

UCLA

UCLA Electronic Theses and Dissertations

Title

A Simulation Approach to Predicting College Admissions

Permalink

<https://escholarship.org/uc/item/8r2695n5>

Author

Lim, Daniel Kibum

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

**A Simulation Approach to Predicting College
Admissions**

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Daniel Kibum Lim

2013

© Copyright by
Daniel Kibum Lim
2013

ABSTRACT OF THE THESIS

A Simulation Approach to Predicting College Admissions

by

Daniel Kibum Lim

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Mark Handcock, Chair

Many educational websites offer tools that predict students' chances of admissions to college using models estimated on crowd-sourced data. Unfortunately, these data are generally not representative of the actual population of college applicants, so admission probability calculators (APCs) which use them offer biased estimates of admission probability, $Pr(A)$. As a solution, I propose a simulation-based framework for estimating $Pr(A)$ that is not modeled on crowd-sourced data. I model the joint distribution of enrollee characteristics at a particular university using a Gaussian copula and official margin data published by the College Board[™] and the university itself. I fit a model for $Pr(A)$ to this data based on several reasonable assumptions about the character of the applicant pool. I show that the estimates of $Pr(A)$ yielded by my framework are excellent predictors of admission outcome for a sample of student data collected from the Internet.

The thesis of Daniel Kibum Lim is approved.

Robert Gould

Nicolas Christou

Mark Handcock, Committee Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

1	Introduction	1
2	Calculating admission probability: approaches and challenges . . .	3
	2.1 Web surveys and biased data	4
	2.2 Insufficient data for accurate simulations	9
3	Available data	9
4	Distributional assumptions	12
	4.1 Margins	12
	4.2 Joint distributions	13
5	Overview of the approach	15
	5.1 Generating the admit pool	17
	5.2 Generating the applicant pool	17
	5.3 Combining the applicant and admittee pools	20
	5.4 Binning and estimating probability of enrolling	20
	5.5 Modeling $Pr(A)$ on $Pr(E)$	21
	5.6 Uncertainty	24
	5.7 Assessing predictive power	26
6	Results	27
7	Conclusion	40
8	Sources	42

LIST OF FIGURES

1	Official and empirical (web survey) marginal data for 4 schools.	8
2	Comparison of CDFs for SAT and HSGPA marginal distribution forms.	13
3	Bounds on possible applicant pools.	18
4	Bounds on marginal distributions for school 1431.	19
5	$\Pr(E)$ for school 1325.	22
6	Stylized illustration of assumptions that constrain $\Pr(A)$ model selection	23
7	95% confidence interval and median prediction surface for school 544	25
8	$\Pr(E)$ and $\Pr(A)$, schools 1-3	29
9	$\Pr(E)$ and $\Pr(A)$, schools 4-6	30
10	$\Pr(E)$ and $\Pr(A)$, schools 7-9	31
11	$\Pr(E)$ and $\Pr(A)$, schools 10-12	32
12	Empirical versus predicted $\Pr(A)$, schools 1-6.	38
13	Empirical versus predicted $\Pr(A)$, schools 7-12.	39

LIST OF TABLES

1	Bias in reported admission rate, percent admits.	6
2	Bias in reported admission rate, percent admits (cont'd).	7
3	List of schools and relevant data.	28
4	Predictions.	34
5	Predictions (continued).	35
6	Reduction ratios.	36

1 Introduction

For the millions of students applying to college each year, assessing chances of admission is an integral part of formulating application strategies. Accurate assessments are valuable because they help applicants intelligently allocate resources - applications cost money and effort, and there is no point in expending either toward schools to which one has no chance of getting in. Even for younger students, understanding how different characteristics affect their future prospects can help decide how to direct their efforts. For example, a student may face the decision of whether to take an extra class to boost their GPA or spend that time preparing for the SAT. By understanding the effect of prospective improvements in each, students can essentially optimize their activity portfolio to maximize probability of admission to target schools.

Given the importance of college admissions, it is not surprising that there are tools available on the Internet that provide estimates of admission probability. These admissions probability calculators (APCs) are offered by many educational websites and work as follows. Users input a variety of data that are pertinent to the admission decision such as academic performance and extracurricular activities. Based on that information, the APCs predict how likely candidates are to be admitted to a certain school. Finally, users are asked to report their decisions at the end of the application season so future calculations can be made with the latest data. While the websites that provide these tools are not very transparent about methodology, it is probably the case that the prediction models they use are estimated on the user submitted data, taking what is essentially a crowd-sourcing approach to the problem.

In one sense, this is a sensible solution and a valuable service. In the absence of such tools, the average college applicant simply does not have sufficient data

to understand what the pool of applicants, i.e. their competition, looks like. Without that information, it is impossible to calculate a probability of admission tailored to a particular individual. Unfortunately, the data used to estimate these models are self-reported so that they are non-random and suffer severely from selection bias. This is because (1) the data are often public so users are subject to social desirability bias, and (2) the characteristics that make certain students more academically successful probably make them more likely to seek out and use the tools. Building a prediction model on such data will lead to biased estimates and, for reasons discussed later, common fixes for selection such as propensity scoring or the Heckman correction cannot be used for these data.

Given these problems, I propose a simulation based framework for estimating admission probability, $Pr(A)$, using a bounding approach and distribution information on enrollees from publically available, *official* sources¹. Characterizing an individual as a vector of student attributes (viz. SAT scores and high school GPA), I first generate a pool of students who enroll at the university using data on marginal distributions published by the College Board and individual colleges. The joint distributions are modeled using Gaussian copulas. Next, I generate bounds on distributional parameters for the full pool of applicants based on two assumptions. The first is that admits and enrollees are, on average, of equal or higher quality than the full pool of applicants; this yields an upper bound. The second assumption is that admits and enrollees are a subset of the applicants, in effect putting a lower bound on the average quality of the full admit pool. A simulated full pool of applicants can be generated using any set of parameters within the feasible region. Once generated, each hypothetical applicant pool is binned along the dimensions of SAT score and HSGPA, and the probability of enrolling $Pr(E)$ within a certain SAT-HSGPA bin is calculated as the ratio of enrollees to

¹Official sources are the College Board, which administers the SAT, and the admitting college themselves.

applicants in the bin. Finally, $Pr(A)$ is estimated by fitting an arbitrary model to the binned $Pr(E)$ data, subject to several constraints. First, the model must yield valid probability estimates (i.e. between 0 and 1). Second, probability of admission must increase monotonically with each predictor. Third, $Pr(A)$ must be greater than or equal to $Pr(E)$ for all possible values.

As a proof of concept, I fit one possible $Pr(A)$ model to demonstrate the fruitfulness of the overall framework². Without actual admission data, it is not possible to truly validate the accuracy of my estimates. However, using web survey data from a commercial APC, I show that my estimates are excellent predictors of admission outcome and possess many of the properties we'd expect from the true $Pr(A)$. I also show how bootstrapping can be used generate confidence intervals on the predictive surfaces, as well as individual predictions.

2 Calculating admission probability: approaches and challenges

The simplest way of estimating probability of admission to a particular school is to divide the number of admits by number of applicants, yielding the admission rate. This estimate is unbiased but also uninformative; it does not take into account any individual characteristics and basically models admission as random selection. Clearly, the next step to take is to account for individual characteristics, which in this study, I constrain to SAT I scores (comprised of math, reading and writing sections) and high school GPA (HSGPA). There are at least two obvious ways of taking these as well as any other quantifiable factors in account.

First, one might take a regression-based approach. Given a dataset of the appli-

²The model I use is but one of many possible models. Others may yield similarly good or even better estimates.

cant pool containing relevant characteristics and admission decisions, it is a trivial task to estimate a regression model which can be used to predict $Pr(A)$. A second approach is even simpler. For some vector of covariate values characterizing the user/student, one can simply take as the estimate of admission probability the ratio of admittees to applicants with similar vectors (where similarity might be defined in a variety of ways). If individual-level data are not available, one can take the same approach using simulation, given some information about the distribution of the applicant and admit pools (e.g. covariance structures). However, practically speaking, both these approaches are hobbled by limitations of data.

2.1 Web surveys and biased data

In the case of regression, the biggest problem is non-randomly sampled data. For regression to yield unbiased estimates, the data being used must be representative of the population (i.e. randomly selected) or the model must correct for resulting biases.

When the data comprise a census or a representative sample, calculating probability is relatively trivial. For example, to study the impact of admission preferences on admit pools, Espenshade, Chung, Walling (2004, 2005) use the National Study of College Experience in which “Ten... colleges and universities... supplied individual-level data on all persons who applied for admission in the fall of 1983 (or a nearby year), 1993, and 1997” (p.1424). Their methodological discussion simply states that they “fit a series of logistic regression models... [and] investigate the potential importance of including interaction terms” (p.1427). Similarly, other studies that examine correlates of college admission do so by running simple logistic or probit models using school-supplied data on full applicant pools (Berk, Freedman and Stark 2002).

Unfortunately, colleges and the College Board do not release individual level SAT and HSGPA data to the general public³. Therefore, it is likely the case that any commercial or non-academic probability calculator that uses “real student data” obtained that data through voluntary submission from students. In essence, these calculators base their analytics on data from a web survey. Unfortunately, web surveys have a variety of quality problems stemming from the nature of the medium including measurement, coverage, and sampling error (Groves, et al. 2009, Couper 2000). All of these issues are potentially at play for the voluntary student data.

Measurement error is “the deviation of the answers of respondents from their true values on the measure” (Couper 2000, p.475). Such errors can arise from a variety of sources including the user, administrator and the instrument, technical or wording problems, and reasons both intentional and unintentional. For student data, one obvious source is social desirability bias, which is “the tendency to present oneself in a favorable light” (Groves et al. 2009, p.168). Because the data are publically available, students may misreport attributes (i.e. higher than actual HSGPA or SAT score) or the actual decision reached (i.e. report accepted when actually rejected). Measurement error may also arise from different uses of the APCs. Some students may use the tools to predict their chances on current scores while others may use them to strategize or goal-set with different sets of potential scores. In short, some or much of the data may be inaccurate.

Coverage error arises when “the target population does not have a... sampling frame that matches it perfectly” (Groves, et al. 2009, p.54). There results a coverage bias if there are differences between those within the frame and those

³“The College Board does not release such data with PII to anyone other than the individual to whom the data pertain without the consent of the individual.” Guidelines for the Release of Data p.5, College Board Office of Research and Development. Retrieved from <http://media.collegeboard.com/digitalServices/pdf/research/RDGuideforReleaseData.pdf> on 10/1/2012

outside on characteristics that affect the outcome of interest. For this application, the coverage problem arises because of mismatch between those who do and do not have Internet access. Fortunately, this is probably less of an issue with our target population of college applicants than with the population at large. Younger individuals are more prone to use the Internet and given the fact that we are looking at college applicants, there is probably some truncation at the lower end of the income distribution.

Perhaps the biggest concern about the survey data has to do with sampling error, which here takes the specific form of self-selection. Selection bias arises in situations where potential survey participants are given the option of not participating, or if there is some mechanism that non-randomly selects participants. The bias actually manifests if a factor that correlates with the probability of participating (whether through self-selection or the external selection mechanism) also correlates with the outcome of interest. In the context of this study, it seems reasonable to think that the characteristics that make some students more academically successful than others (e.g. diligence, concern about academic performance, research skills) make them more likely to seek out advantage in the admissions process by using tools such as APCs. As a result, we would expect to see higher achieving students overrepresented in the data. Then, if regressions were run on this data, we'd expect estimated effects to be biased.

	544	1737	1022	342	1407	1325
official	7.70	9.90	15.60	20.70	23.60	26.10
survey	55.20	55.40	61.60	61.10	62.20	62.50

Table 1: Bias in reported admission rate, percent admits.

Whatever the precise cause, examining a sample of student web survey data clearly shows that it is not representative of the full population of college applicants and

	436	1503	1324	1583	1431	1593
official	27.20	34.80	36.70	36.70	54.80	62.80
survey	68.70	64.50	68.10	64.80	73.20	71.40

Table 2: Bias in reported admission rate, percent admits (cont’d).

liable to yield upwardly biased estimates⁴. To begin with, the proportion of students accepted in the web sample is far greater than actual admission percentage. Tables 1-2 shows the extent of this difference. The first row is the admission rate reported by schools; the second is the admit rate in survey data scraped from a commercial APC. Without exception, the survey data overestimate admission rate, and this is the case across the spectrum of selectivity. This bias could be due to a combination of misreporting as well as selection effects. Further, outcome is not the only measure that is biased; the same is true for student characteristics. Figure 1 shows the distribution of student SAT scores for several schools based on (1) web survey data, and (2) official data. Again, it’s clear that the survey data are not representative of the full population of applicants, with official data implying margins that are more left skewed.

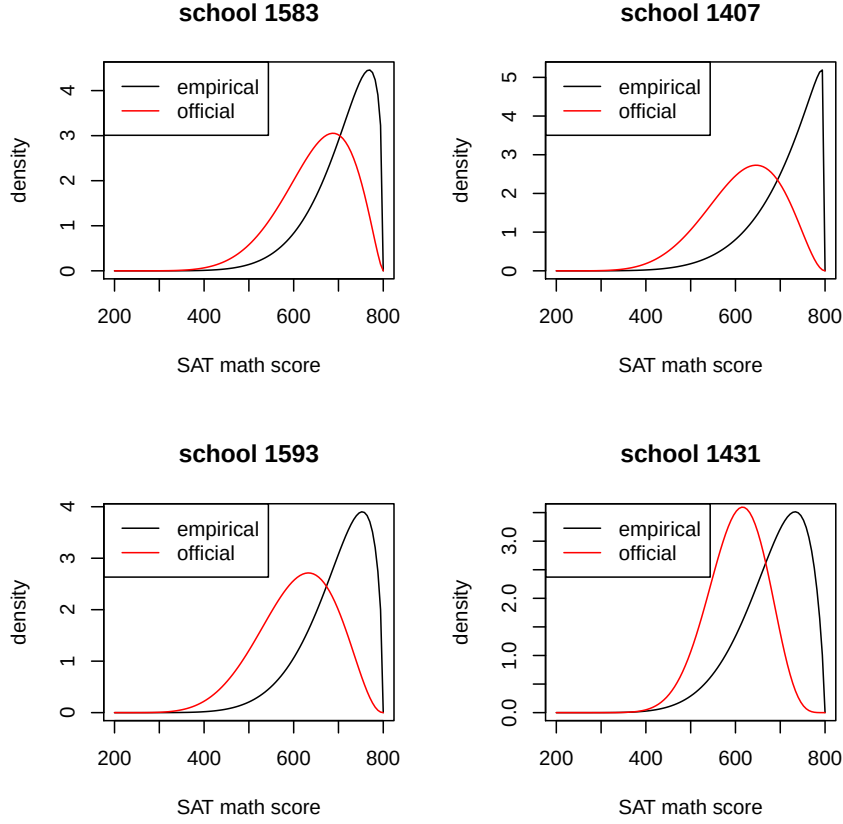
Dealing with selection bias in the context of regression is a well-known problem with several potential solutions such as the Heckman correction and propensity score matching (Heckman 1979, Rosenbaum & Rubin 1984). In the former case, one runs regressions of the form:

$$y_i = \mu + z_i^* \alpha + x_i' \beta + \lambda(\zeta_i, z_i^*) \gamma + \epsilon_i$$

where $\lambda(\zeta_i, z_i^*)$ is a separate first stage regression of the treatment indicator z_i^* on the predictor ζ_i (Obenchain & Melfi 1997, Puhani 2002). Propensity score matching similarly begins with a first stage regression to model selection into the treatment group. Rather than include the calculated propensity for each observation

⁴These data were obtained from Parchment.com (<http://www.parchment.com/c/my-chances/>). All references to web survey data in this paper refer to these data.

Figure 1: Official and empirical (web survey) marginal data for 4 schools.



in a regression (as in the Heckman approach), the analyst matches observations in the treatment and control groups according to their propensity scores. He then calculates the average treatment effect across comparable observations, mitigating the effect of selection (Rosenbaum & Rubin 1985).

Unfortunately, neither of these approaches can be used for predicting admissions. This is because both approaches require, at a minimum, some number of observations to be from the control group, which in this case are the students that apply to a particular school but do not use an APC. To summarize, it is quite likely that voluntarily submitted student data is subject to several types of error associated with web surveys, and while there are methods available for dealing with the most egregious of these, our data are such that those methods cannot be applied.

2.2 Insufficient data for accurate simulations

Problems with web surveys notwithstanding, pursuing a simulation based approach is no easier, again because of lack of data. To see why, consider that in the application process, there are 3 nested pools of students. These are (1) applicants, (2) admittees; and (3) enrollees. Admittees are a nonrandom subset of applicants, and enrollees are a nonrandom subset of admittees. Unlike in the first relationship, enrollees at most schools are selected inverse to their academic quality because those with higher quality will have higher probability of matriculating at a more selective institution. To get a complete picture of the application process, information on each of these pools is required. To calculate probability of admission, one needs information on applicants and admittees.

Unfortunately, the only information published by colleges is quantile data on the marginal distributions of enrollee characteristics (e.g. SAT I component scores, HSGPA). One could, with reasonable assumptions, base a pool of simulated admits on enrollee data. More problematic is the fact that there is basically no data on the full pool of applicants. Without this information, it is impossible to directly infer what the denominator is in any admit to full-pool ratio except at the topmost level (all admits over all applicants). What can be done is to bound the range of possible applicant pool parameters given what we know about the pool of admits. This is the strategy I follow, described in the proceeding sections.

3 Available data

Obviously, there is available a great deal of information about the admissions process, just not what's needed for direct calculation of admission probabilities. Official data are those released by parties with direct access to full applicant data,

while any other data is, broadly speaking, unofficial since it is impractical for others to obtain a *verifiably* representative sample of any meaningful category.

Both the College Board, which administers the SAT, and individual colleges have detailed and complete information about their respective population and all subsets, from which they publish excerpts. Statistics on each year’s population of college-bound seniors are published by the College Board. While that data does not provide the covariance of SAT I scores and HSGPAs, it does provide population level means. For a picture of population level covariance structure, I turn to the College Board report “Validity of the SAT for Predicting First-Year College Grade Point Average” (Kobrin, et al. 2008). Table 4 of Kobrin, et al. 2008 provides the correlation matrix of SAT sections and high school GPA (HSGPA) for the 2006 college-bound cohort. It seems reasonable to assume that the population level correlation structure remains roughly the same from year to year. Therefore, the 2006 correlation matrix may be used in conjunction with annual population-level marginals to generate a population covariance matrix for a given year.

Though the College Board provides population-level data, school-specific statistics must be obtained from the schools themselves. Many schools organize their publicly available data using the Common Data Set (CDS) format, which contains a wide range of information about admission and entering cohort characteristics. Even when schools don’t use the CDS, they generally make such information available on their websites. Colleges typically supply quantile information which, with parametric assumptions, can be used to estimate parameters for marginal distributions. Additionally, two other data points can be obtained from these sources: the total number of applications and the number of acceptances for a given application season. These sources do not provide any information on the overall applicant pool (other than size) nor the shape of the joint distribution for the

admitted pool.

Unfortunately, the only publically available official information on the covariance of HSGPA and SAT I scores is at the population level, and there is good reason to think that these will be significantly different from covariances for a particular school. In the former, the shape of the joint distribution is largely determined by the actual correlation between HSGPA and composite SAT I performance, which was 0.53 in 2006 (Ibid.). This is what we expect from a test that tracks academic ability. School-specific correlations, however, may be significantly different from population figures for a couple of reasons. On the one hand, correlation between HSGPA and SAT performance may be lower to the extent that high performance on one dimension mitigates low performance on the other. On the other, some schools may end up with admittee pools that have higher correlation across characteristics because of boundary effects (e.g. thresholds for admission or extremely selective schools).

An added wrinkle is that selection processes like college admission generate samples that are range restricted relative to their parent populations. For such a sample, it is a well-known problem that its correlation structure can deviate significantly from that of its parent unrestricted population (Sackett & Yang 2000; Birnbaum, Paulson, & Andrews, 1950; Levin 1972). There are a variety of different adjustments that can be made to mitigate this problem depending on the selection mechanism and information available about the range restricted variables (Sackett and Yang 2000). For this thesis, I use the correction derived by Bryant and Gokhale (1972) and Alexander (1990) for situations when there are decent estimates of the unrestricted variances of the covariates but the precise selection mechanism is unknown (Sackett and Yang 2000, p.115).

4 Distributional assumptions

4.1 Margins

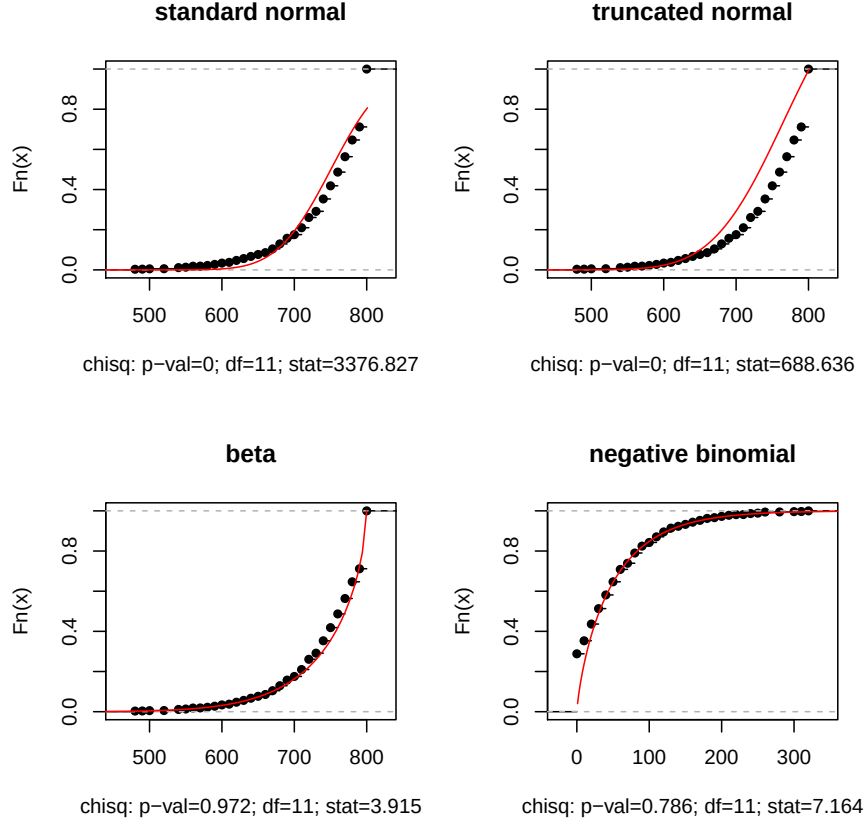
At the population level, the design of the SAT is such that each of the component sections should have a normal distribution. Accordingly, many studies involving the SAT I make a seemingly acceptable assumption that its various sections are distributed multivariate normal (e.g. Clark, Rothstein, Schazenbach 2008). Unfortunately, this assumption gets increasingly unreasonable the further one moves away from the population mean for the simple reason that SAT scores are bounded. For modelling various aspects of student quality, there are several better fitting distributions than the basic normal distribution.

One option is to use a truncated normal distribution which is a normal distribution with one of or both an upper and lower bound. Another is to normalize the range of possible scores and use a beta distribution, which ranges from 0 to 1. Yet another is to begin with the fact that SAT scores are actually discrete values and model deviations from a perfect score as either Poisson or negative binomial distributed. Figure 2 shows the CDFs for each of these distributional forms compared to the empirical CDF for one set of voluntary student data for one of the colleges in my sample⁵.

Performing a chi-squared test for fit shows that we can clearly reject the null hypothesis that the data were drawn from either a normal or truncated normal distribution. On the other hand, both the negative binomial and beta distributions fail to reject the null. Repeating this exercise for HSGPAs from the voluntary data yields a similar conclusion. Between the beta and negative binomial distributions, I found the parameterization of the former to be more tractable for this appli-

⁵Distribution parameters were found through maximum likelihood estimation.

Figure 2: Comparison of CDFs for SAT and HSGPA marginal distribution forms. Black points indicate sample CDF. Red line indicates best parametric fit.



cation. I therefore proceed using beta distributed margins for SAT I component scores and HSGPA in the proceeding simulations.

4.2 Joint distributions

Defining univariate margins for each component of student quality is useless to the goal of simulation if we cannot also relate them in a way that allows identification of individual students. While a multivariate version of the beta distribution does exist, it is not suitable for this purpose because of its restrictive parameterization. Further, if I chose to use different marginal distributions for different components

of quality, it would be difficult if not impossible to define the joint distribution using a standard multivariate distribution. Given the requirements of the task at hand, what is required is a way of defining joint distributions by margins and a correlation structure. Copulas are, by definition, perfect for the task.

A copula is “a joint distribution function of uniform random variables,” the margins of which can be transformed to desired parametric univariate distributions while preserving the rank correlation between all of the variables. Essentially, it provides a mapping from the disparate margins to an underlying easy-to-define multivariate distribution (Hull 2012). Sklar’s Theorem defines the fundamental relationship that is leveraged in practical applications of copulas. Formally stated:

$$H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n))$$

$$C(u_1, \dots, u_n) = H(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n))$$

where H is a joint distribution with $F_1, \dots, F_n$ marginal CDFs, C is a copula and $u \in [0, 1]$ (Embrechts, Lindskog & McNeil 2001, p.4). The foundational result is that “the univariate margins and the multivariate dependence structure can be separated, and the dependence structure can be represented by a copula” (ibid.). When $H = \Phi_R^n$ where Φ_R^n is the standard multivariate normal distribution with correlation structure R ,

$$C(u_1, \dots, u_n) = \Phi_R^n(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n))$$

known as the Gaussian copula. This powerful transformation allows the definition of a joint distribution by its margins and their relationship structure. It should be noted that applying a copula preserves rank correlation but not Pearson’s r because the marginal transformations change the moments of each margin and therefore the overall correlation structure. The degree to which the resulting correlation structure, $\hat{\rho}$, deviates from the desired ρ varies from situation to situation. One could use optimization or MCMC to find the R (correlation structure for the

copula) that leads to the exact target ρ . For the parameters I use, however, the differences between ρ and $\hat{\rho}$ appear minimal and therefore do not merit this extra complication.

5 Overview of the approach

I propose a simulation based framework for estimating probability of admission for a particular school as follows:

1. Estimate bounds on applicant pool parameters.
2. Select applicant pool parameters from feasible region. Generate applicant pool using the implied margins and some correlation structure joined with a copula.
3. Generate a pool of enrollees using (official) enrollee margins and some correlation structure joined with a copula.
4. Combine the enrollee and applicant pools through matching.
5. Estimate $Pr(E)$ using the simulated data.
6. Fit a model for estimating $Pr(A)$ based on $Pr(E)$ (e.g. regression or some non-standard parametric model).
7. Repeat steps 2-6 with random combinations from the feasible region to generate confidence intervals.

For this proof of concept, the above framework is implemented with the following details:

- Enrollee margin parameters are estimated using data from questions C9-C11

of the 2010-2011 Common Data Set submitted by each school. Correlation matrices are estimated from web survey data obtained from a commercial APC.

- Bounds on applicant pool parameters are estimated using enrollee margin parameters above, satisfying the conditions in equations 5.2 (below).
- Applicant pool generated using the midpoint of each feasible region.
- Applicant and enrollee pools are consolidated by matching on Mahalanobis distance with replacement. The result is an $m \times n$ matrix where n is the number of predictors plus 1, and m is the number of observations. $n = 5$ with the following columns: 3 SAT component sections, HSGPA and an indicator for admission.
- The simulation data are reduced to 2 predictor dimensions, SAT (i.e. sum of component scores) and HSGPA, for the sake of interpretability.
- Data are binned along each of the predictors, resulting in a $p \times q$ lattice with each bin containing a subset of simulated students with similar predictor values.
- Within each bin, $Pr(E)$ is estimated as the ratio of admits to applicants in that bin.
- $Pr(A)$ is estimated by fitting model 0.2 to the simulated $Pr(E)$ data (details below).

To be clear, what I am proposing is a *framework* for estimating admission probabilities. I am not advocating one specific model so much as outlining principles which constrain the kinds of models that can be used. The following is a proof of concept for the framework using one possible set of parametric and modeling

choices. At each step of the process other choices can be made which may yield similarly good or even better estimates.

5.1 Generating the admit pool

To estimate marginal parameters from quantile information for the admit pool , I minimize “the sum of squared differences between the given probabilities and the theoretical probabilities of the specified distribution evaluated at the given quantile points” (rriskDistributions manual, p.6). This yields distribution parameters for each of the margins (i.e. covariates). For the correlation structure of the enrollee pool, I use the sample correlation from the web survey corrected for range restriction using the following estimator (Sackett & Yang 2000):

$$\hat{\rho}_{xy} = r_{xy}(s_x/S_x)(s_y/S_y) + [(1 - (s_x/S_x)^2)(1 - (s_y/S_y)^2)]^{1/2}$$

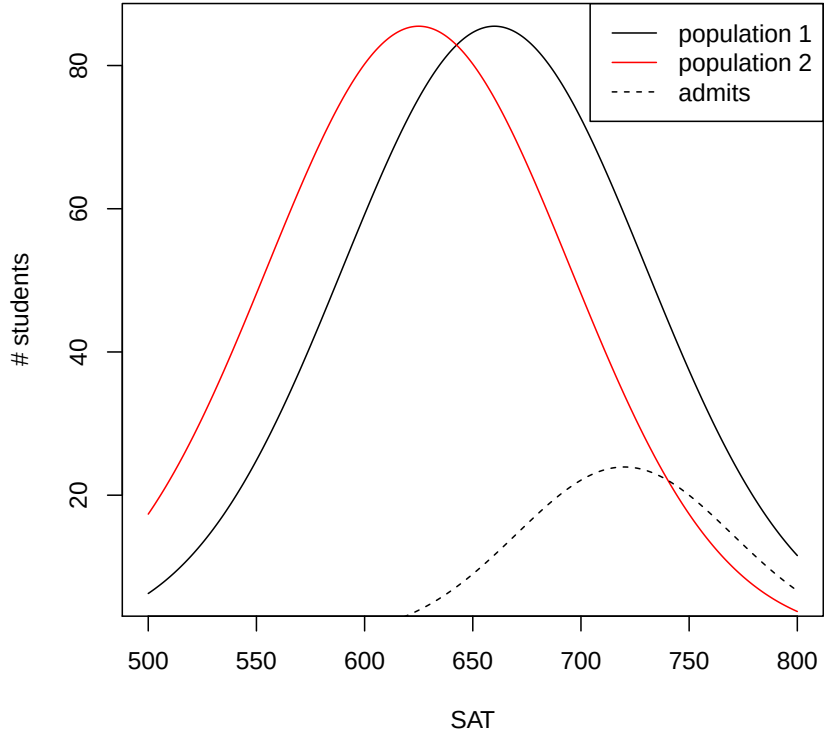
where s_x and s_y are the standard deviations for variables x and y in the range restricted sample, S_x and S_y the same figures for the full population, and r_{xy} the correlation between variables x and y in the restricted sample. For S_x and S_y , I use the population level standard deviations as a reasonable estimate. Using the correlation matrix $\hat{\rho}$, I generate a Gaussian copula: Φ^{-1} of a multivariate normal distribution with mean 0 and $\Sigma = \hat{\rho}$. Finally, each of the margins are transformed to the desired beta distributions by taking F^{-1} with the parameters estimated earlier.

5.2 Generating the applicant pool

Barring further information from admitting institutions, we simply cannot know the *exact* distribution of the applicant pool. However, the range of possible parameters can be reasonably bounded based on what we qualitatively know about the admission process. On the upper end, we know that covariate means for the

admit pool must be greater than or equal to those of the entire pool of applicants because of the admissions decision process. On the lower end, the distribution of the entire pool must be located and shaped such that it is capable of generating the pool of admittees; it cannot be too different. Figure 3 illustrates this point.

Figure 3: Bounds on possible applicant pools. Population 2 could not have generated the admit pool; there simply weren't enough students with high enough scores. Population 1, on the other hand, is a valid applicant population.



Using this intuition, I estimate the parameter bounds of an applicant pool margin given the corresponding admit margin. For covariate j , the range of θ_j , the parameters of the marginal distribution for j , are constrained as follows:

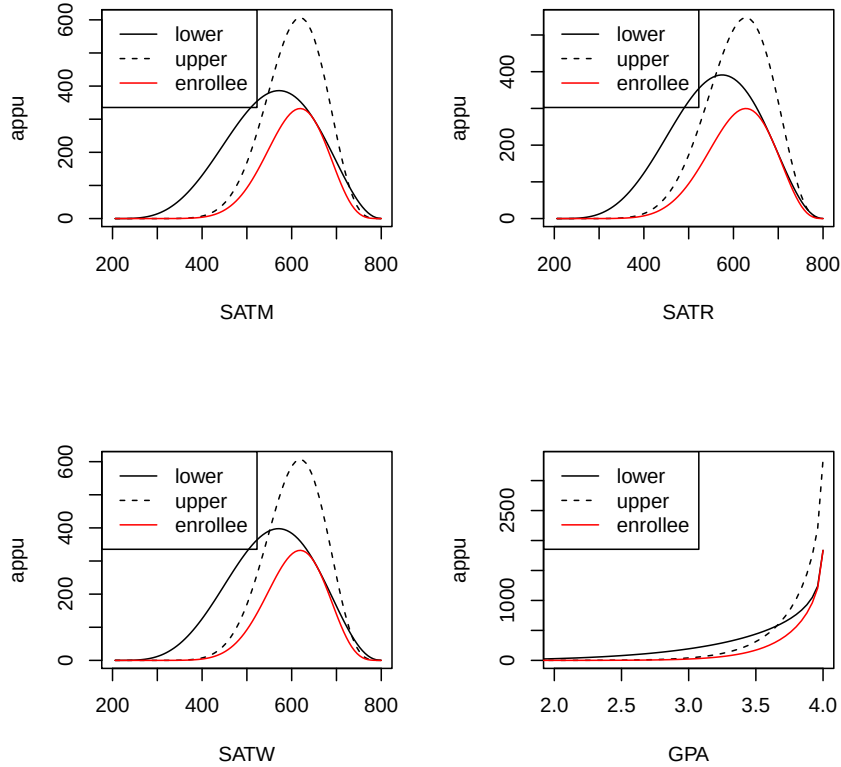
upper: $\mu_j^{(app)} \leq \mu_j^{(adm)}$

lower: $f(x; \theta_j^{(adm)}) \times \#admits < f(x; \theta_j^{(app)}) \times \#applicants \forall x$

where f is the density for margin j ; x is a value on margin j ; $\mu_j^{(app)}$ and $\mu_j^{(adm)}$ are the means of margin j for the applicant and admittee pools, respectively; and $\theta_j^{(app)}$ and $\theta_j^{(adm)}$ are distributional parameters for the same.

The lower bound may be estimated as a one-dimensional minimization problem where the objective function takes a proposal mean (with fixed standard deviation), estimates margin parameters and returns the proposal mean if the above condition is met, and a large positive value otherwise. Once the bounds are estimated, one can simulate the applicant pool using any set of parameters in the feasible region. Figure 4 shows the marginal bounds for one of the schools in my sample.

Figure 4: Bounds on marginal distributions for school 1431.



5.3 Combining the applicant and admittee pools

To combine the enrollee and applicant pools, I match each individual in the former to one in the latter. Matched individuals have their admission indicators set to 1. All others have admission indicator values equal 0. Finding matches for the admittees can be done using one of the many one-to-one matching schemes that are available. For this implementation, I match to the closest non-admitted individual on Mahalanobis distance. Of course, which matching scheme is used can affect the resulting pool in a variety of ways. For example, simple matching as I do will not guarantee that the sum of distances across all matches will be the minimum possible (Diamond & Sekhon 2006). Nor does my choice guarantee preservation of moments or covariate balance. This potential complication will be investigated in future work.

5.4 Binning and estimating probability of enrolling

At this step, to simplify matters, I aggregate the SAT I component scores into a single composite score leaving two predictors - SAT I composite score and HSGPA. This simplification is expedient but not necessary; one could certainly proceed with the original 4 predictors (SAT I math, reading, writing and HSGPA) or even more, provided data were available. The choice is made to facilitate analysis and interpretation. Each pool of admits is binned along the two remaining dimensions for an arbitrary bin size - experimentation suggests that bins of 50 points on the SAT and 0.1 grade points yield a good balance between resolution and tractability. Within each bin, I take the ratio of admits to applicants yielding the probability of enrolling given that bin's corresponding combination of predictor values.

These probabilities are not the final figures of interest because enrollees are a non-random subset of admittees, with lower overall quality than the full admittee

pool. The intuition is that as a student's quality increases, the more likely he is to gain admission to a more selective school than the one under consideration and therefore not enroll. The relationship of interest is:

$$Pr(E = 1) = Pr(E = 1|A = 1, x)Pr(A = 1|x) \quad (0.1)$$

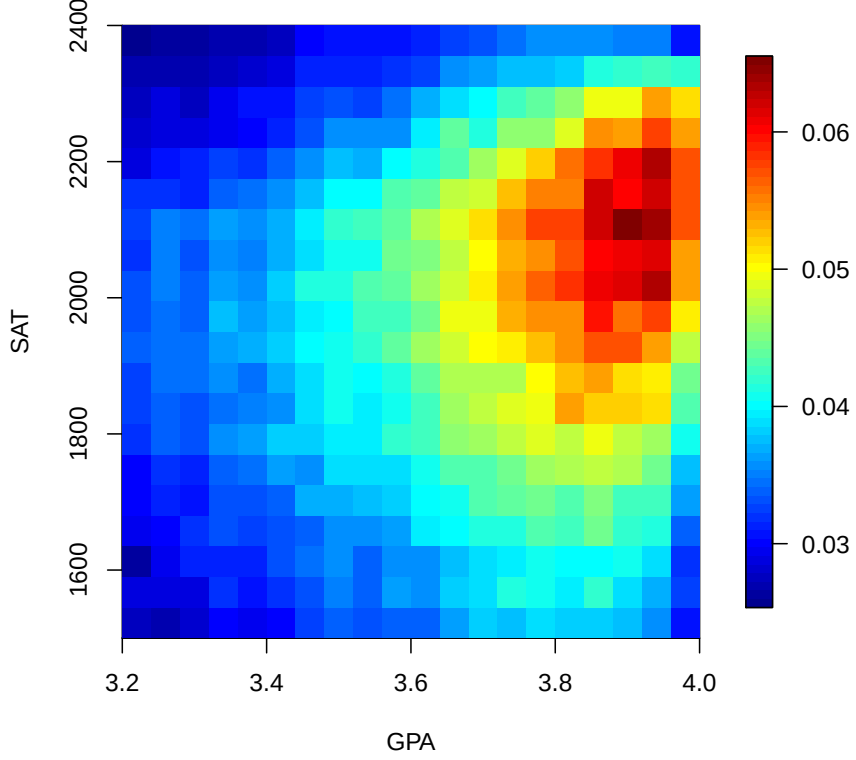
where x is student quality, operationalized here as a function of SAT I score and HSGPA. Figure 5, which is a plot of $Pr(E = 1)$ bin values for one school in my dataset, clearly validates this intuition. In the domain $SAT < 2050, GPA < 3.96$, the estimated probability increases as we would hope. However, it then peaks in the bin for $SAT = 2050, GPA = 3.96$ and decreases there on in both dimensions. Obviously, this is due to admitted students selecting out - not because admission itself gets more difficult with higher quality. Returning then to equation (0.1), what we need is an estimate of $Pr(A = 1|x)$ without explicitly knowing $Pr(E = 1|A = 1, x)$.

5.5 Modeling $Pr(A)$ on $Pr(E)$

I model probability of admission, $Pr(A)$, using the probability of enrolling, $Pr(E)$, on the basis of several qualitative observations about each quantity. First, regardless of the domain of student quality, there will always be some portion of admittees that do not enroll. Part of this will be stochastic (i.e. idiosyncratic reasons for not attending) and part will be systematic (i.e. a function of x) As a result, $Pr(E)$ will be less than or equal to $Pr(A)$ ⁶. The extent of their difference will monotonically increase with student quality because probability of not enrolling given admission has a positive relation with quality. This leads to a peak in $Pr(E)$ at some x_0 after which $Pr(E)$ declines due to increasing $Pr(E = 0|A = 1, x)$. In the extreme case where the school under consideration is the most desirable, no students select out and the only difference between $Pr(E)$ and $Pr(A)$ is due to

⁶This is clearly the case from $Pr(E = 1) = Pr(E = 1|A = 1, x)Pr(A = 1|x)$

Figure 5: $\Pr(E)$ for school 1325.



stochasticity; i.e. $x_0 \rightarrow \infty$. Conversely, as one approaches the lower bound on student quality, $\Pr(enroll|admit, x) \rightarrow 1$ and $\Pr(enroll) \rightarrow \Pr(admit|x)$. These observations lead to the following assumptions.

1. Probability of admission (A) increases monotonically with student quality (Q).

$$d\Pr(A)/dQ > 0$$

2. There is a positive relationship between student quality and selection out of the admittee pool. As a result, $\Pr(E)$ has a parabolic relationship with (Q).

$$d^2\Pr(E)/dQ^2 < 0$$

3. Probability of admission is strictly greater than probability of enrollment and the change in probability of admission w.r.t. quality is strictly greater than change in probability of enrollment.

$$Pr(A) > Pr(E)$$

$$dPr(A)/dQ > dPr(E)/dQ$$

Figure 6: Stylized illustration of assumptions that constrain $Pr(A)$ model selection

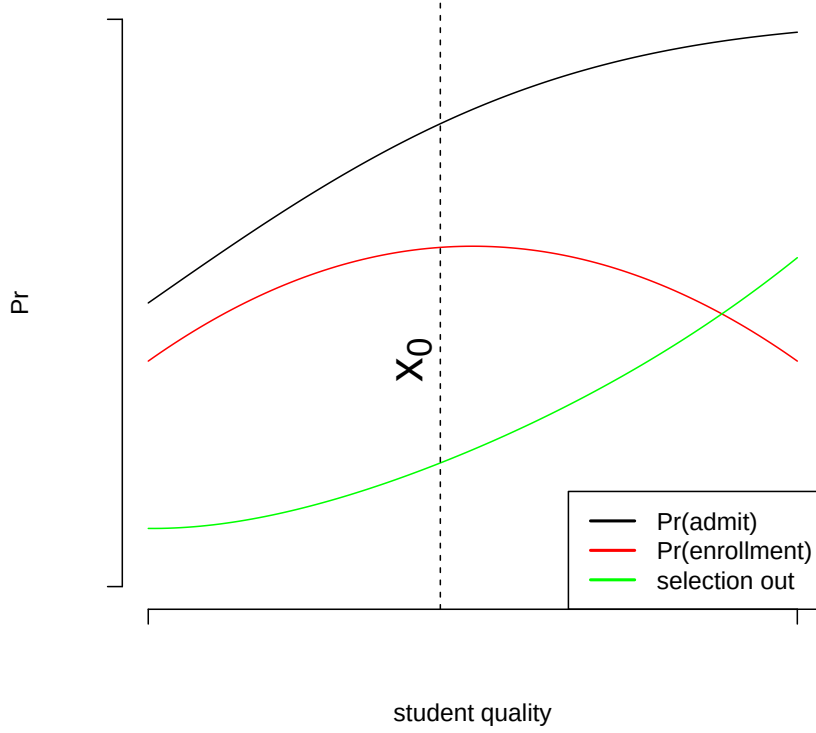


Figure 6 is a stylized illustration of the relationship implied by my assumptions. $Pr(A)$ can be modeled in any manner that satisfies these assumptions. One option is to simply regress $Pr(E)$ in bins $[1 : i_{max}, 1 : j_{max}]$ on the predictors, where bin i_{max}, j_{max} contains the maximum value of $Pr(E)$. In essence, this forces the relationship between the predictors and $Pr(A)$ to be positive, which is what we

want. Predictions may then be made using the regression model. Using this option, the link used should constrain the outcome to $[0,1]$.

Another possibility is to fit a non-standard parametric model to the data to reflect the user's beliefs about the relationship between the covariates and $Pr(A)$. An example might be the following:

$$\begin{aligned}\pi_{admit}(GPA|SAT) &= \Phi(GPA, \mu(SAT), \sigma(SAT)) \\ \mu(SAT) &= \alpha_\mu + \beta_\mu SAT^{-\gamma_\mu} \\ \sigma(SAT) &= \alpha_\sigma + \beta_\sigma SAT^{-\gamma_\sigma}\end{aligned}\tag{0.2}$$

Equations 0.2 model the probability of admission, π_{admit} , as following the normal CDF, Φ . Specifically, π_{admit} is a function of GPA with mean and standard deviation that are functions of SAT score. Experimentation suggests that μ and σ follow linear relationships with SAT^γ where the γ s can be found by minimizing SSR. The overall model reflects the belief that SAT and GPA substitute for each other in the admission decision, with a preference for average performance in both dimension rather than high performance in one (due to the functional forms of μ and σ). In the results section, I present the results of fitting this model. Whether this or any other, once a model has been fit, predicting probability of admission for some vector of student characteristics is simply a matter of plugging it into the chosen model.

5.6 Uncertainty

Within this framework, uncertainty is introduced at three stages: (1) when the parameters of the enrollee and population pools are picked, (2) while creating the applicant pool (i.e. simulating and matching) and (3) when fitting the $Pr(A)$ model. It's unclear whether error estimates can be made analytically, but they can certainly be made through bootstrapping. Referring back to outline 5, error

estimates are produced by repeating steps 2-6 (i.e. everything other than estimating parameter bounds) with random combinations from the feasible region. Once the $Pr(A)$ model has been fit for all of the simulations, confidence intervals can be estimated by predicting with each model and examining the distribution of results. This method can be used to establish CIs on individuals' predictions as well as the prediction surface implied by the $Pr(A)$ model. As an example,

Figure 7: 95% confidence interval and median prediction surface for school 544

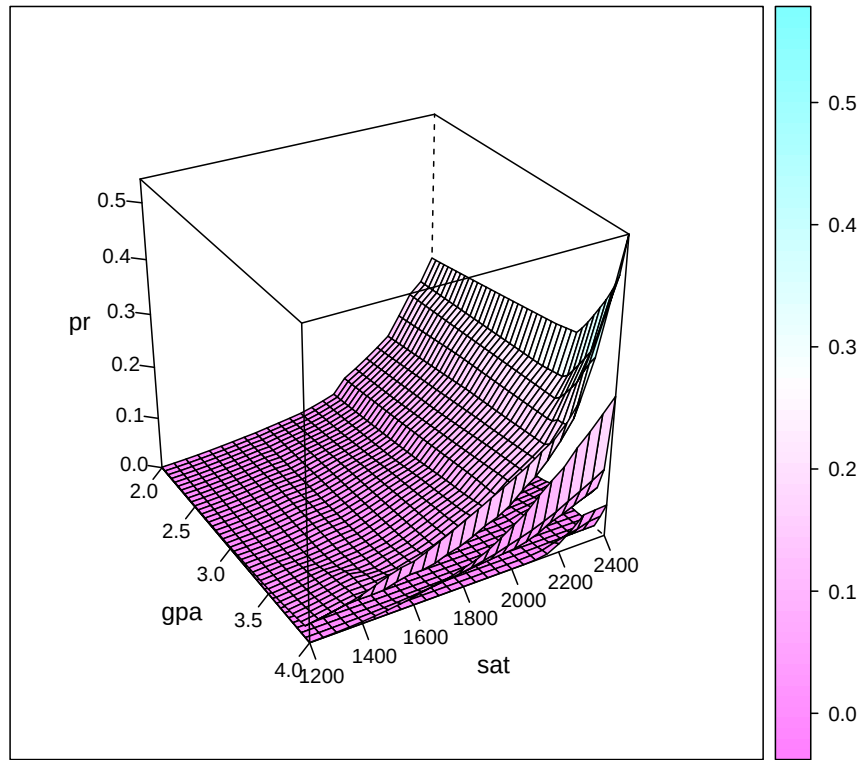


figure 7 shows the 95% confidence interval for the prediction surface of school 544. The predicted probabilities presented in the results section are made using the respective median prediction surface for each school.

5.7 Assessing predictive power

Assessing the predictive power of this model is to ask two different questions. First, how well do the resulting metrics predict application outcome? Second, given some group of students with $X\%$ predicted probability of admission, do $X\%$ of them actually gain admission?

To answer the first question, I use my model to predict admission probability for the individuals from the web survey. I then regress their admission decisions on my predictions in a simple logistic regression. The validity of approach would be supported if my predictions are shown to have a statistically significant effect on outcome. Unfortunately as I showed in tables 1-2, the web survey data suffer from considerable overrepresentation/misreporting of successful applications. The likely consequence is attenuation of predictive power. This is because my approach essentially predicts admission on less qualified applicants and extrapolates to better qualified students using a large number of parametric assumptions. As a result, regression analysis selected for students from the higher end of the spectrum may show my predictions to be less accurate than would be if a truly representative sample were used. If my predictions are statistically relevant despite this handicap, it would only strengthen the support they provide.

As a secondary test, I compare the predictive power of my model with that of a logistic model estimated on the biased data. Comparison is performed by means of cross-validation. I divide the survey data in half, estimate the logistic model on one half and predict on the other, and compare residual deviance of those predictions against those from predictions made using my approach. This process is repeated many times with random cuts of the data. It is worth noting that this test is inherently unfavorable to my approach because random cuts of a biased sample are still biased. My approach would be better if it yields predictions that

lead to greater reductions in deviance.

Unfortunately, without a representative sample of real student data, it is not possible to fully assess how accurate my purported measures of probability actually are. However, if such data were available, one way to answer the question would be to bin students by predicted probabilities and estimate actual admission in each bin. A truly accurate model would yield a 1 to 1 match between modeled and empirical probability in each bin. Running this test with (upwardly) biased data will mean that my model's predictions will tend to be lower than the probabilities implied by the survey data. This is because the students in the web data are presumably of higher quality than average for a particular combination of HSGPA and SAT scores.

Notwithstanding our inability to assess actual probability, there is still an expectation that if my model is predicting accurately, my model's predictions and the probabilities implied by the data should share a positive linear relationship. I present the results of this much weaker test in the results.

6 Results

I demonstrate my approach by testing it on a dozen schools ranging in admission selectivity, with official admission rates ranging from 7.7% to 62.8%. Table 3 shows the official information that is used to generate my models. As mentioned earlier, correlations for the joint distribution of SAT scores and GPAs were approximated using adjusted sample correlations from web surveys.

For each of the 12 schools in my sample, I estimate a model using applicant pool parameters that are at the Euclidean midpoint of the feasible region for each margin. For example, a school might be bounded 1600-1800 for mean SAT scores

school	applied	admitted	admitpct	satm25	satm75	satr25	satr75	satw25	satw75	gp1	gp2	gp3	gq1	gq2	gq3
544	27380	2095	7.70	690	800	700	780	690	790	0.09	0.04	0.01	0.94	0.88	0.81
1737	19323	1911	9.90	700	800	700	790	700	780	0.03	0.00		0.90	0.75	0.50
1022	6293	981	15.60	700	780	690	780	680	770	0.08		0.00	0.90	0.75	0.50
342	33073	6834	20.70	630	730	670	770	640	730	0.10	0.01	0.00	0.90	0.75	0.50
1407	50746	11960	23.60	560	680	600	730	570	690	0.07	0.04	0.02	0.75	0.70	0.65
1325	16739	4367	26.10	630	740	660	770	630	730	0.46	0.19	0.06	0.94	0.88	0.81
436	15366	4175	27.20	640	730	660	740	640	730	0.39	0.08	0.00	0.94	0.88	0.81
1503	20090	6999	34.80	600	700	610	700	590	690	0.05	0.03	0.01	0.75	0.70	0.65
1324	19606	7197	36.70	600	690	600	690	600	690	0.22	0.05	0.00	0.90	0.75	0.50
1583	18363	6735	36.70	600	710	620	730	610	720	0.08	0.03	0.01	0.75	0.70	0.65
1431	16871	9242	54.80	560	650	560	660	560	650	0.33	0.13	0.04	0.94	0.88	0.81
1593	22289	14004	62.80	550	670	610	710	560	670	0.45	0.23	0.10	0.94	0.88	0.81

Table 3: List of schools and relevant data. ‘admitpct’ is the percent admission rate. ‘satXY’ is the Yth quartile for SAT component X among enrollees (e.g. satm25 is the 25th quartile score for the math section). ‘gpX’ is the portion of enrollees with HSGPAs below 4 x ‘gqX’.

Figure 8: $\Pr(E)$ and $\Pr(A)$, schools 1-3

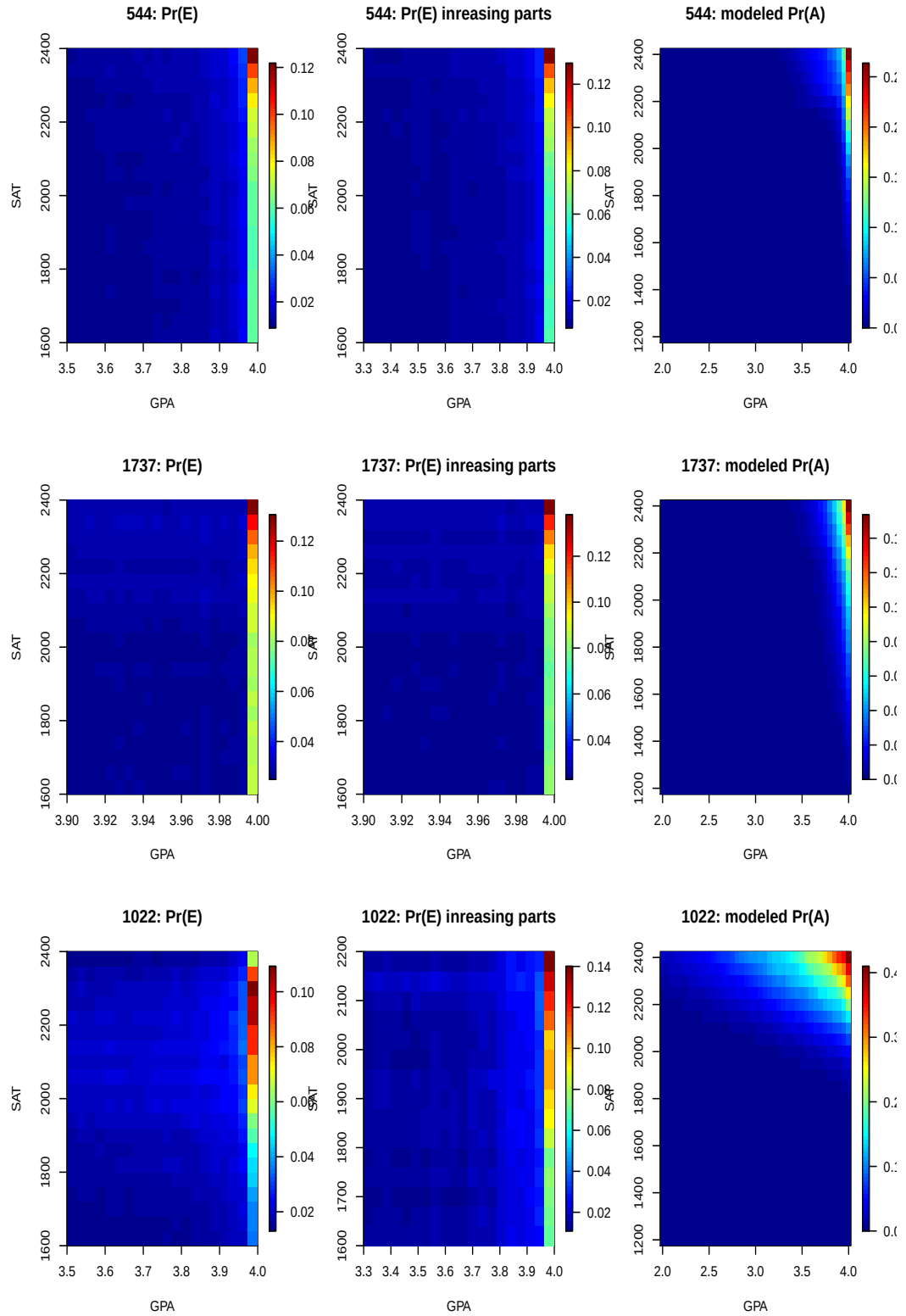


Figure 9: $\Pr(E)$ and $\Pr(A)$, schools 4-6

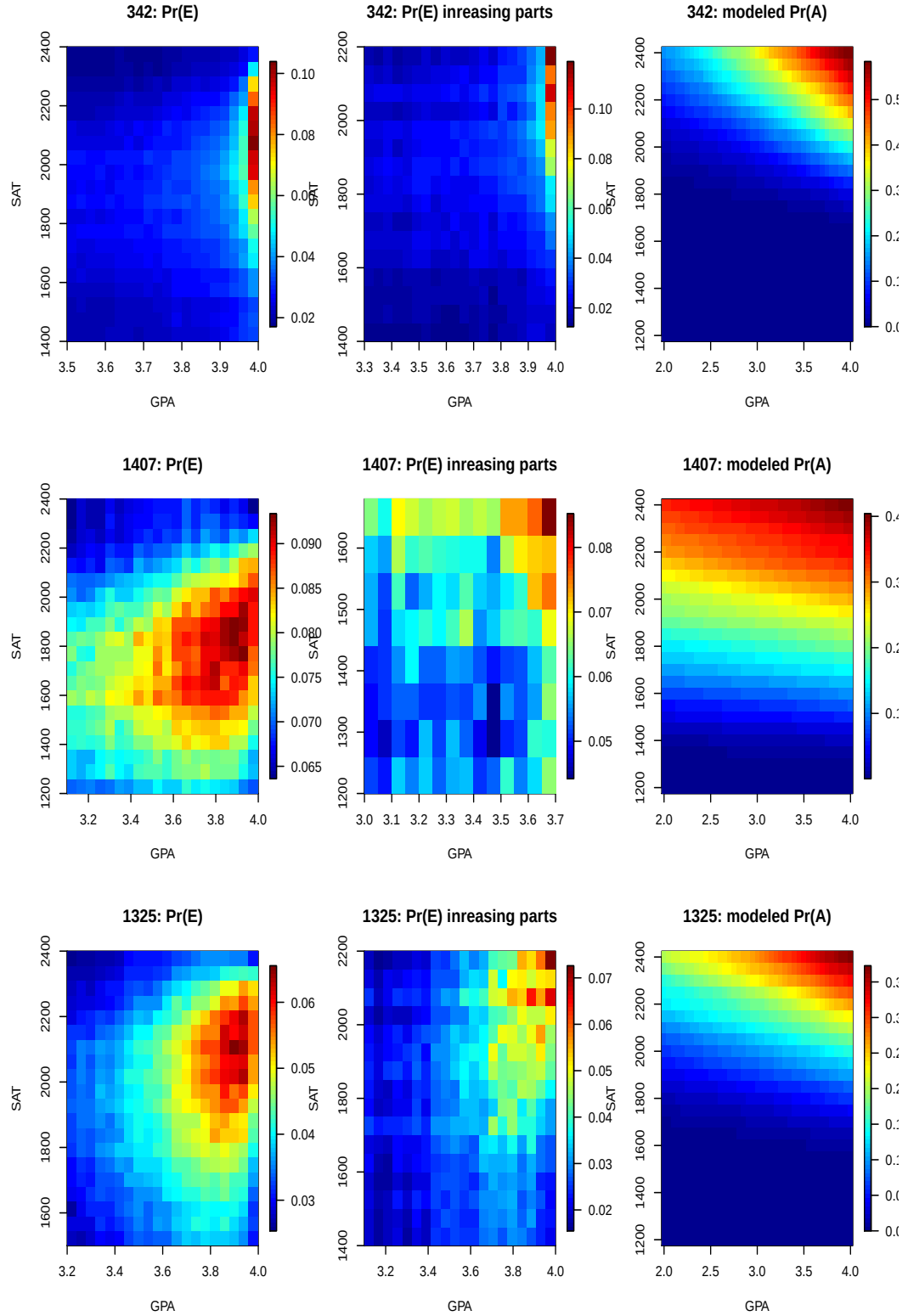


Figure 10: $\Pr(E)$ and $\Pr(A)$, schools 7-9

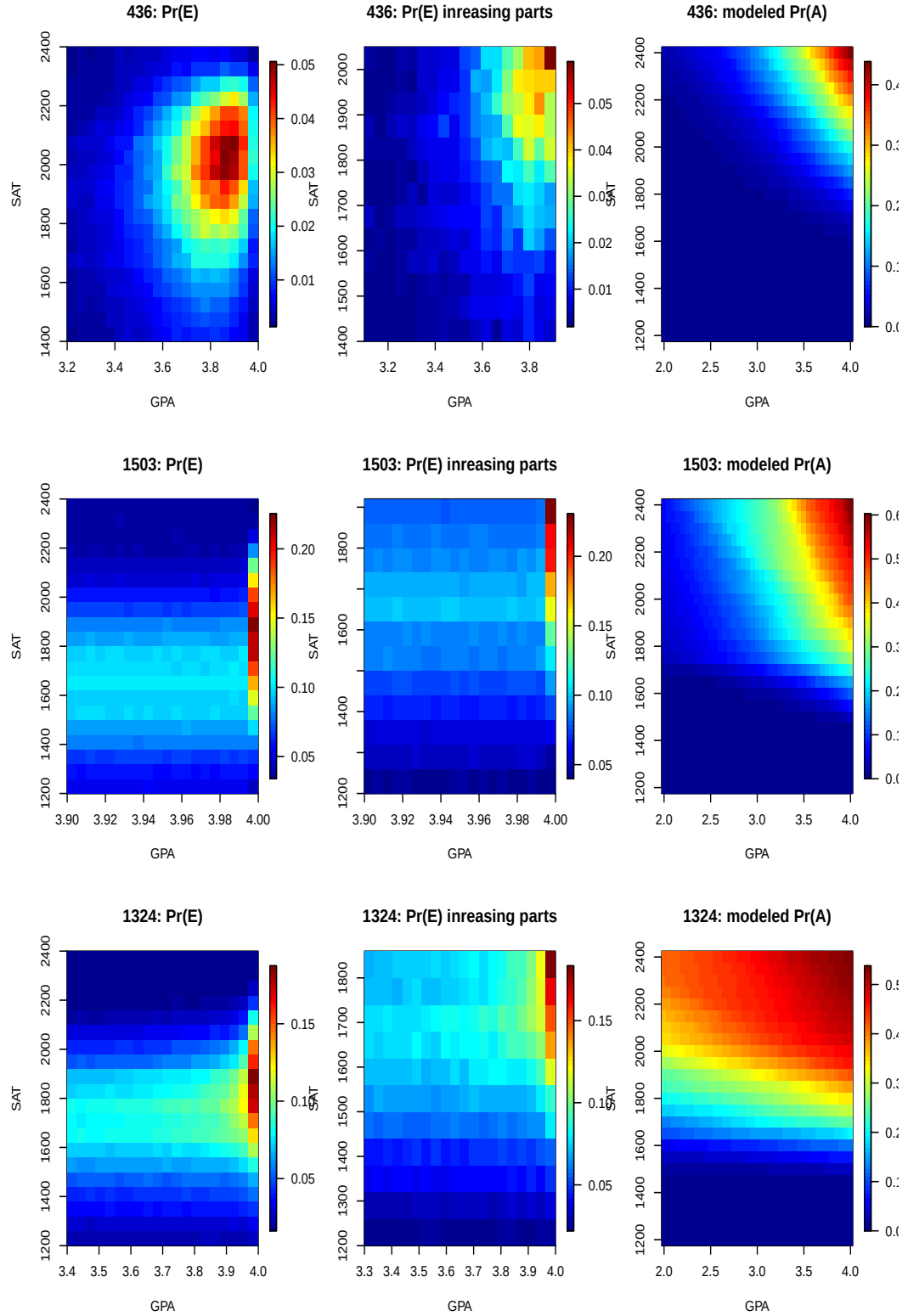
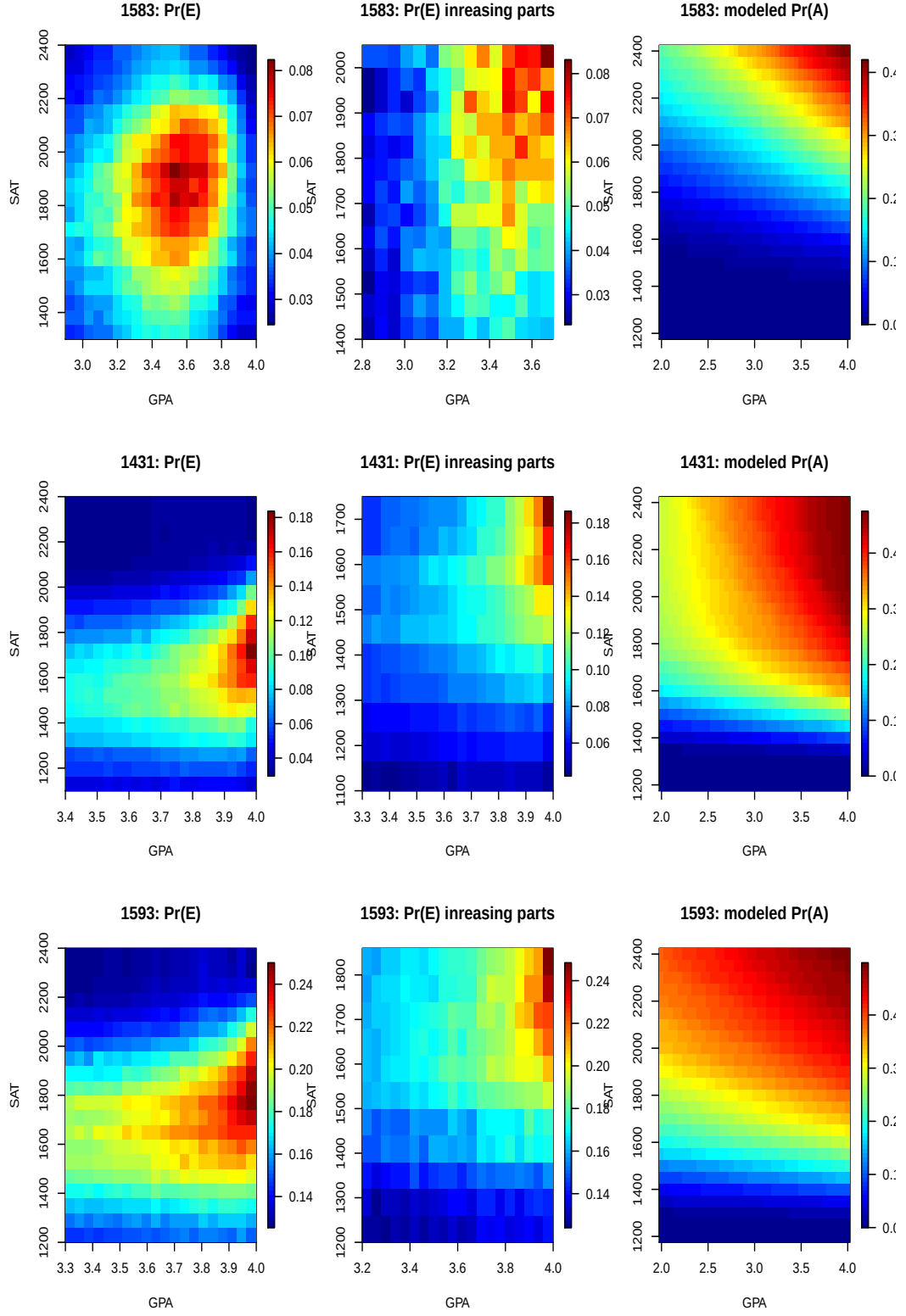


Figure 11: $\Pr(E)$ and $\Pr(A)$, schools 10-12



and 3.55-3.75 for HSGPA; I would then use 1700 and 3.65 as the means for the applicant pool. The following results are based on a simulation of 25,000 students for each school - this number is arbitrary and does not reflect actual application numbers.

The leftmost panels of figures 8 - 11 show the enrollment probabilities implied by the simulated data. The middle panels focus specifically on the parts of that data that are informative with respect to admission, where $Pr(E)$ more or less monotonically increases with predictor values. As seen in the middle panels, all the schools except for 544 and 1737 fit model 0.2 to a subset of the full range of possible SAT and HSGPA values. This reduction of the data that are used to fit the $Pr(A)$ model decreases with selectivity. School 1407 with an admission percentage of 23.6%, for example, is fit to data in the domain $GPA < 3.8$ and $SAT < 1900$. School 1583 which admits 36.7% of applicants is another step down, using $Pr(E)$ data only where $GPA < 3.5$ and $SAT < 1800$. As will be seen later, this has an impact on predictive power, with less selective (w.r.t. admission rate) schools being more sensitive to noise in the simulation data because there is more extrapolation done.

Finally, the rightmost panels of figures 8 - 11 show the predicted surfaces of $Pr(A)$ that are extrapolated from the middle panels by fitting model 0.2. One of the interesting takeaways from these figures is the variation across schools of the effect of SAT score and HSGPA on $Pr(A)$. Some of the schools display a substitutive relation between the two, where high performance on one may mitigate lower performance on the other. This is seen in schools like 544, 1737, 1022, and 1583 where the contours are essentially linear. Other schools have what appear to be threshold SAT scores, where the gains in $Pr(A)$ with increasing score substantially decrease after the threshold. Examples of such schools include 1503, 1431, and 1324 to some extent. Of course this may be an artifact of the model that is fit,

but it could also reflect the way admission is handled at these schools. It would be interesting to corroborate this inference with interviews with admission officers from these schools.

To assess the predictive power of my approach, the models estimated above are now used to predict admission probability for a sample of web survey data. These data contain 5 variables: 3 SAT component scores, HSGPA, and admission result. N's range between 417 and 3146 observations for the 12 schools modeled above. I begin by regressing admission result on my model's predictions to see whether they are statistically and substantively significant.

Table 4: Predictions.

	544	1737	1022	342	1407	1325
(Intercept)	0.021	0.062	0.172	0.110	−0.365**	0.148
	(0.777)	(0.433)	(0.322)	(0.285)	(0.010)	(0.377)
mine	1.764**	2.283*	1.875*	0.981***	2.983***	1.581*
	(0.001)	(0.019)	(0.036)	(0.000)	(0.000)	(0.021)
N	1866	1992	417	2352	3146	1223
null dev.	2566.619	2738.464	555.313	3144.643	4170.892	1618.566
residual dev.	2555.839	2732.982	550.834	3131.549	4130.521	1613.242
$\log L$	−1271.920	−1360.491	−269.417	−1559.775	−2059.260	−800.621

p values in parentheses

[†] significant at $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 5: Predictions (continued).

	436	1503	1324	1583	1431	1593
(Intercept)	0.456**	−0.129	0.306	−0.072	0.228	0.016
	(0.009)	(0.567)	(0.288)	(0.704)	(0.689)	(0.972)
mine	1.711*	1.760***	1.106	2.529***	1.858	2.190*
	(0.031)	(0.001)	(0.104)	(0.000)	(0.170)	(0.042)
N	594	1014	746	1362	555	616
null dev.	738.441	1319.223	934.201	1767.695	645.720	737.068
residual dev.	733.721	1307.931	931.594	1753.594	643.895	733.020
$\log L$	−360.861	−647.966	−459.797	−870.797	−315.947	−360.510

p values in parentheses

[†] significant at $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Tables 4-5 shows the results of regressing admission on the predicted probabilities. For this particular set of simulations and model fits, there is a statistically significant positive relationship between my predictions and actual outcomes for 10 of the 12 schools ($\alpha = 0.05$). Across different iterations of this process, I found that the number of schools with statistically significant prediction values ranged from 9 to 12 (all), with my model consistently underperforming for less selective schools. This again is due to the fact that $Pr(E)$ peaks at a lower combination of predictor values and therefore extrapolation occurs for a greater range of values.

I perform my cross-validation test by executing the following steps 100 times for each school:

1. Select a random half of the biased data
2. With the selected half, regress application result on SAT scores and HSGPA

3. Predict results for the remaining half using the model from step 2
4. Predict results for the remaining half using my model
5. Compute reduction in deviance (null deviance - remaining deviance) for each set of results
6. Compute the reduction ratio $\frac{d_{mine}-d_{logistic}}{d_{logistic}}$ where d_{mine} and $d_{logistic}$ are the reductions in deviance resulting from my model and the logistic model, respectively.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
544	0.51	1.38	2.69	4.90	4.97	26.22
1737	0.04	0.79	2.14	3.40	4.04	18.11
1022	-0.03	0.89	2.07	9.18	10.21	101.40
342	-0.37	-0.27	-0.14	-0.12	-0.02	0.36
1407	-0.49	-0.43	-0.38	-0.37	-0.30	-0.16
1325	-0.61	-0.44	-0.29	0.10	0.26	3.87
436	-0.64	-0.47	-0.33	-0.18	-0.11	2.52
1503	-0.25	-0.13	-0.02	0.06	0.16	0.90
1324	-0.66	-0.34	0.02	0.97	1.09	12.26
1583	-0.52	-0.44	-0.36	-0.31	-0.21	0.15
1431	-0.89	-0.27	0.20	1.25	1.12	15.09
1593	-0.80	-0.69	-0.54	-0.39	-0.12	0.69

Table 6: Reduction ratios.

Table 6 summarizes the distribution of reduction ratios for each school resulting from the above steps. For example, a reduction ratio of 0.6 would indicate that my model explained 60% more of the deviance than the logistic model. Positive figures are desirable since they indicate that the deviance reduction from my predictions

is greater than that from the comparable logistic model.

The results of this assessment are mixed. On the one hand, my approach performs consistently better for several schools. For example, my model shows a median deviance reduction for school 544 that is over 2.7 times greater than that of the logistic model. Similar points can be made about schools 1737, 1022, 1324 and 1431. While the remaining schools show negative median reduction ratios, significant portions of their distributions of reduction ratios appear to straddle the origin. This means that they often perform better than the logistic model as well, though not a majority of the time. For these schools, my model performs -24% worse on average. While this is not ideal, this underperformance does not seem egregious considering the naïve models are estimated on data that are similarly biased to the out-of-sample data while my model is tailored to the unrestricted population.

Figure 12: Empirical versus predicted $\Pr(A)$, schools 1-6. Dotted line indicates what we'd see if my predictions were a perfect 1 for 1 fit.

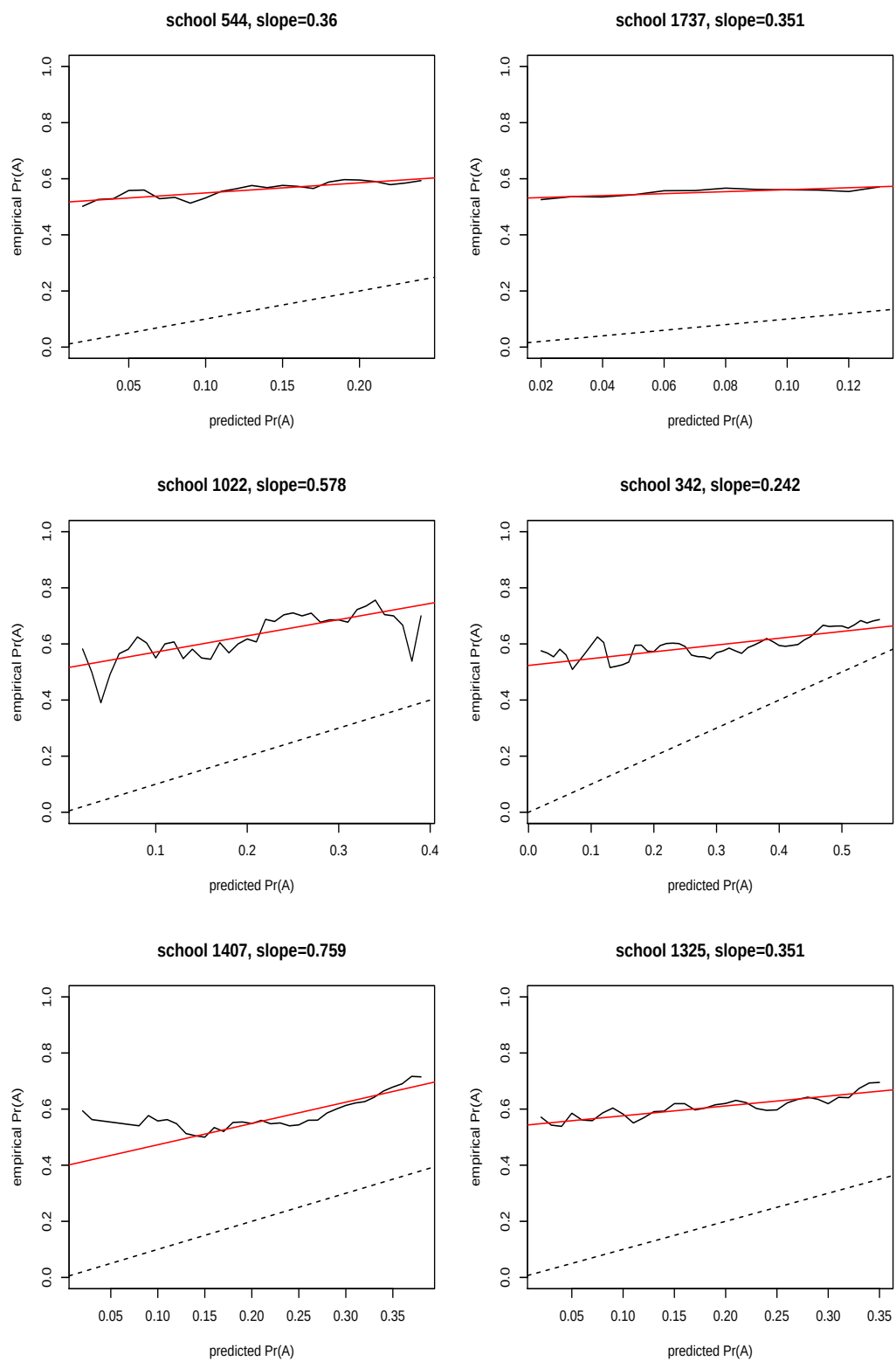
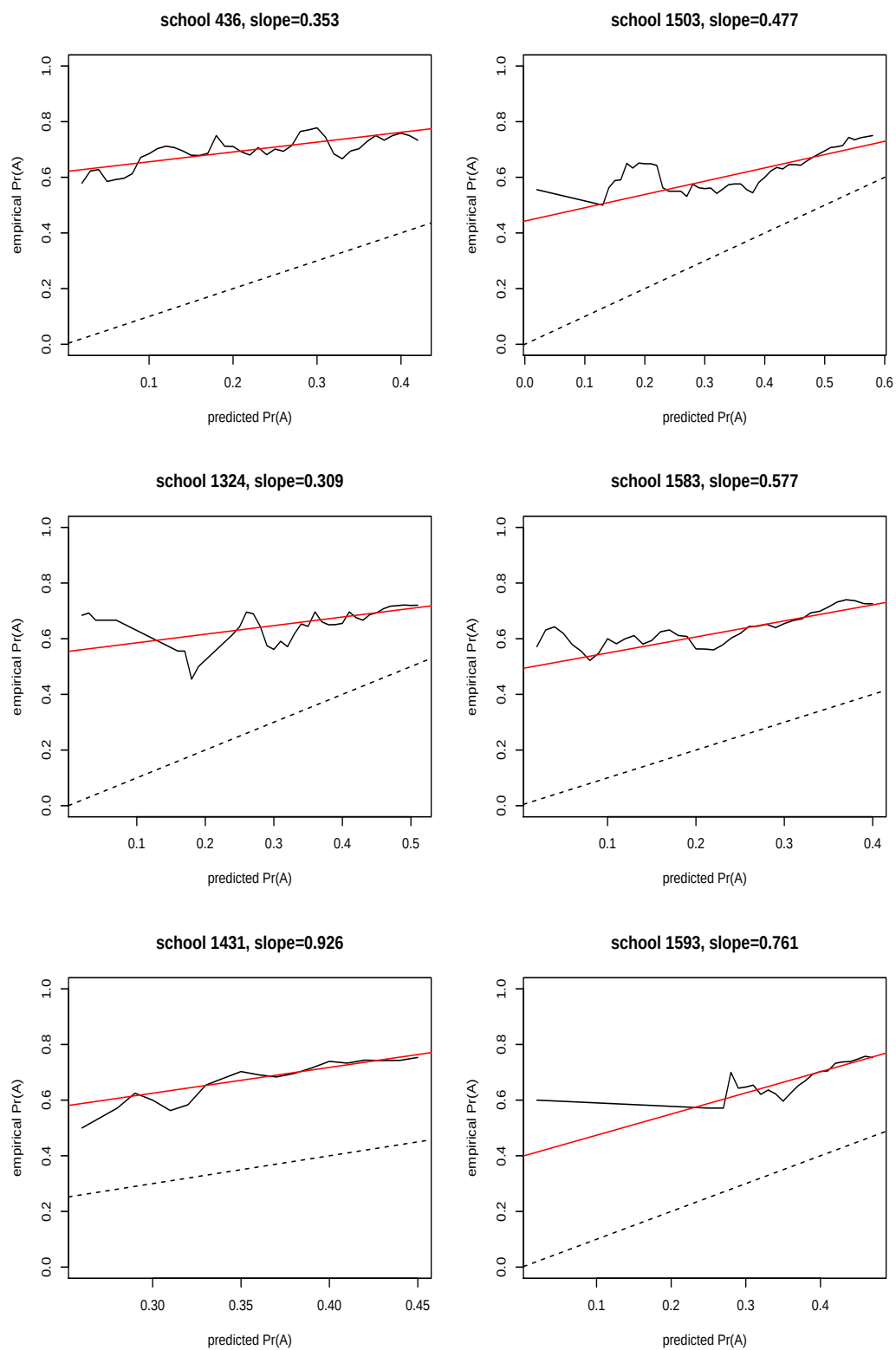


Figure 13: Empirical versus predicted $\Pr(A)$, schools 7-12. Dotted line indicates what we'd see if my predictions were a perfect 1 for 1 fit.



Finally, figures 12 and 13 show the relationship between my predicted $Pr(A)$ s and those implied by the survey data for all 12 schools in my sample. In each figure, the dotted black line is the relationship we would see if predicted and empirical values lined up perfectly. The solid red line is the relationship implied by a weighted least squares regression of empirical $Pr(A)$ on predicted $Pr(A)$ ⁷. As expected, my models drastically underpredict for the available data. In school 544, for example, my model yields predictions that are 43% lower on average. As mentioned earlier, using the available data, it cannot be ascertained how much of this inaccuracy is due to biases present in the data and how much results from the (poor) quality of the model.

On the other hand, all the schools show a strong positive relationship between predicted $Pr(A)$ and empirical $Pr(A)$. The slope of the regression of empirical $Pr(A)$ on predicted $Pr(A)$ yields slope ranging between 0.242 and 0.926. Certainly these figures are not perfect (i.e. slope of 1); however, they do suggest that my models do a good job of assessing relative chance of admission. This test does not validate the exact assumptions used (e.g. model 0.2, particular parameter values) but it does suggest that the overall approach has merit and provides one approach for future work on this topic, particularly if some unbiased student data does become available.

7 Conclusion

In this thesis, I proposed a framework for modeling college admission probability using official data available to the general public. This framework provides estimates of $Pr(A)$ as well as confidence intervals on those estimates. The main merit of my approach is that it avoids the naïve use of biased data, a sin which

⁷WLS is used instead of OLS because of severe heteroscedasticity. The numbers of observations in each bin are used as weights.

approaches relying solely on survey data are almost certainly guilty of. I showed that with the right assumptions and modeling choices, my approach is capable of rendering predictions that are statistically and substantively significant. Unfortunately, in the absence of verifiably real individual admission data, it is not possible to rigorously test my approach. Nonetheless, I showed through a series of tests that the predictions yielded by my model possess many of the properties we would expect if they were accurate.

While the $Pr(A)$ model I used here (equations 0.2) performed admirably, it clearly had at least 3 obvious shortcomings. For one thing, fitting covariate margins to a normal CDF was an arbitrary choice based solely on desired response values. Further investigation might compare that with other functional forms for the margins. Another issue is that there isn't an obvious way to generalize this particular model to more than two predictor dimensions. Multiple regression obviously solves this problem but does so at the cost of flexibility in functional form. If more predictors (e.g. HSGPA and each of the SAT components separately) are to be considered using a non-standard model, a more generalizable one will need to be developed. Finally, given the assumed relationship between $Pr(A)$ and $Pr(E)$, this model will consistently underpredict $Pr(A)$ since it extrapolates $Pr(E)$ without fully accounting for $d^2Pr(E)/dQ^2$.

A shortcoming on the estimation side of things was that the $Pr(A)$ model was estimated in a single sequential iteration: I first fit the marginal PDFs, then estimated $\mu(SAT)$ and $\sigma(SAT)$ on the margin parameters. Better fits to the data may be achieved by implementing an MCMC solution where I fit the model over many iterations allowing me to home in on the best fit. A final topic for future investigation is how to pick the correct parameters for the applicant pool. Here, I took the median prediction surface yielded by 100 simulations with randomized combinations of applicant pool parameters. I did not address the question of how

to infer what the ‘true’ parameters actually are. It may be possible to do so by optimizing for some desirable property of the model - minimizing prediction error for example. This exercise is not undertaken here solely for the reason that there would be no way to validate the resulting numbers. Again, if real data are made available, this is certainly a worthwhile avenue for investigation.

From a substantive point of view, it is unclear how much better the predictions afforded by my model are compared to predictions based on biased survey data. As I showed in table 6, the performance of my approach varies across schools and is somewhat correlated with school selectivity. The latter issue may be mitigated with different modeling choices, but ultimately, a complete assessment is still prevented by lack of data. In one sense, this is not that large of a problem: the average user will not care whether his true probability of admission is 0.25 or 0.35 so long as the estimate is in the right ball park, and therefore good enough to inform decision-making. If nothing else, my model yields estimates of admission probability that are (largely) independent of those found on online APCs⁸. It can therefore serve its purpose by providing a second opinion to inform student behavior.

8 Sources

1. Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, 27(2), 187-189.
2. Belgorodski N., Greiner, M., Tolksdorf, K., Schueller, K. “Package `riskDistributions`” (Reference Manual). Retrieved from <http://cran.r-project>.

⁸This implementation did use school correlation matrices estimated from survey data, but no other information was used from those sources.

3. Berk, R., Freedman, D., & Stark, P. (2002). SATs and UCLA Admissions. Unpublished.
4. Birnbaum, Z. W., Paulson, E., & Andrews, F. C. (1950). On the effect of selection performed on some coordinates of a multi-dimensional population. *Psychometrika*, 15, 2, 191-204.
5. Bryant, N. D., & Gokhale, S. (1972). Correcting Correlations for Restrictions in Range Due to Selection on an Unmeasured Variable. *Educational and Psychological Measurement*, 32, 2, 305-10.
6. Clark, M. A., Rothstein, J., Schanzenbach, D., & National Bureau of Economic Research. (2008). Selection bias in college admissions test scores. Cambridge, Mass: National Bureau of Economic Research.
7. Couper, M. P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, 464-94.
8. Diamond, A., Sekhon, J. S. (2006). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. UC Berkeley: Institute of Governmental Studies.
9. Embrechts, P., Lindskog, F., McNeil, A. (2001). Modelling Dependence with Copulas and Applications to Risk Management. Unpublished.
10. Espenshade, T. J., Chung, C. Y., & Walling, J. L. (2004). Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities. *Social Science Quarterly*, 85, 5, 1422-1446.
11. Espenshade, T., Hale, L., & Chung, C. (2005). The Frog Pond Revisited: High School Academic Context, Class Rank, and Elite College Admission.

Sociology of Education, 78, 4, 269-293.

12. Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. Hoboken: John Wiley & Sons.
13. Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47: 153-161.
14. Hull, J. (2012). *Risk management and financial institutions*. Hoboken, N.J: John Wiley & Sons.
15. Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., Barbuti, S. M. (2008). *Validity of the SAT for Predicting First-Year College Grade Point Average*. College Board Research Report No. 2008-5.
16. Levin, J. (1972). The Occurrence of an Increase in Correlation by Restriction of Range. *Psychometrika*, 37, 1, 93-97.
17. Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., Barbuti, S. M. (2008). *Differential Validity and Prediction of the SAT*. College Board Research Report No. 2008-4.
18. Obenchain, R. L., & Melfi, C. A. (1997). Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies. *Proceedings-Biopharmaceutical Section American Statistical Association*, 297-306.
19. Puhani, P. (2000). The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys*, 14, 1.
20. Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41-55.

21. Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, 39, 1, 33-38.
22. Sackett, P. R., & Yang, H. (January 01, 2000). Correction for range restriction: an expanded typology. *The Journal of Applied Psychology*, 85, 1, 112-8.