An Application of Binary Logistic Regression to College Admissions Data

Michael Sulock
Department of Mathematical Sciences
Montana State University

May 15th, 2009

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Michael Sulock

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Mark Greenwood
May 15th, 2009	Writing Project Director

1. Introduction

Accurately predicting the outcome of whether or not an accepted student will enroll at a particular school aids in efficient allocation of available limited resources, such as time and money. Knowing approximately how many students will enroll based on how many students are accepted can shape admissions policies, determine on-campus housing, and generally assist an institution in being more prepared to meet the academic needs for its incoming freshmen.

One of the most practical applications of the statistical sciences is the ability to create models using existing data in order to predict the outcome of future events. Statistical modeling is a data-based science that uses patterns in observed variables to provide an overall mathematical structure to real-world happenings. Topics such as global warming, consumer spending, and cancer risk have all been studied extensively using statistical models.

The general linear model is one of the oldest frameworks for the statistical modeling of an existing phenomenon. The more recent and similarly named generalized linear model (GLM) was developed to model commonly occurring responses that violate the distributional assumptions of ordinary least squares (OLS) regression.

This paper will use a specific generalized linear model, logistic regression, to formulate a model that can be used to make such predictions regarding college enrollment. The use of a generalized linear model is necessary, as the binary distribution of the results of whether or not a given student enrolls (a set of 0's and 1's) must be modeled using the binomial distribution. Binary data clearly do not follow a normal distribution, making ordinary least squares inadequate for dealing with such a response. A generalized linear model, on the other hand, can directly use the Binomial distribution to model the non-normal response.

2. The Generalized Linear Model

2.1 The Model

First formulated in a 1972 paper by Nelder and Wedderburn, the generalized linear model (GLM) is an essential part of modern statistics. Linear regression, ANOVA, probit, logistic regression, and Poisson log-linear models are all types of GMLs. Due to their ability to be applied in situations where the usual assumptions regarding the normal distribution for linear models are clearly not met, statisticians use GLMs for a wide variety of applications in a number of different fields, from medical to actuarial sciences.

As an example, suppose one is interested in developing a model for a situation with a response variable that takes on only two possible outcomes, termed binary or dichotomous. Important decisions such as determining whether cancer is present or absent from the results of a medical test, or predicting if a startup company will or won't succeed are examples of situations based on binary response variables. Ordinary linear regression requires an assumption that the response variable follows the continuous normal distribution with variance that is constant across all mean values. For modeling such dichotomous categorical responses, the binomial distribution becomes the appropriate assumed distribution (Hosmer, 1989). Allowing the response to follow a discrete distribution such as the binomial or Poisson cannot be done in an OLS model, but is possible with GLM.

For the GLM, the random component of response variable is assumed to follow any member in the class of distributions known as the exponential family, which is described later in this section. Important exponential family distributions for modeling frequently occurring processes are the Poisson, binomial, geometric, exponential, and normal distributions (Casella, 2002). The general framework for the model is given by the following equation (Hosmer, 1989)

$$g(\mu_i) = g(E(y_i|\tilde{x}, \theta)) = x_t'\beta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

As in any linear model, a GLM consists of a linear combination of the explanatory variables and their regression coefficients. A GLM has the distinction that this linear predictor is related to the expected value of the response through a function called the link function (Hosmer, 1989). This appears as the function g() in the above equation. The only necessary requirements for the link function are that it be monotonic and differentiable (Myers, 2002). Many different links are used, from the identity function for a normal linear OLS model, to an exponential function. The logarithmic function is a simple example of a frequently used non-linear link function:

$$g(\mu_i) = log(\mu_i)$$

Closely related to the log function is another non-linear link function, the logit, described in detail in the following section. Logistic regression uses the logit as its link function. The ability to use both linear and non-linear link functions enable researchers to model a variety of relationships between the mean of the response and predictors. As a result, the GLM is a unification of linear and non-linear models (Myers, 2002).

As previously mentioned, another important aspect of the GLM is that it allows the assumed distribution of the response to be any distribution included in the exponential family (Hosmer, 1989). Given parameters θ and data, x, the exponential family is defined as any variable with a probability density function (or mass function), as follows

$$f(x|\theta) = h(x)c(\theta)\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}.$$

Where h and t are functions depending solely on the data, and c and w are functions depending solely on the parameters (Casella, 2002).

Regression coefficients are estimated in a GLM by the method of maximum likelihood (Ramsey, 2002). Generally speaking, the maximum likelihood estimate for a given parameter is the value making the observed data most likely. The likelihood function calculates the probability that the parameter takes on a particular value given the observed data. The maximum likelihood estimate is obtained by finding the value that maximizes this likelihood function (Casella, 2002). As a result, the method of maximum likelihood can be applied to a variety of situations with different distributional assumptions as long as the likelihood function can be explicitly stated.

Depending on the distribution specified for GLMs, it is possible to model responses where the variability depends on the mean, unlike models created under the assumption of the normal distribution that require the variance to remain constant for all mean values. Again suppose that Y is a dichotomous response variable and follows a Bernoulli distribution, a simple distribution describing the outcome of a single observation of a binary random variable. Denote the probability of success for an individual observation equal to π. It can be shown that (Casella, 2002):

$$E(Y) = \pi$$

and

$$Var(Y) = (\pi)(1-\pi).$$

These results make it clear that the variance is not constant, but rather is a function of the mean. For binary outcomes, the Bernoulli distribution describes the outcome of a single observation with a given probability of success. The binomial distribution can be used to model the counts of successes of independent and identically distributed Bernoulli trials for a fixed number of observations. As a result, the Bernoulli distribution is special case of the Binomial distribution where the number of trials is equal to one.

As another example of non-constant variance, for modeling the probability of the number of events occurring for a fixed amount of effort, the response is often assumed to follow a Poisson distribution. A random variable with a Poisson distribution has variance exactly equal to the mean, or the expected value (Casella, 2002). Therefore, with a response variable assumed to follow either the Bernoulli, binomial, or Poisson distribution, the variance is a function of the expected value. By using such distributions, a GLM can allow for the variance of the response to change with the value of the mean, whereas an OLS model assumes the variance in the response remains constant.

2.3 The Logit Function

Probabilities are commonly used to describe the likelihood of random events based on the long-run frequency of occurrence. An alternative expression is the use of odds. Denote the probability of a successful outcome for a given binary event as π. By definition, the odds in favor of a success is

$$\frac{\pi}{1-\pi}.$$

In other words, it is the ratio of the probability of a successful outcome to an unsuccessful one. For example the odds in favor of rolling a six with one roll of a fair six-sided die is

$$\frac{\pi}{1-\pi} = \frac{\frac{1}{6}}{1-\frac{1}{6}} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5} = 1:5.$$

The interpretation is that that the event of rolling a six is five times less likely than the event of not rolling a six. The odds in favor, or against, a certain outcome is a concept often used in the context of gambling.

The logit function, in the context of a binary outcome, is defined simply to be the log of the odds (Hosmer, 1989).

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right).$$

The logit function serves as the link function in the logistic regression model.

3. Logistic Regression

3.1 Overview

The logistic regression model is a commonly implemented GLM, fit to a response that is assumed to follow a binomial distribution. Such responses include an individual binary response (Bernoulli trial) or responses that are binomial counts, meaning they are the counts of successes of independent Bernoulli trials for a group of a known size. Logistic regression can also be extended to model multinomial counts. This paper will deal strictly with its binomial applications.

Binary logistic regression on an individual assumes the response variable conforms to the following distributional assumption:

$$Y_i \sim Bin(m, \pi_i) \; where \; i = 1, ..., n \; and \; m = 1 \; \forall i$$

In other words, the individual observations of the response are Bernoulli trials, each with a potentially distinct probability of success (Casella, 2002). These are two of the most widely used applications for a GLM. Often these individual binary outcomes are indicators of a "success" or "failure". With this type of data, binary logistic regression is frequently used to estimate the probability of a successful outcome. Binomial logistic regression is also used for modeling the success for a given sample size of identically distributed Bernoulli random variables (Casella, 2002)It is in these practical and simple applications that logistic regression has been found to provide a myriad of uses in a broad range of fields such as biology, epidemiology, and economics.

2.2 The Logistic Regression Model

For a logistic regression model using *p* different explanatory variables, denote the probability of a success for a given set of realizations of these *p* variables as $\pi(\tilde{x})$. If g() is the logit function, then the framework of the model is given by the following equation (Hosmer, 1989)

$$g(\pi(\tilde{x})) = log\left[\frac{\pi(\tilde{x})}{1-\pi(\tilde{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

The logit function links the linear predictor to the expected value of the response. Like all link functions, the logit is one to one and is therefore invertible. By solving for $\pi(\tilde{x})$ the

relationship between the linear predictor and the probability of success is evident. This equation is known as the logistic function (Hosmer, 1989):

$$\pi(\tilde{x}) = \frac{e^k}{1+e^k} = \frac{1}{1+e^{-k}}$$

where

$$k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

According to basic axioms of probability, it is clear that $\pi(\tilde{x})$ cannot be negative or greater than one. Therefore an appropriate model will restrict the possible values of $\pi(\tilde{x})$ to this interval. The logistic function achieves this as its domain is equal to the real numbers, while its range is restricted to values between zero and one (Hosmer, 1989).

Another reason why logistic regression is a good model for many different real-world phenomena is the shape of the logistic curve. It is an S-shaped curve, well-suited for modeling many observed processes such as population growth that are characterized by rapid change during the middle values of the argument between periods of slower change at the extremes.

Figure 3.0 illustrates this S-shape by graphing a simple case of the logistic function where k = x.

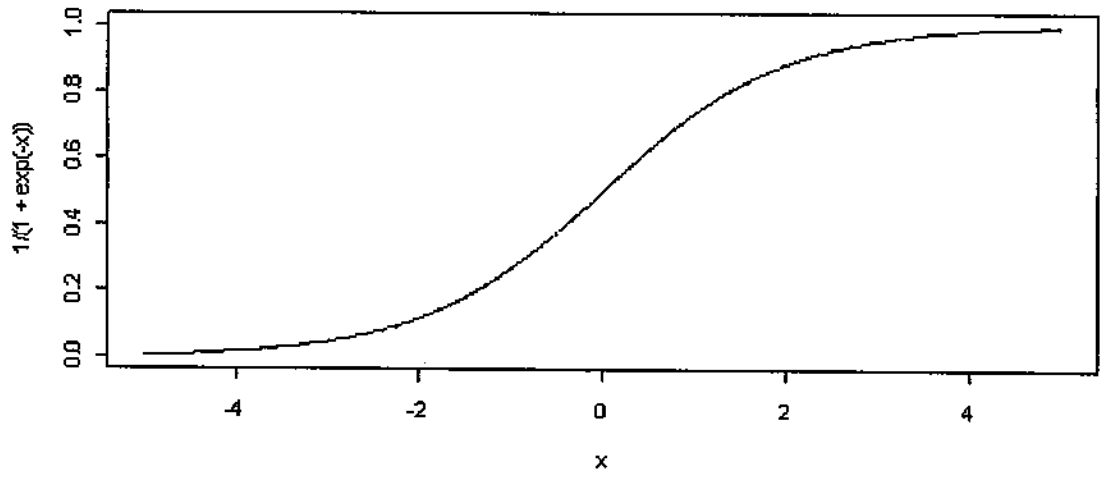

Figure 3.0: Curve of the logistic function

In contrast, using OLS to model $\pi(\tilde{x})$ makes this probability exactly equal to a linear combination of the predictor variables.

$$E(\pi(\tilde{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The OLS model where the mean response is the probability of a success has no restrictions on its output, and therefore could lead to nonsensical predictions, such as predicted probabilities greater than one (Hosmer, 1989). Also, an OLS model assumes that the random error, after accounting for the above mean structure, is normally distributed. This is clearly not correct when the only values the response can take are 0 and 1.

3.3 Assumptions

For a researcher to be confident in any inferences made with logistic regression, certain assumptions must be met. A major assumption shared with OLS is that the observations need to be independent from one another. Also, as with an OLS model,

logistic regression requires a linearity assumption. While OLS assumes the covariates to be linearly related to the mean of the actual response itself, logistic regression assumes that the linear combination of the independent variable is linearly related to the logit of the mean response, or the log odds in favor of a success (Myers, 2002).

For continuous explanatory variables there are a limitless number of possible values, and as a result, multiple observations sharing the same value is unlikely. Binning a continuous variable into categories enables one to estimate the proportion of success in each category. Therefore the probability of a success for a certain range of values can be estimated using the proportion of successes in the response for each corresponding range of the explanatory variable (Hosmer, 1989). The assumption of linearity of the log odds can be checked for each continuous variable individually by transforming these estimated probabilities to the log odds scale, called empirical logits.

3.4 Interpretation of Logistic Regression Coefficients

In addition to calculating the estimated probability of a success, the regression coefficients from a logistic model provide information about how each covariate is related to the response. As noted previously, the logistic regression model equates the linear combination of the explanatory variables and their coefficients to the log odds. Therefore, the regression coefficient for a given variable is the change in the log odds for a one unit change in that particular explanatory variable, assuming all other variables are held constant (Allison, 1999).

Since the change in log odds is a difficult quantity to give a meaningful interpretation to, regression coefficients are often transformed to produce a more applicable value, an odds ratio. An odds ratio is the ratio of the odds of an event occurring for a given realization of the variable to the odds of the same event occurring when the realized value is increased by

one unit. Since logistic regression often models the probability of a success given particular values of the explanatory variables, the logistic regression model can be used to derive the odds ratio for changing values of a particular explanatory variable.

As a simple example, suppose you want to determine the odds ratio when x_1 is an indicator variable for group and takes on the possible values of 0 and 1. For this example assume that all other independent variables are held constant. By manipulating both sides of the logit function we can see that the odds of a success when $x_1 = 1$ is given by

$$e^{log\left[\frac{\pi(\tilde{x})}{1-\pi(\tilde{x})}\right]} = log\left[\frac{\pi(\tilde{x})}{1-\pi(\tilde{x})}\right] = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p},$$

and therefore the odds of a success when $x_1 = 0$ is

$$e^{\beta_0+\beta_1(0)+\beta_2 x_2+\ldots+\beta_p x_p}.$$

Finally, the odds ratio of these two particular groups is given by

$$\frac{e^{\beta_0+\beta_1(1)+\beta_2 x_2+\ldots+\beta_p x_p}}{e^{\beta_0+\beta_1(0)+\beta_2 x_2+\ldots+\beta_p x_p}} = e^{\beta_1}.$$

Notice that all the other terms in the exponent cancel because only the value of x_1 was changed. Therefore, calculating e^{β_1} results in an odds ratio and therefore can be interpreted

as how much more likely (if $e^{\beta_1} > 1$), or less likely (if $e^{\beta_1} < 1$), the outcome of interest is depending on the group that observation falls into (Allison, 1999).

3.5 Prospective vs. Retrospective Studies

Generally, there are two types of approaches when collecting binary data. To illustrate these types, suppose one is interested in using the presence of or absence of lung cancer as a response variable. The first approach is a prospective study randomly selecting individuals with certain values of explanatory variables, such a frequency of smoking, and watching them over time to determine whether or not they develop cancer. In this situation the number of successes and failures is not fixed by the researcher. The second approach is a retrospective study, where the individual patients are chosen based on having lung cancer or not. Thus, the number of successes for a retrospective study is fixed by the researcher, whereas in a prospective study it is not. Retrospective data has the advantage of being cheaper and easier to collect, as a potentially lengthy study involving observing subjects over time is not necessary. The disadvantage to retrospective data is that prospective probabilities cannot be estimated.

While probabilities regarding future observations can only be made using prospective data, the odds ratio is the same regardless of whether the study is prospective or retrospective. Furthermore, it is the only calculable quantity that has the same interpretation whether one is referring to events that have already happened or events that are yet to happen (Ramsey, 2002). As we have seen, for a logistic model, an odds ratio is a simple function of the regression coefficients for explanatory variables, and thus inference can be made for both types of studies using logistic regression. Since making predictions from a model involves estimating a probability, not simply an odds ratio, you cannot estimate the probability of an event happening in the future from a retrospective study.

4. Validation

4.1 The ROC Curve

A receiver operating characteristic (ROC) curve is a graphical, nonparametric method for measuring the accuracy of a predictive model. First used by radar operators in World War II to correctly classify whether a given signal was a plane or just random noise, ROC curves are a way to evaluate and compare the ability of predictive models to correctly classify observations from a set of given predictors. Although ROC curves can be constructed using a variety of model validation techniques, this paper will implement the cross-validation technique used by SAS in PROC LOGISTIC, a leave one out process.

The estimated probability of success produced by the logistic function is a continuous variable. In order to use this continuous probability to classify an observation as one of two possible outcomes, a predetermined cutoff value must be specified. An estimated probability above this cutoff will result in a prediction of a success. An estimated probability below the cutoff will result in a prediction of a failure (SAS Institute Inc., 1995).

4.2 Binary Classification

For understanding the mechanism behind an ROC curve it is helpful to start with a confusion matrix. A confusion matrix is simply a truth table placing the outcome of a prediction into its possible categories. Confusion matrices come in several forms. Only the 2 x 2 case categorizing binary outcomes will be examined, as it is the basis for constructing the ROC curves corresponding to our candidate models.

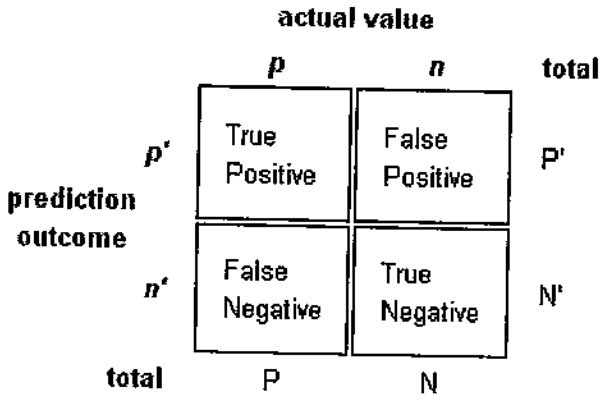

Figure 4.0: Confusion matrix for a binary outcome

The labels on top row of figure 4.0 refer to the truth of whether an observation is positive (a success) or negative (a failure). The labels on the left side correspond to a prediction from the model of either success or failure. An event correctly classified as positive is referred to as being a true positive outcome. Likewise, an event incorrectly classified as positive is referred to as being false positive. For negative predictions, a correct prediction is classified as being true negative, and an incorrect prediction is false negative.

A ROC curve looks at how the predictive power of a particular model changes as the cutoff value for making predictions is altered. As mentioned previously, the cutoff value is necessary to transform a continuous estimated probability of success into a binary outcome. In particular, a ROC curve examines the tradeoff between correctly and incorrectly classifying positive outcomes. For a given cutoff value, the proportion of actual positives that the model correctly categorizes is known as the sensitivity.

$$sensitivity = \frac{true\ \ positives}{true\ positives + false\ negatives}$$

Similarly, the proportion of actual negatives that the model correctly classifies is known as the specificity (SAS Institute Inc. 1995).

$$specificity = \frac{true\ \ negatives}{true\ negatives + false\ positives}$$

A ROC curve is simply a connected line graph of the sensitivity vs. (1 – specificity) for different values of the cutoff point. Another interpretation is that a ROC curve plots the true positive rate against the false positive rate. The unit square shown in figure 4.1 graphing the sensitivity vs. (1-specificity) is the sample space for a ROC curve.

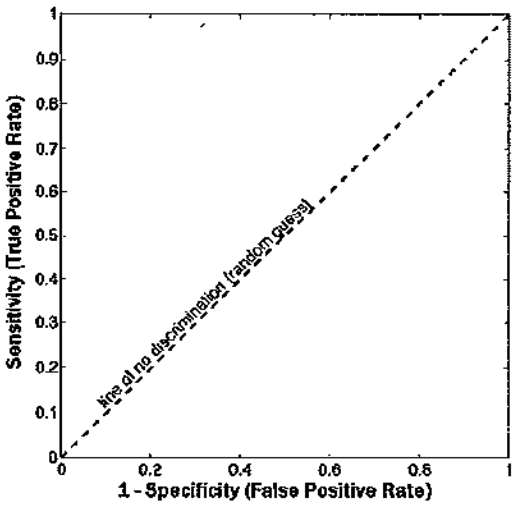

Figure 4.1: The ROC Curve sample space.

Simply randomly classifying points for various cutoffs results in a 45° line across the ROC space. For an illustration of this, imagine the cutoff point for any chosen model is a probability of zero. As a result, every observation will be classified as a success and the sensitivity will be equal to one. However, the specificity will be zero as no actual negatives are correctly identified. This results in a point on the uppermost right-hand corner of the

ROC space. A similar argument illustrates that setting the cutoff point to one results in a point on the lower left hand corner of the ROC space.

4.3 Model Evaluation

The best predictive models are those with both high sensitivity and high specificity for a given cutoff value. This results in a point towards the upper left-hand corner of the ROC space. Therefore, the higher a ROC curve is above the 45° line, the more accurate that model is at making predictions for different cutoff values. Area under the curve (AUC) is a statistic that evaluates individual models and can compare models to one another. The closer an AUC value is to one for a given model the more predictive accuracy that model has. The interpretation of the AUC itself is that for a given model, if one randomly selects an individual with a observed response corresponding a success, and an individual with an observed response of failure, the AUC is the probability that the model assigns a higher predicted probability of success to the individual with the observed success than to the one without. (Delong, 1989)

5. Application of Binary Logistic Regression

5.1 A College Enrollment Problem

Every year colleges in the United States undergo a process of collecting applications from potential students and deciding which individuals to accept. Out of these individuals that are accepted, only a subset actually enrolls. Colleges are interested in controlling the number of incoming students each year, as either too many or too few attending students can result in substantial problems such as a decrease in government funding or a lack of on-

campus housing for the incoming students. Therefore, a model for estimating the probability that a student who is accepted will enroll is a valuable asset.

Logistic regression will be implemented for modeling the probability than an accepted student at the University of North Carolina at Asheville (UNCA) will actually enroll based set values of the explanatory variables for that student. The data set was generously provided by Archer Gravely, the head of institutional research at UNCA. The focus of this paper is limited to the prediction of whether or not a student will attend, and not on interpreting the regression coefficients for particular explanatory variables included in the model.

5.2 The Data

This paper will use a data set collected from all applicants to UNCA during the period of 2005 to 2008. The response is binary, with categories corresponding to whether or not the student enrolled at UNCA. The data set includes 14 possible covariates, ranging from categorical variables such as race, sex, and NC residency status to continuous variables such as high school GPA, SAT math score, and ACT score.

A possible hindrance to the stated goal of making an accurate, usable predictive model is that many data values are missing from the UNCA data. For example, often students did not take the both the SAT and ACT, or a student came from a high school where class rank is not computed. The missing data is fairly widespread. Out of 7158 individuals in the entire data set, 6598 of them have at least one missing value for the 14 possible independent variables. Further complicating matters, the proportion of missing values for different variables varies widely. For example, only 56 students did not report a total SAT score, while 5621 students did not report an ACT score.

While it is true the data set contains a significant proportion of missing data, many combinations of the possible covariates are likely to contain similar information and therefore have significant correlation. Examples of this include GPA with weighted GPA, and total SAT score with ACT score. Missing information between correlated variables is less significant than between independent variables, as the information contained in the missing variable may still be present in any correlated variables. A more significant problem exists with any individual missing all the variables that contain the same type of information. Missing data can lead to a biased model if there are systematic differences between those students who provided a value for a particular variable and those who did not.

5.3 Types of Missing Data

Generally speaking, there are three possible types of missing data. The easiest type to work with is data missing completely at random (MCAR). These data exist when missing values occur randomly among all observations. In other words, every possible observation has an equally likely chance to be absent. A fortunate result of MCAR data is that since the missing data is a random sample of the entire sampled population, no systematic differences exist between those individuals with missing values and those without. Therefore MCAR data can be ignored with little or no effect on the accuracy of a particular model. Unfortunately, true MCAR data is a relatively rare phenomenon.

A second category is data missing at random (MAR). Data are said to be MAR if the probability of a value being absent is distributed uniformly among a subset of the entire sample. Within each of these subpopulations, the probabilities of any particular data value to be missing are equal. However, these probabilities are not equal among every individual in the sample. An example of MAR data would be if all females in the UNCA data set are more likely to have taken the ACT and report their scores than the males.

The third type of missing data, and the hardest to correct for, is non-ignorable missing data. Here, the individuals with missing data points are systematically different from those without missing data points. However, the cause of the missing data cannot be fully explained by individuals belonging to certain subsets of the sample. Non-ignorable missing data is likely to create bias in any results and is extremely difficult account for (Cherry, 2008).

5.4 Exploratory Data Analysis

The full UNCA data set has 4866 individuals with complete information for the eight variables as well as the response. Using these individuals and these nine variables, a new data set is constructed that will be used for the analysis in the remainder of this paper. The continuous variables with complete information are high school percent rank, high school yield (the proportion of individuals from a particular high school who enrolled at UNCA over the last 10 years), weighted high school GPA, and scores on both the math and verbal sections of the SAT. Three categorical variables are also used, the first being geographic region of residence, with three levels for inside of NC (Eastern, Piedmont and Western) and one pertaining to out of state. The final two categorical variables considered are sex and race.

For our modeling purposes, the population of interest will be seen as any student with complete information for these nine variables. Within the data set corresponding to this population, no individuals were left out, and the data can be seen as a representative sample. Therefore, any models made from this dataset can be used to make inference on the entire population.

Examining the distribution of the response over each of the categorical variables is important as having too few individuals for any particular combination can cause problems in the numerical methods used to calculate the regression coefficients for the logistic model.

In particular, any variable that has a category with all successes or all failures will lead to serious problems for the iterative numerical techniques used to estimate regression coefficients (Hosmer, 1989). For example, preliminary model building resulted in convergence problems as the category for RACE pertaining to Native Americans had fewer than 10 observations total across all four years. To avoid this problem, this category was condensed into "other".

Figures 3.1, 3.2, and 3.3 each display a possible categorical explanatory variable with color coding according to whether or not a student enrolled. These figures are further divided according to year. Year was not included as a potential variable, as the goal is for a predictive model. However, examining the distribution of each variable separated by year provides a quick visual check on any possible strong differences between years.

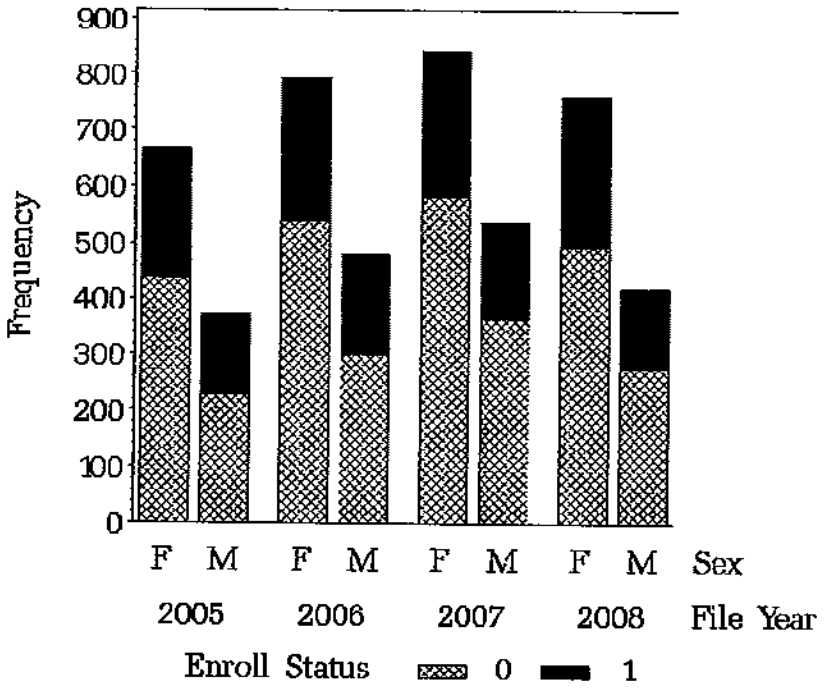

Figure 5.1: The Distribution of gender separated by enrollment and grouped by year.

From figure 5.1, the main difference in the distribution of gender between years is the total number of individual applicants, with 2007 having the most and 2005 with the least. Proportionally, the distribution of enrolling students between males and females is fairly similar across the four years the data was collected. Also, within each gender, the proportion of enrolling students appears to be relatively similar, casting doubt on the usefulness of using sex in to predict whether a given student will enroll.

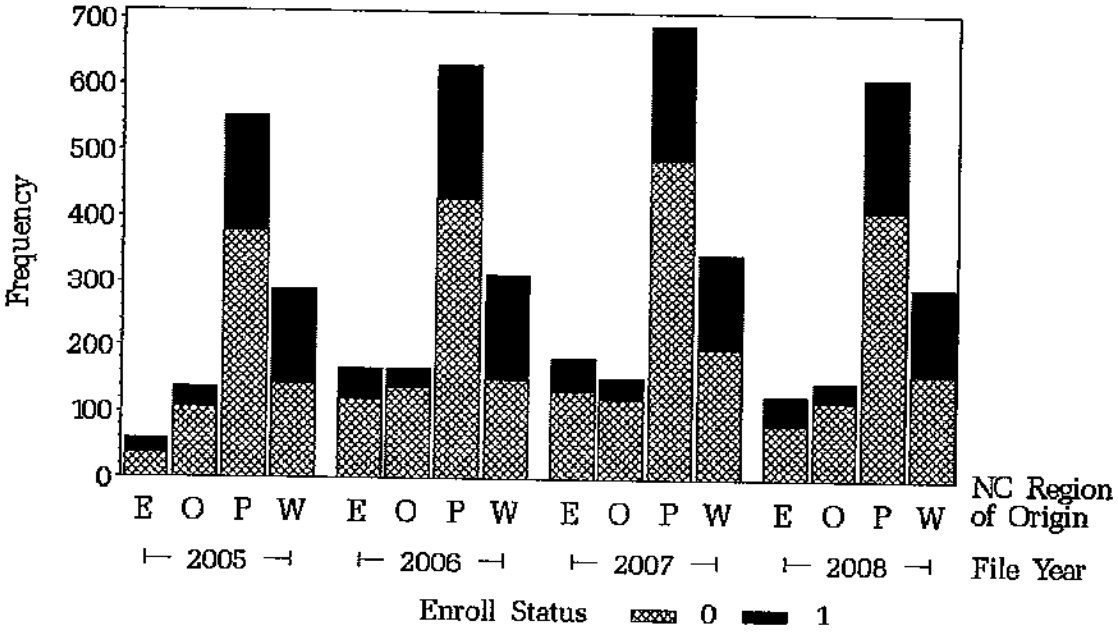

Figure 5.2: Region of origin separated by enrollment and grouped by year. The categories for within North Carolina are E = Eastern, P = Piedmont, and W=Western. The 'O' corresponds to out of state.

From figure 5.2, the distribution of region of origin has slightly more variability from year to year than gender does. Most notably, the proportion of applicants from the eastern part of North Carolina was substantially lower in 2005 than the other three included years.

The distribution of enrolling students within a given region is relatively homogenous from year to year. Notably, the proportion of accepted students from western North

Carolina for any given year is substantially higher than for any other region. This is an indicator that region may be a significant predictor of whether or not a student enrolls.

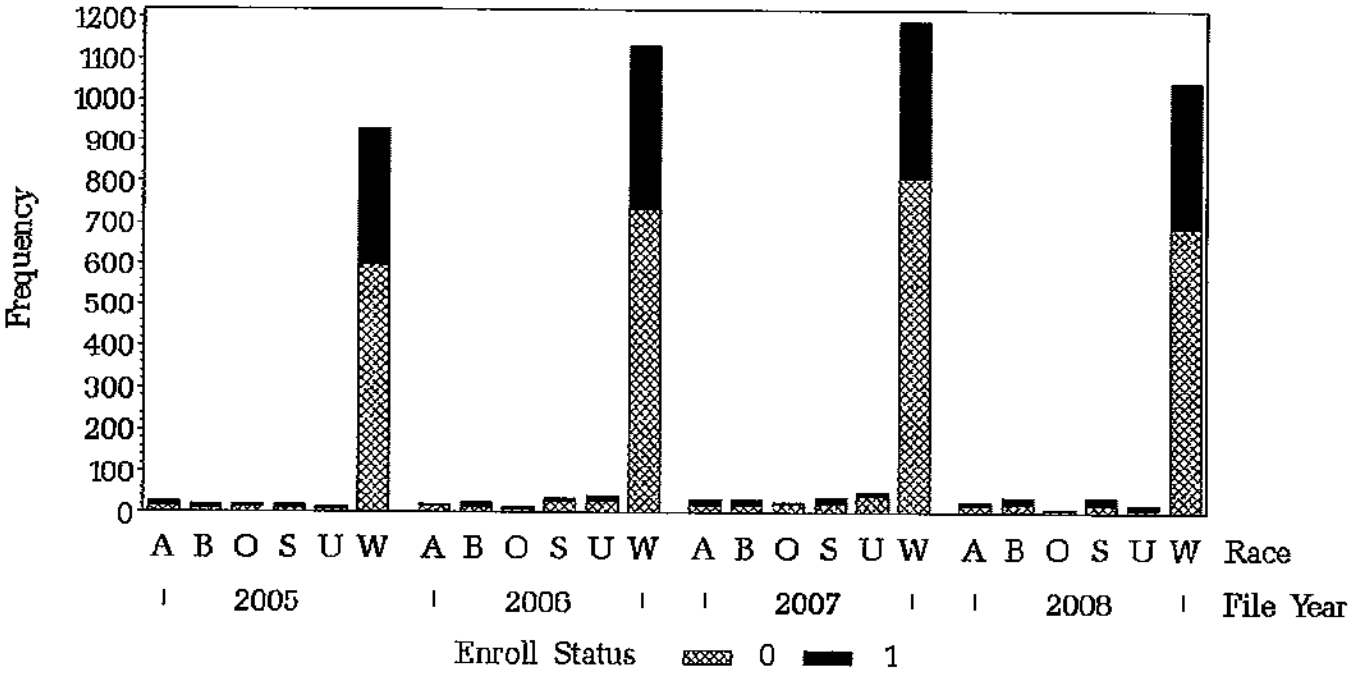

Figure 5.3: Race separated by enrollment and grouped by year. The categories are A=Asian, B=Black, O=Other, S = Hispanic, U=Not Reported, and W=White.

The trend for the distribution of race that is immediately evident is the predominance of white applicants. Every year has at least 900 white applicants, with no other race having more than 50 applicants. Similarly to gender, the year to year differences in the distribution of race by enrolling appear minimal. Due to the extremely low level of non-white applicants, a visual comparison of the distribution of enrolling students between the different races is difficult. As a result, figure 5.3 provides little information as to the significant of including race in a predictive model.

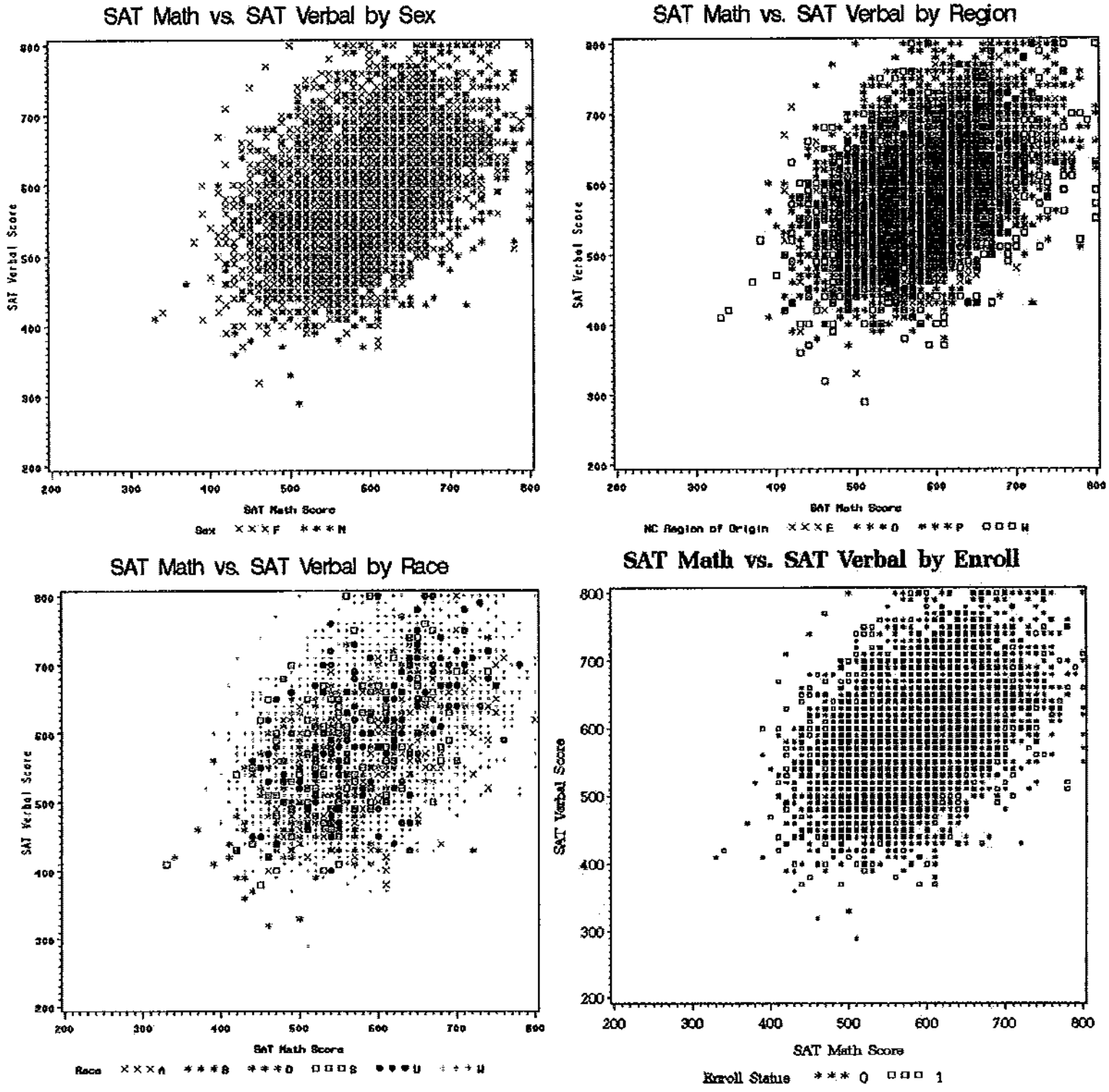

Figure 5.4: Math SAT score vs. Verbal SAT score separated by each categorical variable. Starting at the top left and moving clockwise, the categorical variables used are sex, region of origin, whether or not the student enrolled, and race.

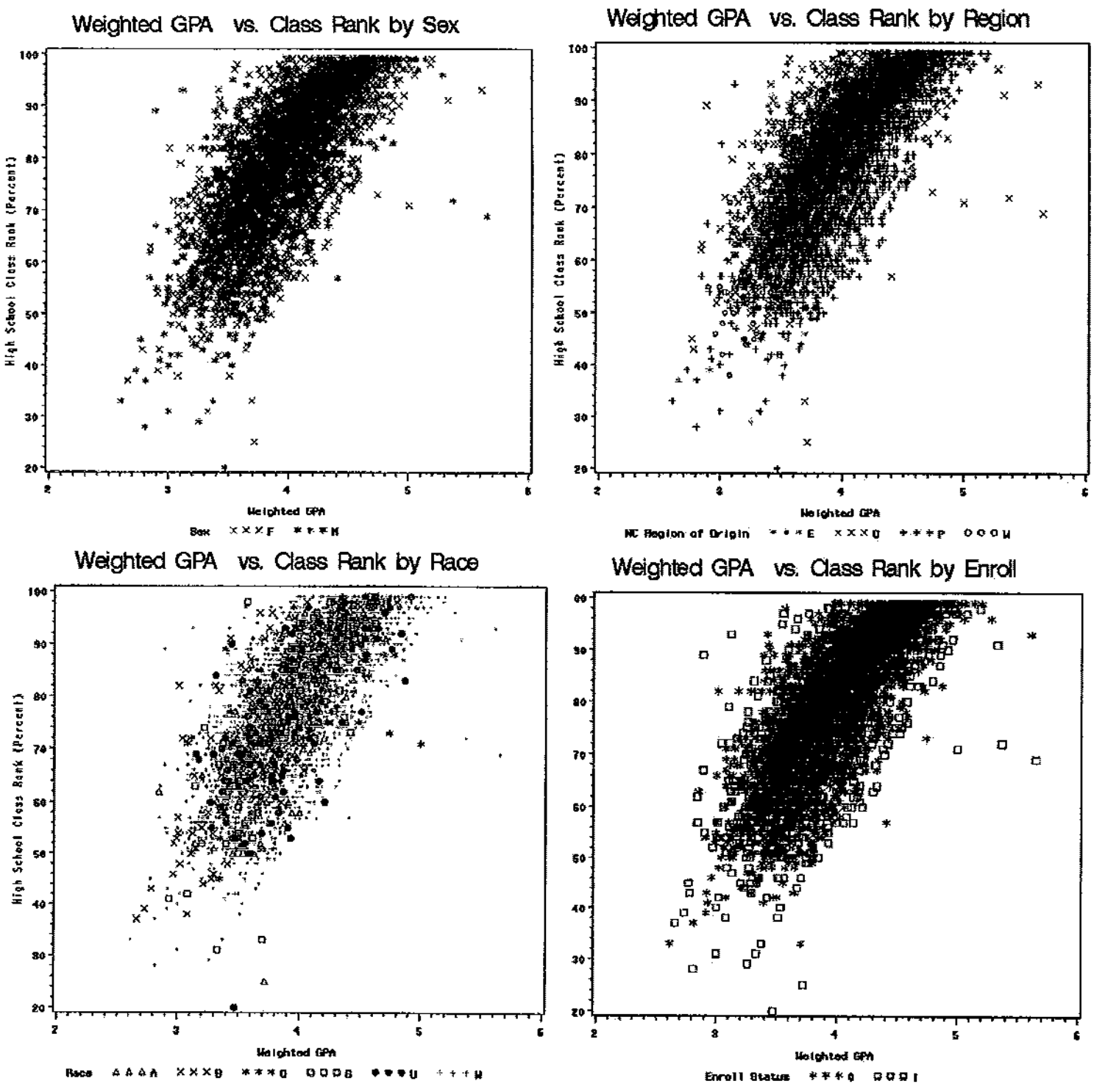

Figure 5.5: Weighted GPA vs. High School Class Rank separated by categorical variables. Starting at the top left and moving clockwise, the categorical variables used are sex, region of origin, whether or not the student enrolled, and race.

The plots of weighted GPA against class rank from figure 5.5 exhibit a denser, more linear pattern than the previous plots in figure 5.4 showing SAT scores. Few strong differences appear when examining the distribution of GPA and class rank separated into different categories. However, one apparent trend is that for a given value of class rank, students from the Piedmont appear to have higher weighted GPAs, on average. Since weighted GPA seems to depend, at least partially, on which region the individual is from, including an interaction term between weighted GPA and region in our model could increase its predictive power.

As a visual check of the linearity in the logit assumption, each of the five continuous variables is binned into six ordered categories. A graph of the empirical logits for each of these six categories for all five continuous variables is given in figure 5.6.

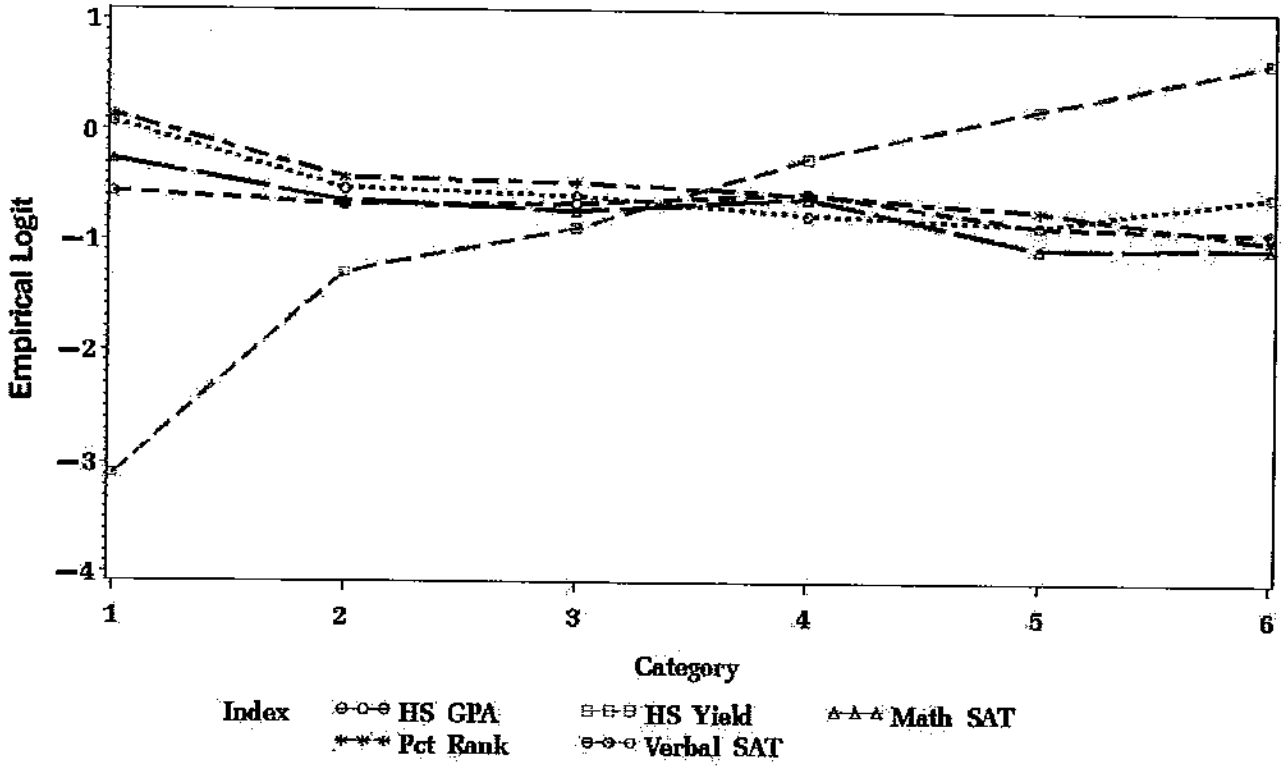

Figure 5.6: Empirical logits of five continuous variables binned into six categories.

Figure 5.6 gives evidence that the linearity assumption is a reasonable one. No strong curvature exists in the empirical logits between the different categories for any of the five examined variables. This plot also makes it clear that, aside from HS Yield, no variables have serious changes in the empirical logits across their different categories. The dynamic nature of the change in high school yield with relation to enrollment is expected, as it is a variable corresponding to the proportion of enrolled individuals from that high school over the last 10 years. This general lack of change in most variables can be interpreted as there is little difference in the sample log odds in favor of a student enrolling as the category for that variable changes. A lack of strong differences in empirical logits for changing values of a given explanatory variable is an indication that the predictive accuracy for any model based on these variables may be limited.

5.5 Model Building

Since including significant interaction effect is likely to enhance the predictive accuracy of our model, a rich interaction model using the eight original variables is fit first. This full interaction model uses all possible two-way and three-way interactions, as well as all possible quadratic effects. Although predictive ability is the main goal of our model, for the sake of simplicity, a reduced interactive model is created by removing highly non-significant terms from the full interactive model.

A likelihood ratio test comparing the rich interaction model to one without all the three-way interactions is conducted. This test gives strong evidence that the two models are not significantly different, so it is concluded that the three-way interaction are unnecessary. As with the main effects only model, backwards elimination based on the Wald Statistics is used to further reduce this interactive model one variable at a time. Any term with a p-value less than 0.5 is left in the model, as anything with even a slight contribution to predictive

power is desired. Quadratic effects for high school yield and verbal SAT score are removed, as well as seven two-way interaction effects. A likelihood ratio test comparing the full interaction model with the reduced interactively model confirms the results of the individual Wald Statistics.

After preliminary investigation of the individual variables, a full additive model using the eight variables previously defined is fit to the data set. The individual Wald statistics provided by SAS are a preliminary indicator of whether or not each individual predictor is statistically significant given this particular model. However, many authorities prefer the likelihood ratio test as it tends to be more reliable, because it is not based on a normal approximation (Allison 1999). Using the drop-in deviance to remove one variable at a time through backwards elimination, this full additive model is reduced to a six variable model with RACE and REGION removed. This reduced additive model along with estimated coefficients and standard errors is given

$$\widehat{logit(\pi(\tilde{x}))} = \underset{}{-0.83} + \underset{(0.37)}{4.63}(High\ School\ Yield) + \underset{(0.26)}{0.35}(Weighted\ GPA)$$

$$\underset{(0.14)}{-0.03}(High\ School\ Percent\ Rank) - \underset{(0.0038)}{0.0028}(Math\ Sat\ Score)$$

$$\underset{(0.00055)}{-0.0012}(Verbal\ Sat\ Score) - \underset{(0.00045)}{0.063}(Sex).$$

Overall, this model selection process results in four candidate models that will be compared in the context of prediction:

1) Full Interaction Model
2) Reduced Interaction Model

3) Full Additive Model

4) Reduced Additive Model

The question of how successful any of these models are at actually predicting whether or not future students enroll is one that has yet to be addressed. Evaluating the predictive accuracy of any candidate model is necessary as even the best model from a candidate set can turn out to be an unsatisfactory predictor of the response. ROC curves will be used as a method of evaluating the accuracy of the four candidate models.

5.6 ROC Plots

The remainder of the analysis focuses on using ROC curves and the resulting AUC values to estimate the predictive accuracy and provide comparisons for the four candidate models that can be used to predict whether or not an individual student will enroll at UNCA. Also included, as reference points for the four models created earlier, are a simple logitistic regression model using weighted GPA only, and a model using only an intercept.

Figure 6.1 is the actual ROC curve from the leave one out validation conducted by PROC LOGISTIC on the UNCA data for the full additive model using all eight explanatory variables. The SAS macro ROC PLOT is used to plot the output from PROC LOGISTIC. The estimated AUC is included.

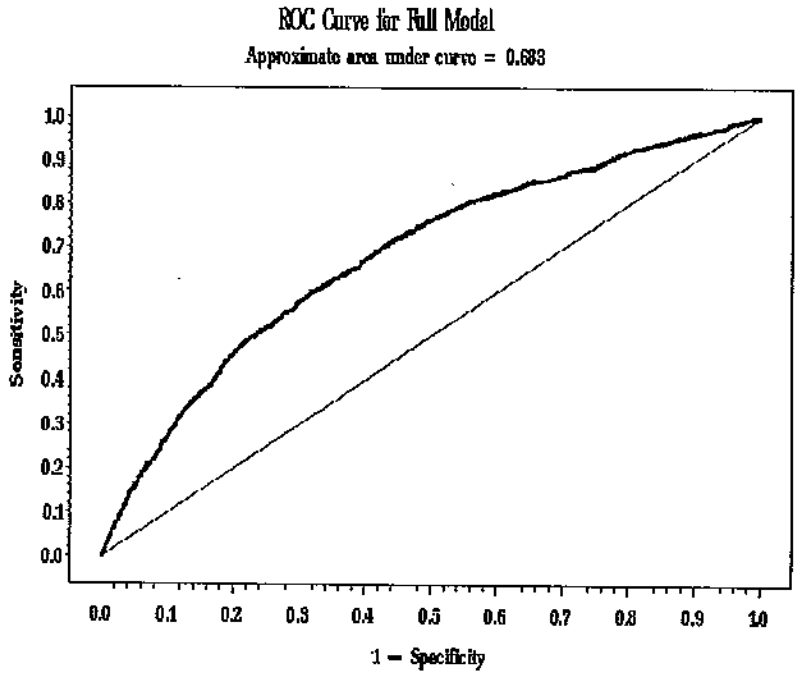

Figure 5.7: ROC curve for full additive model

As with all ROC curves, the curve in figure 5.7 starts at the lower left-hand corner of the ROC space corresponding to a cutoff value of one, and ends at the upper right hand corner as the cutoff point is gradually lowered to zero. ROC curves for these six different models overlaid on the same plot are given.

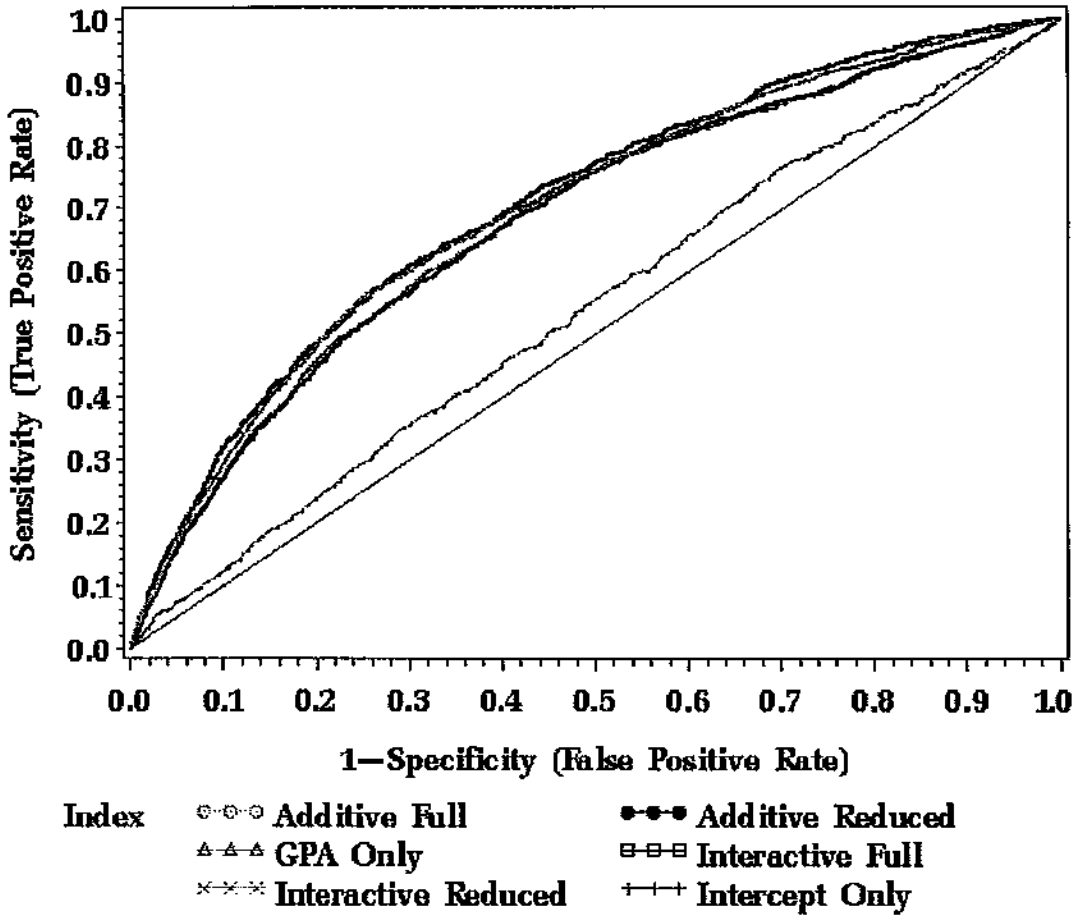

Figure 5.8: Overlaid ROC curves from four candidate models and two reference models

As evidenced by figure 5.8, the four candidate models all have similar ROC curves. The full interaction model, shown in black, has the highest ROC curve for most cutoff points. The reduced interaction model, shown in pink, has a curve that is slightly lower than the full interaction model for most cutoff values. The ROC curves for the full and reduced additive models lie just below the curves for the interaction models. The reference models have significantly lower ROC curves. The model using only GPA, in green, has a ROC curve much closer to the line of no discrimination. Using the intercept only model results in a

ROC forming a 45 ° line across the ROC sample space. This is expected as an intercept only model is equivalent to randomly classifying information.

5.7 AUC Contrasts

While an ROC curve is a quick way to assess the predictive power of a given model, a statistical examination of the AUC results is necessary for scientifically valid conclusions as to how models compare in terms of predictive accuracy. This analysis is done for the six models examined previously by implementing the SAS macro "ROC" to find AUC values with standard errors for the each of the individual models. The ROC macro is also used to test all fifteen pairwise comparisons between the models using contrasts. The analysis used by the ROC macro to compute these values is based on theories from a 1989 paper by E. Delong, D. De Long and D. Clarke-Pearson. The following table provides individual estimates of AUC along with standard errors and 95% confidence intervals:

ROC Curve Areas and 95% Confidence Intervals				
Model	ROC Area	Std Error	Confidence	Limits
Interactive Full	0.7065	0.0078	0.6912	0.7218
Interactive Reduced	0.6990	0.0079	0.6835	0.7146
Main Effects Full	0.6829	0.0081	0.6670	0.6988
Main Effects Reduced	0.6818	0.0081	0.6659	0.6977
Weighted GPA Only	0.5392	0.0087	0.5221	0.5563
Intercept Only	0.5000	0.0000	0.5000	0.5000

The full interaction model has the highest overall AUC with a value of 0.7065, while the reduced interaction model trails slightly with an AUC of 0.6990. The full and reduced

additive models have similar AUC values of 0.6829 and 0.6818 respectively. The model using only weighted GPA has a substantially lower AUC of 0.5392. As expected, the intercept only model resulted in an AUC of 0.5 with zero variability.

Contrasts are used to evaluate whether or not these observed differences in the calculated AUC values are statistically significant. Of the fifteen contrasts tested, all the pairwise comparisons are found to be significantly different except for the comparison of the full additive model to the reduced additive model. Overall, the full interaction model has the highest AUC curve by a statistically significant margin, suggesting it is better than all other models at classifying whether or not a student will actually enroll.

For the two main effects model models, the contrasts suggest there is no significant difference in predictive accuracy. Since the reduced model only uses five variables, ROC analysis suggests that the reduced main effects model is preferred over the full model. This is in agreement with the conclusions from the drop in deviance tests performed previously. Using ROC curves, in addition to likelihood ratio tests, provide a non-parametric measure to compare different models.

6. Conclusion

6.1 Model Selection

Four candidate models are presented here, each with their own strengths and weaknesses. For example, the three-way interactions in the full interaction model provide an additional amount of predictive power, but at the expense of complicating the model. The reduced interaction model is less complicated than the full interaction model and is still significantly than either of the additive models at predicting enrollment. It is notable that the finding of a statistically significant difference between models according to a hypothesis

tests may be largely due to the size of the same. Using almost 5000 individuals allows AUC values to be estimated with a high degree of precision, resulting in statistical significance where the practical implications might be minimal. As a visual reminder of this, on the ROC plot in figure 6.2, the curves for the two interaction models are not easily distinguishable, yet the AUC suggested that they are significantly different.

If the lack of parsimony is not a serious issue to an administrative official, the full interaction model provides the highest degree of predictive accuracy as determined by the hypothesis tests on the AUC values. Assuming the cost of a false negative and false positive are equal, a cutoff point can be selected that maximizes both the sensitivity and the specificity. Interestingly, these values simultaneously reach a maximum of approximately 0.63 when a cutoff value of 0.33 is selected. Visually, a maximum value of specificity and sensitivity for a given model occurs at the point of intersection between the ROC curve and a 45° tangent line. Applying this idea to figure 6.2 makes it evident this intersection occurs at a cutoff point of about a third.

6.2 Model Predictions

Using this full interaction model, it is possible to examine specific estimated probabilities of enrolling for individuals. The highest estimated probability of enrolling from any individual in the UNCA data was 0.99954, with a 95% confidence interval of 0.98281 to 0.99999. He is a white male from an out of state high school with a yield of 24%. His high school rank was the 69th percentile, with a weighted GPA of 5.65 and no reported unweighted GPA. His standardized test scores were 580 on the math section of the SAT, 520 on the verbal section, and a 20 on the ACT. For any reasonable cutoff this individual is classified as enrolling at UNCA, which, it turns out, is a true positive prediction.

The lowest estimated probability of enrolling is 0.10964, with a 95% confidence interval of 0.00051 to 0.43999. This individual is a black female from a high school in eastern North Carolina with a 18% yield rate. She graduated in the 97th percentile of her class, with a weighted GPA of 4.45 and an unweighted GPA of 3.87. Her scores on the math and verbal sections of the SAT were 640 and 690 respectively. She did not end up enrolling at UNCA, which is also the conclusion the model makes for a cutoff value of one-third. Another interesting point is that the cutoff point of 0.33 lies within the 95% CI for the probability of enrolling for this individual. Basing predictions on simply the point estimate of enrolling is necessary in order to make sure every individual is classified into one of the two possible categories.

There are a number of distinctions regarding these two individuals with very different estimated probabilities of enrolling. Variables such as gender, SAT scores, and high school year rate are all different. This simple examination makes it impossible to determine if the difference in estimated response is caused by a single variable or a combination of variables working separately or together.

6.3 Discussion

Parametric techniques requiring distribution assumptions such as the likelihood ratio tests and non-parametric graphical measures such as ROC curves have been used to evaluate different binary predictive models. The most complex interaction model was determined to have the highest absolute predictive power as measure by the area under its ROC curve. However, further investigations are necessary to determine how practically useful this achieved level of predictive power is. Alternative methods such as generalized additive models (GAM) or regression trees might also be successful for making predictions from this type of data.

Using multiple statistical methods to analyze a situation is typically a sound practice. The full interaction model had a significantly higher predictive power than all the others as judged by a test statistic and a p-value. However, the visual comparison provided by the ROC plot casts doubt upon the practical importance of these differences. Also, the analysis in this paper is limited to validation using leave one out, and could possible be improved by conducting similar ROC analysis using other types of validation, such as external validation, or leave many out. ROC curve analysis based on multiple validation techniques that picked the same model would give further credibility to that model being preferred.

Regarding the creation of a model that provides the greatest predictive power when used on incoming students, the missing data should be addressed further. A model that is able to make predictions on all applicants, not just those providing the right combination of variables, will have much greater utility. Problems arise when attempting to predict whether or not a student will enroll using any model including variables that student does not provide. From a data collection standpoint, a more standardized admissions survey might result in a higher overall completion rate for a certain variables. Even merely reducing the proportion of missing data to a manageable quantity allows for the potential use of imputation techniques for filling in missing values.

The generalized linear model provides a substantial framework for many different types of data. For a binary response variable, logistic regression is well-suited for making future predictions based on previously observed data. Furthermore, in this modern era, computer programs such as SAS can be used to quickly fit a variety of models and implement sophisticated model assessment techniques, to aid in the development of predictive model that can potentially save time and money.

Predicting whether or not a student will enroll is a complex issue, as many different variables are simultaneously affecting whether or not that student enrolls. As an example of

this, consider the weighted GPA variable. When a simple logistic model is fit with only this variable, the higher the GPA the less likely a student is to enroll. However, the individual with the highest estimated probability of enrolling was also the individual with the very highest weighted GPA out of almost 5000 individuals. A useful statistical model must successfully, and simultaneously, include the effects provided by multiple measureable characteristics of that individual in order to make accurate and reliable predictions. A successful predictive model could potentially result in the savings of hundreds of thousands of dollars for a given university.

References

- Agresti, Alan. An Introduction to Categorical Data Analysis.Hoboken: John Wiley & Sons, 2007

- Allison, Paul D. 1999 Logistic Regression Using SAS © : Theory and Application. Cary, NC: SAS Institute Inc.

- Casella, George and Roger Berger. Statistical Inference. Duxbury, 2002

- Delong, Elizabeth and David, and Daniel Clarke-Pearson "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach."Biometrics 44: 837-845

- Hosmer, David and Stanley Lemeshow. Applied Logistic Regression. New York: John Wiley & Sons, 1989

- Myers, Raymond, Douglas Montgomery, and Geoffrey Vining. Generalized Linear Models: With Applications in Engineering and the Sciences. New York: John Wiley & Sons, 2002

- Ramsey, Fred and Daniel Shafer. The Statistical Sleuth: USA: Duxbury, 2002

- "Receiver operating characteristic." *Wikipedia, The Free Encyclopedia.* 16 Apr 2009, 23:34 UTC. 21 Apr 2009 <http://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=284312148>

- SAS Institute Inc., Logistic Regression Examples Using the SAS® System, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1995.

- Stokes, Maura E., Davis, Charles S., Koch, Gary G., Categorical Data Analysis Using the SAS System, Cary NC: SAS Institute Inc., 1995.