti(, size(small)) yti(, size(small))

*check for heteroskedasticity
estat imtest

***********************ORDINAL/MULTINO-
MIAL**********************
*pse_enroll_sel is the dependent var we created.

ologit pse_enroll_sel i.student_gender $iv, or
estimates store ord

*get some marginal effects
estimates restore ord
```

```
margins, dydx(gpa) post
estimates store ord_me

predict nocol_log lsel_log sel_log msel_log

*test if we need multinomial
oparallel, ic
brant, detail

***run it as multinomial
mlogit pse_enroll_sel $iv, rrr
estimates store multi

*get a marginal effect
margins , dydx(gpa) post
estimates store multi_me

*tests of IVs
estimates restore multi
mlogtest, lr
estimates restore multi
mlogtest, wald

*Test of categories - can we collapse them?
mlogtest, combine
estimates restore multi
mlogtest, lrcomb
estimates restore multi
mlogtest, hausman

*Interpretation
estimates restore multi
listcoef student_gender student_race_combo stugpa_10 stu_mathirt
apcred, gt adjacent

*Pred Probs for select subgroups
estimates restore multi
mtable if student_gender==2 & parental_ed==1 & family_income==1,
atmeans noci rowname(lowinc firstg) clear brief

************************COUNT************************
poisson apps $iv, irr
estimates store pois

estat ic
```

```
prcounts pois, max(20) plot
label var poispreq "Poisson"
labe var poisobeq "Observed"
label var poisval "# of apps"

nbreg apps $iv, irr
estimates store nb

estat ic

countfit apps $iv, nbreg prm
```

References

Addo, F. R., Houle, J. N., & Simon, D. (2016). Young, black, and (still) in the red: Parental wealth, race, and student loan debt. *Race and Social Problems, 8*(1), 64–76. https://doi.org/10.1007/s12552-016-9162-0

Allison, P. D. (2002). *Missing data: Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage.

Angrist, J. D., & Pishke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Archer, K. J., & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal, 6*(1), 97–105.

Arcidiacono, P. (2005). Affirmative action in higher education: How do admission and financial aid rules affect future earnings? *Econometrica, 73*(5), 1477–1524. https://doi.org/10.1111/j.1468-0262.2005.00627.x

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*(4), 726. https://doi.org/10.1037/0893-3200.21.4.726

Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. In J. C. Smart (Ed.), *Higher education: handbook of theory and research, VIII* (pp. 379–410). New York: Agathon Press.

Bahr, P. R. (2008). Does mathematics remediation work?: A comparative analysis of academic attainment among community college students. *Research in Higher Education, 49*(5), 420–450. https://doi.org/10.1007/s11162-008-9089-4

Bastedo, M. N., & Flaster, A. (2014). Conceptual and methodological problems in research on college undermatch. *Educational Researcher, 43*(2), 93–99. https://doi.org/10.3102/0013189X14523039

Bastedo, M. N., & Gumport, P. J. (2003). Access to what? Mission differentiation and academic stratification in US public higher education. *Higher Education, 46*(3), 341–359. https://doi.org/10.1023/A:1025374011204

Bastedo, M. N., & Jaquette, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. *Educational Evaluation and Policy Analysis, 33*(3), 318–339. https://doi.org/10.3102/0162373711406718

Baum, C. F. (2008). Stata tip 63: Modeling proportions. *Stata Journal, 8*(2), 299.

Belasco, A. (2013). Creating college opportunity: School counselors and their influence on postsecondary enrollment. *Research in Higher Education, 54*(7), 781–804. https://doi.org/10.1007/s11162-013-9297-4

Belasco, A. S., Rosinger, K. O., & Hearn, J. C. (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis, 37*(2), 206–223. https://doi.org/10.3102/0162373714537350

Bielby, R., House, E., Flaster, A., & DesJardins, S.L. (2013) Instrumental variables: Conceptual issues and an application considering high school coursetaking. In M. Paulsen (Ed.), *Higher education: Handbook of theory and research, XXVIII* (pp. 263–321). Dordrecht, The Netherlands: Springer.

Bielby, R., Posselt, J. R., Jaquette, O., & Bastedo, M. N. (2014). Why are women underrepresented in elite colleges and universities? A non-linear decomposition analysis. *Research in Higher Education, 55*(8), 735–760. https://doi.org/10.1007/s11162-014-9334-y

Blume, G. H. (2016). *Application behavior as a consequential juncture in the take-up of postsecondary education*. Doctoral dissertation, University of Washington.

Borooah, V. K. (2002). *Logit and probit: Ordered and multinomial models*. Thousand Oaks, CA: Sage.

Brasfield, D. W., Harrison, D. E., & McCoy, J. P. (1993). The impact of high school economics on the college principles of economics course. *The Journal of Economics Education, 24*(2), 99–111. https://doi.org/10.2307/1183159

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research, X* (pp. 225–256). Bronx, NY: Agathon Press.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.

Carnevale, A. P., & Strohl, J. (2013). *Separate and unequal: How higher education reinforces the intergenerational reproduction of white racial privilege*. Washington, DC: Georgetown University Center on Education and the Workforce.

Carnevale, A. P., & Van der Werf, M. (2017). *The 20% solution: Selective colleges can afford to admit more Pell grant recipients*. Washington, DC: Georgetown University Center on Education and the Workforce.

Ceja, M. (2001). Understanding the role of parents and siblings as information sources in the college choice process of Chicana students. *Journal of College Student Development, 47*(1), 87–104. https://doi.org/10.1353/csd.2006.0003

Cha, K.-W., & Weagley, R. O. (2002). Higher education borrowing. *Financial Counseling and Planning, 13*, 61–74.

Cha, K.-W., Weagley, R. O., & Reynolds, L. (2005). Parental borrowing for dependent children's higher education. *Journal of Family and Economic Issues, 26*, 299–321. https://doi.org/10.1007/s10834-005-5900-y

Chen, X., Ender, P., Mitchell, M. & Wells, C. (2003*). Regression with Stata*. Retrieved from https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/

Cheng, S., & Starks, B. (2002). Racial differences in the effects of significant others on students' educational expectations. *Sociology of Education, 75*(4), 306–327. https://doi.org/10.2307/3090281

Chung, A. S. (2012). Choice of for-profit college. *Economics of Education Review, 31*, 1084–1101. https://doi.org/10.1016/j.econedurev.2012.07.004

Clinedinst, M., Koranteng, A., & Nicola, T. (2015). *The state of college admission*. Arlington, VA: National Association for College Admission Counseling. Retrieved from: https://indd.adobe.com/view/c555ca95-5bef-44f6-9a9b-6325942ff7cb

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica, 39*(5), 829–844. https://doi.org/10.2307/1909582

Cramer, J. S. (2003). The origins and development of the logit model. In J. S. Cramer (Ed.), *Logit models from economics and other fields* (pp. 149–158). Cambridge, UK: Cambridge University Press.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software, 34*(2), 1–24. https://doi.org/10.18637/jss.v034.i02

DesJardins, S. L. (2002). An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher Education, 43*(5), 531–553. https://doi.org/10.1023/A:1020162014548

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education, 34* (5), 569–581. https://doi.org/10.1007/BF00991920

Doyle, W. (2007). Public opinion, partisan identification, and higher education policy. *The Journal of Higher Education, 78*(4), 369–401. https://doi.org/10.1080/00221546.2007.11772321

Dynarski, S. M. (2004). Does aid matter? Measuring the effect of student aid on college attendance and completion. *The American Economic Review, 93*(1), 279–288. https://doi.org/10.1257/000282803321455287

Eagan, K., Lozano, J. B., Hurtado, S., & Case, M. H. (2013). *The American freshman: National norms fall 2013*. Los Angeles: Higher Education Research Institute, UCLA.

Eagan, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., & Garibay, J. C. (2013). Making a difference in science education: The impact of undergraduate research programs. *American Educational Research Journal, 50*(4), 683–713. https://doi.org/10.3102/0002831213482038

Eliason, S. R. (1993). *Quantitative applications in the social sciences: Maximum likelihood estimation*. Thousand Oaks, CA: SAGE.

Engberg, M. E., & Allen, D. J. (2011). Uncontrolled destinies: Improving opportunity for low-income students in American higher education. *Research in Higher Education, 52*(8), 786–807. https://doi.org/10.1007/s11162-011-9222-7

Engberg, M. E., & Gilbert, A. J. (2014). The counseling opportunity structure: Examining correlates of four-year college-going rates. *Research in Higher Education, 55*(3), 219–244. https://doi.org/10.1007/s11162-013-9309-4

Engberg, M. E., & Wolniak, G. C. (2010). Examining the effects of high school contexts on postsecondary enrollment. *Research in Higher Education, 51*(2), 132–153. https://doi.org/10.1007/s11162-009-9150-y

Federal Student Aid, U.S. Department of Education. (2016). *Official cohort default rates for schools*. Washington, DC: Author. Retrieved from https://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.

Freeman, K., & Thomas, G. (2008). Black colleges and college choice: Characteristics of students who choose HBCUs. *The Review of Higher Education, 25*(3), 349–358. https://doi.org/10.1353/rhe.2002.0011

Furquim, F., & Glasener, K. M. (2016). A quest for equity? Measuring the effect of QuestBridge on economic diversity at selective institutions. *Research in Higher Education, 58*, 646. https://doi.org/10.1007/s11162-016-9443-x

Furquim, F., Glasener, K. M., Oster, M., McCall, B. P., & DesJardins, S. L. (2017). Navigating the financial aid process: Borrowing outcomes among first-generation and non-first generation students. *The Annals of the American Academy of Political and Social Science, 671*(1), 69–91. https://doi.org/10.1177/0002716217698119

Goldrick-Rab, S. (2006). Following their every move: An investigation of social-class differences in college pathways. *Sociology of Education, 79*(1), 67–79. https://doi.org/10.1177/003804070607900104

Gonzales, R. G. (2011). Learning to be illegal: Undocumented youth and shifting legal context in the transition to adulthood. *American Sociological Review, 76*, 602–619. https://doi.org/10.1177/0003122411411901

Gonzalez, J. M., & DesJardins, S. L. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education, 43*(2), 235–258. https://doi.org/10.1023/A:1014423925000

Greene, W. H. (2002). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hahn, E. D., & Soyer, R. (2005). Probit and logit models: Differences in the multivariate realm. *The Journal of the Royal Statistical Society, Series B*, 1–12.

Hart, N. K., & Mustafa, S. (2008). What determines the amount students borrow? Revisiting the crisis–convenience debate. *Journal of Student Financial Aid, 38*(1), 17–39.

Hillman, N. W. (2013). Economic diversity in elite higher education: Do no-loan programs impact Pell enrollments? *The Journal of Higher Education, 84*(6), 806–833. https://doi.org/10.1353/jhe.2013.0038

Hillman, N. W. (2014). College on credit: A multilevel analysis of student loan default. *The Review of Higher Education, 37*(2), 169–195. https://doi.org/10.1353/rhe.2014.0011

Horace, W. C., & Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters, 90*, 90321–90327. https://doi.org/10.1016/j.econlet.2005.08.024

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.

Howell, J. (2010). Assessing the impact of eliminating affirmative action in higher education. *Journal of Labor Economics, 28*(1), 113–166. https://doi.org/10.1086/648415

Hurtado, S., Inkelas, K. K., Briggs, C., & Rhee, B. S. (1997). Differences in college access and choice among racial/ethnic groups: Identifying continuing barriers. *Research in Higher Education, 38*(1), 43–75. https://doi.org/10.1023/A:1024948728792

Hurwitz, M. (2012). The impact of institutional grant aid on college choice. *Educational Evaluation and Policy Analysis, 34*(3), 344–363. https://doi.org/10.3102/0162373712448957

Ishitani, T. T., & McKitrick, S. A. (2016). Are student loan default rates linked to institutional capacity? *Journal of Student Financial Aid, 46*(1), 17–37.

Kelchen, R., & Li, A. Y. (2017). Institutional accountability: A comparison of the predictors of student loan repayment and default rates. *The Annals of the American Academy of Political and Social Science, 671*(1), 202–223. https://doi.org/10.1177/0002716217701681

Kim, J., DesJardins, S., & McCall, B. (2009). Exploring the effects of student expectations about financial aid on postsecondary choice: A focus on income and racial/ethnic differences. *Research in Higher Education, 50*(8), 741–774. https://doi.org/10.1007/S11162-009-9143-X

Kim, J., Kim, J., DesJardins, S. L., & McCall, B. P. (2015). Completing algebra II in high school: Does it increase college access and success? *The Journal of Higher Education, 86*(4), 628–662. https://doi.org/10.1353/jhe.2015.0018

Lin, T. F., & Schmidt, P. (1984). A test of the Tobit specification against an alternative suggested by Cragg. *The Review of Economics and Statistics, 66*(1), 174–177. https://doi.org/10.2307/1924712

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, TX: Stata Press.

McDonough, P. M. (1994). Buying and selling higher education: The social construction of the college applicant. *The Journal of Higher Education, 65*(4), 427–446. https://doi.org/10.2307/2943854

McDonough, P. M. (1997). *Choosing colleges: How social class and schools structure opportunity*. Albany: State University of New York Press.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine, 14*, 2143–2160. https://doi.org/10.1002/sim.4780141908

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. Thousand Oaks, CA: Sage.

Morrison, E., Rudd, E., Picciano, J., & Nerad, M. (2011). Are you satisfied? PhD education and faculty taste for prestige: Limits of the prestige value system. *Research in Higher Education, 52* (1), 24–46. https://doi.org/10.1007/s11162-010-9184-1

Myers, S. M., & Myers, C. B. (2012). Are discussions about college between parents and their high school children a college-planning activity? *American Journal of Education, 118*(3), 281–308. https://doi.org/10.1086/664737

Niu, S. X., & Tienda, M. (2008). Choosing colleges: Identifying and modeling choice sets. *Social Science Research, 37*(2), 416–433. https://doi.org/10.1016/j.ssresearch.2007.06.015

Norton, E. C., Wang, H., & Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal, 4*(2), 154–167.

O'Connor, N., Hammack, F. M., & Scott, M. A. (2010). Social capital, financial knowledge, and Hispanic student college choices. *Research in Higher Education, 51*(3), 195–219. https://doi.org/10.1007/s11162-009-9153-8

Office for Civil Rights, U.S. Department of Education. (2016). *Securing equal opportunity: Report to the president and secretary of education*. Washington, DC: Author. Retrieved from: https://www2.ed.gov/about/reports/annual/ocr/report-to-president-and-secretary-of-education-2016.pdf

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis, 56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Palardy, G. J. (2015). High school socioeconomic composition and college choice: Multilevel mediation via organizational habitus, school practices, peer and staff attitudes. *School Effectiveness and School Improvement, 26*(3), 329–353. https://doi.org/10.1080/09243453.2014.965182

Pallais, A. (2015). Small differences that matter: Mistakes in applying to college. *Journal of Labor Economics, 33*(2), 38. https://doi.org/10.1086/678520

Pampel, F. C. (2000). *Logistic regression: A primer* (Series Number 07-132). Thousand Oaks, CA: Sage.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*, 619–632. https://doi.org/10.1002/(SICI)1099-1255

Peng, C. Y. J., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in Higher Education, 43* (3), 259–293. https://doi.org/10.1023/A:1014858517172

Perna, L. W. (2006). Studying college access and choice: A proposed conceptual model. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research, XXI* (pp. 99–157). Dordrecht, The Netherlands: Springer.

Perna, L. W., & Titus, M. A. (2004). Understanding differences in the choice of college attended: The role of state public policies. *The Review of Higher Education, 27*(4), 501–525. https://doi.org/10.1353/rhe.2004.0020

Perna, L. W., & Titus, M. A. (2005). The relationship between parental involvement as social capital and college enrollment: An examination of racial/ethnic group differences. *The Journal of Higher Education, 76*(5), 485–518. https://doi.org/10.1080/00221546

Porter, S., & Umbach, P. (2006). College major choice: An analysis of person-environment fit. *Research in Higher Education, 47*(4), 429–449. https://doi.org/10.1007/sl1162-005-9002

Posselt, J. R., Jaquette, O., Bielby, R., & Bastedo, M. N. (2012). Access without equity: Longitudinal analyses of institutional stratification by race and ethnicity, 1972–2004. *American Educational Research Journal, 49*(6), 1074–1111. https://doi.org/10.3102/0002831212439456

Pryor, J. H., Hurtado, S., Saenz, V. B., Santos, J. L., & Korn, W. S. (2007). *The American freshman: Forty year trends*. Los Angeles: Higher Education Research Institute, UCLA. Retrieved from http://heri.ucla.edu/PDFs/40TrendsManuscript.pdf

Roderick, M., Coca, V., & Nagaoka, J. (2011). Potholes on the road to college: High school effects in shaping urban students' participation in college application, four-year college enrollment, and college match. *Sociology of Education, 84*(3), 178–211. https://doi.org/10.1177/0038040711411280

Rowan-Kenyon, H. T., Bell, A. D., & Perna, L. W. (2008). Contextual influences on parental involvement in college going: Variations by socioeconomic class. *The Journal of Higher Education, 79*(5), 564–586. https://doi.org/10.1353/jhe.0.0020

Scott, M., Bailey, T., & Kienzl, G. (2006). Relative success? Determinants of college graduation rates in public and private colleges in the US. *Research in Higher Education, 47*(3), 249–279. https://doi.org/10.1007/s11162-005-9388-y

Scott-Clayton, J. (2011). On money and motivation: A quasi-experimental analysis of financial incentives for college achievement. *Journal of Human Resources, 46*(3), 614–646. https://doi.org/10.3368/jhr.46.3.614

Smith, J. (2014). The effect of college applications on enrollment. *E. Journal of Economic Analysis & Policy, 14*(1), 151–188. https://doi.org/10.1515/bejeap-2013-0002

Smith, J., Pender, M., & Howell, J. (2013). The full extent of student-college academic undermatch. *Economics of Education Review, 32*, 247–261. https://doi.org/10.1016/j.econedurev.2012.11.001

Sribney, W. (n.d.). *Why should I not do a likelihood-ratio test after an ML estimation (e.g., logit, probit) with clustering or pweights?*. Retrieved from http://www.stata.com/support/faqs/statistics/likelihood-ratio-test/

Stratton, L. S., O'Toole, D. M., & Wetzel, J. N. (2007). Are the factors affecting dropout behavior related to initial enrollment intensity for college undergraduates? *Research in Higher Education, 48*(4), 453–485. https://doi.org/10.1007/s11162-006-9033-4

Taggart, A., & Crisp, G. (2011). The role of discriminatory experiences on Hispanic students' college choice decisions. *Hispanic Journal of Behavioral Science, 33*(1), 22–38. https://doi.org/10.1177/0739986310386750

Teranishi, R. T., & Briscoe, K. (2008). Contextualizing race: African American college choice in an evolving affirmative action era. *The Journal of Negro Education, 77*(1), 15–26.

Titus, M. A. (2007). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education, 48*(4), 487–521. https://doi.org/10.1007/s11162-006-9034-3

Wells, R. S., Lynch, C. S., & Siefert, T. A. (2011). Methodological options and their implications: An example using secondary data to analyze Latino educational expectations. *Research in Higher Education, 52*(7), 693–716. https://doi.org/10.1007/s11162-011-9216-5

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49*(4), 512–525.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2008). *Introductory econometrics: A modern approach*. Ontario, Canada: Nelson Education.