Journal of Student Financial Aid

Volume 26 | Issue 3 Article 1

12-1-1996

Econometric Modeling of Enrollment Behavior

Stephen H. Brooks

Econometric Modeling of Enrollment Behavior

By Stephen H. Brooks

Stephen H. Brooks is President of S.H. Brooks Co., Inc., in Massachusetts, an economic consulting firm that specializes in analytic and advisory services to higher education clients.

This paper describes using econometric modeling techniques to analyze the enrollment behavior of admitted students at four-year colleges and universities. It is based on the author's work over the last four years with approximately two dozen institutions seeking to understand more about the behavior of their admitted students–especially those requesting financial aid. This approach to modeling enrollment behavior has been discussed frequently in the literature and has been adopted or attempted at a number of institutions. Here, schools' experiences with such an approach are examined. We look at a wide spectrum of institutions differing by size, selectivity, and program focus.

Econometric models, and the analysis surrounding their creation, can provide institutions with valuable insights in three important enrollment management areas:

- Understanding the competitive marketplace within which an institution operates;
- Assessing how financial aid influences enrollment decisions; and
- Predicting student enrollment, financial aid cost, and the composition of an entering class.

However, it will come as no surprise to anyone familiar with the financing of higher education in the 1990s that the primary motivation for the use of these techniques has been financial. Administrators want to understand more about the role financial aid plays in the enrollment process and, with that knowledge, to insure that financial aid policy effectively and optimally supports institutional enrollment goals.

Enrollment management successes include:

- A large, regional, university discovered that it had been hurting itself by providing less than optimal aid to its first-year students. By expanding its financial aid budget, it was able to increase the size and quality of its class while increasing net revenue.
- A small liberal arts college was able to consolidate a large and unwieldy financial-aid packaging matrix into a few targeted categories. By eliminating a wastefully complex awarding policy, packaging time was shortened, and the effectiveness of the financial aid budget was increased.
- A college moved reluctantly from being ability-blind to a differential financial-aid packaging policy, using an econometric approach to focus aid dollars more systematically in areas that support enrollment goals.
- A college in a budget squeeze made more aggressive use of financial aid to achieve a net revenue goal. Econometric modeling also helped identify the limits of using financial aid alone to achieve net revenue goals.

Why Model Econometrically

The principal alternative to econometric modeling (for either predicting enrollments or assessing the impact of financial aid) is to build a table of historical yields. In these yields, students are grouped, for instance, by their ability and the size of their institutional grant. The historical yields become the best indicator of the yields for the coming year. The impact of alternative financial aid rules are estimated by seeing what would happen to enrollments if admitted students were moved from one grant cell to another. Table 1 shows a very simple example of such a table.

This tabular approach, while useful, has some serious limitations. Without testing statistically to determine the importance of each of these categories, this approach remains more or less arbitrary. There are countless ways to break up the class based on student characteristics or interests. Suppose that yields are different by sex, race, or major. One would need tables (similar to Table 1), broken down by Scholastic Aptitude Test (SAT) and grant size, for women and men, different ethnic groups, or different majors. As one splits the class up into ever finer cells with ever fewer students, confidence in the predictive accuracy of the yields in each cell falls fast.

In contrast, the econometric approach uses all of the information about each student and does not force one to break the admitted applicant pool into smaller and smaller slices to do it. It calculates the independent effect on yield of each student characteristic (college aid, SAT, grade point average [GPA], rank, ethnicity, geographic origin, expected major, etc.). It allows one to test statistically whether or not a particular characteristic "matters" in determining yields. This method also estimates the probability of each student's matriculation and can estimate the effect of changes in institutional aid awards on matriculation probabilities.

The econometric approach consists of three distinct phases that are discussed here:

- Data acquisition and verification,
- Model estimation and validation, and
- Model simulation and prediction.

Data Acquisition and Verification

It is impossible to overestimate the importance of getting the data right. The most frequent lament among researchers studying admission and financial aid is that "the data are difficult to work with." The author's experiences suggest that there is an inexhaustible supply of potential problems. A short list of problems could include: changing computer systems resulting in historical files that are potentially inconsistent with current data; changing field definitions within an existing historical data base; changing definitions of underlying concepts (did we calculate "family contribution" and "need" differently before the last Reauthorization?); changing procedures and approaches to entering original data; data entered at alternative points during the year (are these data as of June or as of October?); inconsistency in the way matriculant and nonmatriculant data are entered; insufficient care insuring the accuracy of underlying information during data entry.

Moreover, data concerns go beyond questions of accuracy and completeness. Suppose that all of the relevant data are accurate and complete somewhere

TABLE 1
Historical Yields by Financial Aid Status and Ability (in percent)

	No Aid Requested		Aid Requested		
Ability groups:	**No Grant**	**Merit Award**	**$ 0-5,000**	**$ 5,001-10,000**	**$ 10,001+**
SAT less than 900	20%	NA	27%	43%	65%
SAT: 900-1100	18	40	42	62	57
SAT: 1100-1300	15	35	34	40	43
SAT: above 1300	10	29	24	28	31

within the institution. The financial aid and admissions offices maintain databases to satisfy their own needs and purposes, which may or may not coincide with the needs and purposes of others, such as institutional researchers or faculty members. Different agendas imply different ways of looking at the data that may or may not be consistent with the way in which the data are currently maintained. The trick is to balance the competing needs of the different interest groups so that each has access to the most up-to-date, reliable, and accurate information possible.

What to Collect?

Although what is available and relevant will vary from institution to institution, the kind of information to gather can be grouped into several rough categories:

Ability

Measures of ability include SAT or American College Testing program (ACT) scores, class rank, GPA, academic honors or awards, and academic ratings (either subjective or objective) determined by the admissions office.

Financial Aid

Whether the applicant requested aid; if so, whether the student completed the application; the size and components of the financial-aid package; detail on negotiation of the award between the student/family and the institution; and whether the student received a merit award.

Demographics

Ethnicity, religion, sex, geographic origin, if the applicant is a son or daughter of an alumnus, the type of high school the applicant attends, family income, need, and family composition.

Interests

Major or program of study, extracurricular activities, clubs, career plans, etc.

Student Actions

Ideally, any action taken by the student in his or her interaction with the

institution ought to be recorded and maintained. The list would include whether the student was an early decision candidate; where else the student applied; who initiated the first contact and in what manner; how the student learned about the institution; what decisions were made by the college and how the student responded; and whether the student had an interview (either on-campus or off-campus).

"Real" Financial Aid And Income Concepts

It is important to look not only at the aid package offered to the applicant, but the amount of aid relative to the cost of attending the institution. Annual inflation, in tuition as well as other price levels, tends to distort trends in financial aid and makes year-to-year comparisons difficult unless the values are standardized in some way.

The preferred approach is to standardize on a single year's education cost, usually the most current academic year. This is done by dividing the actual values (whether they are aid, income, or need) by tuition and then multiplying the resulting ratio by the tuition in the chosen base year. For example, suppose a student received a grant of $8,000 in a year when tuition was $10,000; his or her award was thus 80% of that year's tuition. If today's tuition is $20,000, that original $8,000 award would be restated as $16,000 (80% of this year's tuition) so that it could be compared with this year's awards.

Other measures of inflation could be used to "deflate" the actual yearly values so that year-to-year comparisons could be made. The Consumer Price Index or other nationally-derived economic statistics are obvious candidates. However, to analyze decisions involving the purchase of a college education, deflation (using the institution's own cost of education) makes the most sense. This is because education is the single product that the financial aid award is designed to help purchase.

Nonmatriculant Data Is Often Elusive

A frequent and very significant data problem encountered is bias resulting from inconsistencies in the collection and reporting of matriculant and nonmatriculant data. Generally institutions have complete and detailed data on those who enroll. Data on the nonmatriculants are much more sketchy, particularly financial aid data. Many financial aid information systems require the deletion of data on awards that are not accepted. If the system does not have an ability to track the original offers, then all information on awards to students who choose to enroll elsewhere is lost. From a modeling standpoint, complete nonmatriculant data is every bit as important as matriculant data. Without it, there is no chance of capturing a complete and accurate profile of students who, for whatever reason, choose not to enroll. Finding out why they chose not to enroll is, of course, a major goal of any enrollment modeling exercise.

Occasionally some data will be available for nonmatriculants but the data may be less accurate, derived differently, or from a different point in time than for matriculants. Outside grants from private sources are an example of a data source that is often quite different for matriculants and nonmatriculants. For matriculants the data are typically updated and accurate as of fall registration: any outside grants that they have earned or been awarded will be recorded

accurately. But for nonmatriculants, some of whom have had no contact with the institution since early spring, the information is quite sparse. If any of those outside awards were received at high school graduation (after the last contact a nonmatriculant would have had with the institution), they will not be recorded.

This is just one example of a very prevalent and potentially quite severe bias problem in a data set. Careful attention to how information is received and recorded is critical in understanding the limitations of the research database.

Data Timing

In the best of worlds, a researcher would have copies of the admission and financial aid databases from several points during the admissions season: one when initial letters and awards are sent, another after the bulk of the negotiation has taken place (perhaps by the end of June), and another as of fall registration. Unfortunately, the reality usually falls short of the ideal. Typically the most accurate and complete data are as of fall registration.

Model Estimation

The models are constructed by testing the significance of all relevant variables from a merged admissions and financial aid database. The purpose is to see which variables improve the accuracy of the enrollment prediction. The model estimates the separate influence of each relevant factor–from SAT scores to geographic location to intended field of study. Special attention is, of course, focused on the role of financial aid offers.

Estimation Techniques

The modeling described in this paper is performed with a combination of ordinary least squares (OLS) and probit regression techniques. Because it is a much faster estimating technique, ordinary least squares is used for preliminary analysis of the data set to get a notion of which variables are most important in determining enrollment at a particular institution. Once the basic structure of the model has been identified using OLS, the models are reestimated and tested using probit regression.

Other researchers have used logit techniques to estimate enrollment models. There does not seem to be any clear winner in this estimation technique horse race. However, it is clear that the nature of the matriculation data (1 or 0 indicating whether the student enrolled or not) requires the use of some limited dependent variable technique.

Regression Factors

The models are built by testing the significance (the ability to improve enrollment predictions) of variables indicating ability, financial aid, interests, student actions, student demographics, etc. Coefficients on measures of ability (test scores, class rank, grade point average, etc.) will always have negative signs (regardless of the selectivity of the institution). This universal negative relationship between ability and enrollment probability comes from the fact that more able students will have more options and are, thus, less likely to enroll at any single institution. Moreover, in a world of competitive financial aid awards, these students will also have more generous competing financial aid offers from other institutions.

Experience has shown that the most significant variable describing finan-

cial aid is the amount of an institutional grant awarded. As noted above, data on outside grants are typically plagued by bias. The same is true for student loan and work-study data. For matriculants, loan and work-study data are usually what has been accepted as of fall registration. Nonmatriculant data are generally as of the initial spring offer. Ignoring this fact can often lead to unbelievable coefficient estimates suggesting that loans (or work-study) have a more powerful positive impact on enrollment than institutional grants.

Missing Data

The most significant econometric problem facing the researcher is the unavailability of data on competing financial aid offers from other institutions. Unfortunately there is no way around this fundamental problem. The coefficient estimates are thus likely to be less precise than they would be if the data on competing awards were available. This is a compromise that is inherent in the way the data are collected and maintained at all institutions. The models can often capture the impact of competing awards indirectly through information (typically from the financial aid application) about other institutions to which the student has applied. Student ability (for instance, the SAT score) can also play a role in capturing the impact of competitors' awards. More able students tend to receive higher awards from competing institutions.

This data failing is common to any approach that is based primarily on a single institution's own data collected over an admissions cycle.

Defining Sample Sets

The analyst must make some initial decisions about the scope of the sample set. Should there be one model for all admitted applicants or several models depending on financial aid status? Should there be separate models for aid-requesters and full-payers? Should engineering students be grouped with humanities students? Should women have a different model from men? Should the data set include early admits, or athletes, or tuition remission students?

In addition to answering these questions about individual student groups, there are similar sets of questions to be asked about grouping data across time. Should the model be built on last year's data alone, the last two or three years' data, or longer? How much change has there been over the last several years in the competitive marketplace faced by the institution? Does the model that describes the enrollment process of two years ago "work" well to explain enrollment last year?

Unfortunately there are no simple answers to these questions without reference to a particular data set. The analyst must test the reasonableness of the exclusion or inclusion of different sets of students and different years of data. This is done by using techniques that measure the validity of the assumption that behaviors (across student groups and across time) come from the same underlying model.

How to Handle Missing Observations

Frequently the modeler encounters a variable that is only available for a portion of the sample but that is clearly important in determining overall enrollment. The most obvious example is high school class rank. Many secondary schools do not provide class rank to college admissions offices. Lacking this information (often on a sizable fraction of the admitted applicant pool) poses a

TABLE 2
Enrollment Probability Equation

Dependent Variable: Probability of Matriculation
Periods: 1994, 1995
Number of Observations: 1,389

	Coefficient	T-Statistics
Constant	-0.0377	(-1.03)
Grants (in thousands) > $7,500	0.0220	(7.57)
Grants (in thousands) between $5,000 and $7,500	0.0286	(5.43)
Grants (in thousands) <= $5,000	0.0352	(3.92)
SAT (in thousands)	-.0214	(-5.42)
Appeal (denied)	0.1410	(3.63)
Appeal (granted)	0.1796	(4.28)
Asian	0.0670	(2.55)
Early decision	0.4236	(8.23)
Rank (actual)	0.5274	(3.82)
Rank (estimated))	0.6839	(4.38)
Resident of state 1	0.2223	(5.92)
Resident of state 2	0.1779	(5.83)
R-squared	0.154631	
Standard Error	0.44631	

dilemma: Should one restrict the sample to only those with class rank (along with everything else) or use the full sample and drop class rank as a determining variable? If class rank is shown to be an important explanatory variable (by using it in a regression restricted to students with valid ranks), then a secondary model should be developed to estimate the ranks of those who do not have them.

This procedure involves building a satellite model linking class rank to other variables that can be shown to play a role in determining class rank (SAT, grade point average, sex, type of secondary school, etc.). This model is then used to estimate a national class rank for those students who do not report rank but who do report SAT and GPA. That variable is used to fill in the holes for those with missing class ranks.

A Representative Enrollment Probability Equation

Table 2 displays the ordinary least squares coefficients from a very simple representative enrollment probability equation. The OLS coefficients are shown because they can provide a reasonably intuitive representation of the changes in matriculation probability in each variable. However, in actual prediction or simulation, the coefficients are reestimated using the probit technique.

The model was estimated over academic years 1994-95 and 1995-96. It uses two measures of ability: SAT and class rank. Class rank enters twice using reported data "Rank (actual)" and estimated data "Rank (estimated)." The positive coefficient on rank implies a negative relationship between enrollment and ability (as measured by rank) since higher class rank is actually measured as a lower percentage amount.

The grant variable enters in three separate projections, each with its own coefficient. The results from this analysis suggest strongly that additional

grants have a smaller impact on enrollment probability as the grant size gets larger. This institution noted whether the student tried to negotiate for higher grants and whether the appeal was denied or granted. Both granted and denied appellants were more likely to enroll than all other students (everything else being the same), but those whose appeals were granted were even more likely to enroll. This impact (entered as a "dummy" Boolean variable) measures whether a grant was given or denied, independent of the actual dollars granted–which is picked up directly in the grant variable.

In this equation, Asian students were more likely to enroll as were early decision admits. Residents of nearby states also had a higher enrollment probability.

The above equation is specific to a single institution. Its structure and coefficients could not be used in any way as predictors of enrollments at another institution.

Model Validation

Within-Sample Prediction Accuracy

The primary method for model validation during estimating is the accuracy of the model over the estimation sample. The traditional way to test the accuracy of limited dependent variable models is to pick a critical value for the estimated dependent variable. Then, assign a "yes" (in this case enroll) to all those whose estimated values lie above that value and a "no" (do not enroll) to all those below. Comparing these enroll/don't-enroll estimates with the actual enrollment decisions helps forecast the precision of the models.

Unfortunately, doing this will drastically lower the effectiveness and usefulness of the forecasts generated by the enrollment probability model. Suppose the critical value is the yield for that year's class (say 30%). Using the traditional approach, all those with estimated probabilities above 30% are forecasted to "enroll," and all those with probabilities below 30% are forecasted to "not enroll." But the model is actually predicting the enrollment probability of an admitted applicant, not whether he or she is going to actually enroll–an important distinction. A much more effective use of the model, for forecasting, is to group students by probabilities and estimate what percentage of each group will enroll, based on an average probability assigned to each group.

Enrollment Prediction

Suppose we take all of the students whose estimated enrollment probability lies between 40% and 50%. Rather than saying that all of those students will enroll (as the traditional approach would, because their probabilities lie above the critical value) the alternative approach says that 45% of them will enroll. Likewise, suppose we take all students with probabilities between 10% and 20%. The traditional approach says none of them will enroll. The alternative approach says 15% of them will.

The preferred alternative approach leaves the modeler with the task of having to pick which 15% of the students with probabilities between 10% and

20% will enroll and which 45% of the students with probabilities between 40% and 50% will enroll. The procedure adopted by the author involves several steps. For each probability range (10% to 20%, for instance), approximations of class composition are generated by randomly selecting the appropriate number of students some large number of times (typically 50). The 50 approximations are then used to derive estimates of expected class characteristics (average SAT, percent minority, financial aid costs). This process has the additional advantage of measuring variability by inspecting the distribution of the 50 separate outcomes. The forecasted and actual enrollment by the decile (the ten groups) for the one institution in this sample were very close.

The model predicts enrollment for each of the ten groups and then adds up the predictions to get total enrollment. Actual enrollment is simply the count of the matriculants whose estimated enrollment probability put them into that group. The figure is laid out by decile so that one can see where the errors occurred. Note that the high counts of students in the higher probability band primarily represent the early decision admits whose probabilities are very high. The overall within-sample error in this case was 2 students, far less than if the traditional "critical-value" approach had been used.

An additional approach to model validation is to assess the stability of the coefficients estimated over subsamples of the estimation interval. Significant variability in the coefficients suggests some potentially spurious correlation that may not hold up in the future. This, in turn, raises the question of over-parameterization.

Over-Parameterization

There is a reasonably significant potential for over-parameterization in the modeling of enrollment probability. Typically the researcher is provided with a very large data set, with sufficient degrees of freedom to allow using anything that might improve enrollment prediction. Many of these variables may "work" (in the sense of being statistically significant according to the usual criteria). In addition, there may be a strong temptation to model subgroups of the sample (by sex, major, or race) when the data do not necessarily warrant it.

Relentless efforts to find all the variables that "explain" this sample's enrollment behavior can easily lead one astray in the efforts to provide a useful tool for understanding financial aid and enrollment decisions. Ultimately we are looking for a general model that captures the underlying "truth" in the relationships among student characteristics, financial aid, and enrollment. Such a model is useful in planning the general design of award packages whether or not it can predict activity with utter precision across all years or among a subgroup of students in any one year. By "over-parameterizing" the model, one may end with an equation that is only useful in analyzing a specific year (indeed one that has already passed and is thus of little more than historical interest).

Using Models to Optimize Aid

The enrollment probability models can be simulated in many "what-if" exercises to see the effect selected changes in financial aid awards will have on

TABLE 3
Selected Simulation Results

Increase grants by $1,000 to:	Yield	Matriculated Students In Group	Matriculated Students Change	Net Revenue Change
All admitted financial aid requesters	35.3	251	13	-$46,000
Admit/deny with family income above $50,000	21.2	26	3	$20,000
SATs between 1300 and 1500	29.5	33	2	$ 1,000

enrollment and net revenue for a particular institution. For example, one could test whether or not a group of high-yielding students would still have matriculated had their grant awards been slightly smaller, or whether a group of low-yielding students would have matriculated had their awards been slightly higher. One could determine whether either of these changes would have generated additional net revenue, or whether the resulting changes in class composition would have helped or hurt enrollment goals.

Optimizing financial aid consists of many of these "what-if" experiments, trying to find groups of students for whom higher grants would raise enrollment enough to increase net revenue despite the higher apparent financial aid costs. In addition, we look for groups of students for whom reductions in grants would not lower enrollment so much as to cut net revenue.

Table 3 shows a simple example of several of these "what-if" simulations at a small New England liberal arts college.

In the first simulation, all financial aid requesters are given an additional $1,000 in institutional grants. If the simulated levels of enrollment and net revenue had been higher than actual levels, the institution could have been too stingy with its actual financial aid packaging. It could have attracted more students and increased its net revenue had it offered higher grants in the first place. In this case, the model estimates that the change would have produced 13 more matriculants, but a net revenue loss of $46,000. This amount was determined by the following calculations. The 13 new students would have brought with them a total of $205,000 in net revenue (tuition, fees, room and board less their grants). Tallied against this, however, is the $251,000 that would have been "wasted" by offering higher grants to those who would have enrolled anyway. The difference ($251,000 minus $205,000) yields a net loss of $46,000.

The second shows a more targeted example in which $1,000 in institutional grants are offered to students who reported more than $50,000 in family income and who had requested, but were denied, institutional aid. Twenty-six of these students had enrolled without the aid, so this action would have "wasted" $26,000. The additional aid would have boosted enrollment by 3 students who would have brought with them $46,000 in net revenue for a total gain of $20,000.

In the third example, students with SATs between 1300 and 1500 have

their institutional grants increased by $1,000. There were 112 applicants in this group, 33 of whom enrolled. As a result of the grant increase, 2 additional students would have matriculated. This would have "wasted" $33,000, but the additional two students would have brought with them $34,000 in net revenue for an increase of $1,000. Thus, this policy would have been virtually self-financing: slightly more financial aid would have generated slightly higher enrollment of a desirable group of students with no net financial aid cost.

A Delicate Balance

Running simulations to find groups of students for whom an institution may have under- or overspent will optimize financial aid. However, since changes in financial aid will inevitably result in enrollment changes, the optimization process also requires balancing any proposed award changes with institutional enrollment goals. The knowledge gained by these modeling, estimation, and simulation techniques can be very useful. It can enable institutions to make informed decisions about their financial aid budgets, including how to use those budgets to achieve their enrollment goals.

References

Ehrenberg, R. and Sherman, D. Optimal Financial Aid Policies for a Selective University. *Journal of Human Resources*, 1984, Vol. 19, pp. 202-230.

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

Moore, R.L., Studenmund, A.H., and Slobko, T. The Effect of the Financial Aid Package on the Choice of a Selective College. *Economics of Education Review*, 1991, Vol. 10, No. 4, pp. 311-321.