2007-01-02

Regression Analysis of University Giving Data

Yi Jin
Worcester Polytechnic Institute

Regression Analysis of University Giving Data

by

Yi Jin

A Project Report

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

____

December 2006

APPROVED:

____

Joseph D. Petruccelli, Advisor

____

Bogdan M. Vernescu, Department Head

To My Parents

Abstract

This project analyzed the giving data of Worcester Polytechnic Institute's alumni and other constituents (parents, friends, neighbors, etc.) from fiscal year 1983 to 2007 using a two-stage modeling approach. Logistic regression analysis was conducted in the first stage to predict the likelihood of giving for each constituent, followed by linear regression method in the second stage which was used to predict the amount of contribution to be expected from each contributor. Box-Cox transformation was performed in the linear regression phase to ensure the assumption underlying the model holds.

Due to the nature of the data, multiple imputation was performed on the missing information to validate generalization of the models to a broader population.

Concepts from the field of direct and database marketing, like "score" and "lift", were also introduced in this report.

Acknowledgments

First and foremost, I would like to thank Dr. Joseph D. Petruccelli, my academic and project advisor, for his guidance and patience throughout this project. The illuminating ideas from each meeting, although at certain stage made the project seem endless, turned out to be the most fun and rewarding part of this exploration.

I also want to thank the people at the Office of Development and Alumni Relations, especially Mr. Dexter A. Bailey Jr. (Vice President for Development and Alumni Relations) and Ms. Lisa Corinne Maizite (Assistant Vice President for Development) for their consent in granting me access to the data as well as faith in recruiting me for the examination of the file.

Thanks also go to Dr. Jason D. Wilbur and Dr. Balgobin Nandram, whose excellent teaching helped me step into a wonderful world of advanced statistics and be better prepared for this capstone project and things beyond.

The internship I had in Epsilon Data Management over the summer, also under the supervision of Dr. Petruccelli, opened my eye to the statistical application in business world and practically led to the Development Office's sponsorship of this project. I enjoyed the three months spent with the Epsilon team of industrial statisticians and look forward to joining them after graduation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Project Overview

1.1.1 Background

As a private institution, Worcester Polytechnic Institute (WPI) has relied on the generosity of its alumni, parents and many friends to help provide the fundamental support that enhances the school's overall operations since its very founding in 1865.

The Office of Development and Alumni Relations (Development Office) is the university administrative unit that has as one of its missions reaching out to the community to secure financial support for the institution.

Since WPI had its database system computerized in 1983, information has been collected on the giving history plus other aspects of the university's alumni and broader constituents (parents, neighbors, foundations, etc.). With the accumulation of data and the recognition of statistical analysis techniques, the Development Office initiated a project to examine the giving patterns quantitatively in an effort to achieve deeper understanding of the constituents and better results in its solicitation efforts. The Center for Industrial Mathematics and Statistics (CIMS) at WPI's Mathematical Sciences Department was invited to partner in the project.

1.1.2 Expectations

The records include constituents who have given to the school, whom we will call *contributors*, as well as those who have not given, whom we will call *prospects*. The two main questions for which the Development Office is seeking answers are:

1) *What are the characteristics that distinguish contributors from prospects?* and
2) *What are the key factors that drive the contributors' amount of contribution?*

By answering the first question, the office is hoping to obtain a clearer image of a "typical" contributor and prospect, along with a set of predictors effective in identifying prospective contributors. The answer to the second question will lead to more effective allocation of resources and increased magnitude of support.

1.2 Data Description

The original data file was extracted by WPI's Computing and Communications Center (CCC) from the "Banner" system and delivered in the format of Microsoft Excel spreadsheet. A quick initial data browsing was then done followed by meetings with Ms. Lisa Maizite of the Development Office, and Ms. Paula Delaney and Mr. Kevin Sheehan of CCC to discuss quality issues and place further requests. Based on these meetings, an updated version of the data was prepared and used for this project.

1.2.1 Data Overview

The data set consists of 48,604 observations (constituents) and 102 variables. A data dictionary was also supplied. The file includes all living WPI constituents and their gifts recorded in the computerized "Banner" system beginning in 1983. The values for 1983 represent the cumulative giving up to the end of that fiscal year. After 1983, the yearly gift data and giving club membership are listed by fiscal years.

1.2.2 Data Dictionary

Explanations for the 102 original variables are presented in Table 1.1.

Table 1.1 Original Data Extract Key

1	PERSON_NUM	Person number for data extract
2	CATEGORY	See **Table 1.2**
3	GENDER	M/F/NA
4	BIRTH_YEAR	4-digit year of birth
5	MARRIED	Married/Single/etc.
6	LEGACY	Yes: the person's admission record indicated a legacy relationship (no details available)
7	GPA**[1]**	Numbers for those available, spaces for those unavailable, "N/A" for those not applicable
8	BS_YEAR	WPI B.S. year
9	BS_MAJOR	WPI B.S. major
10	MS_YEAR	WPI M.S. year
11	MS_MAJOR	WPI M.S. major
12	PHD_YEAR	WPI Ph.D. year
13	PHD_MAJOR	WPI Ph.D. major
14	CERT_YEAR	WPI certificate year
15	CERT_MAJOR	WPI certificate major
16	HONOR_YEAR	WPI honorary degree year
17	HONOR_DEG	WPI honorary degree
18	NON_WPI_DEG	value if known (formatted as institution : degree code : year : major)

19	WPI_SPS	Yes: the spouse is a constituent
20	NUM_OF_CHILD	Count of children
21	PREF_CLAS	Preferred class year
22	HAD_SCHOLARSHIP	Yes: had scholarship while at WPI
23	PRES_FND	Yes: a Presidential Founder
24	LIFETIME_PAC	Yes: a lifetime PAC**[2]** member
25	TRUSTEE	Yes: a trustee of WPI
26	ADM_VOL	Yes: involved in alumni/admissions
27	CLS_AGENT	Yes: involved in a solicitation structure
28	REUNION	Yes: constituent attended reunion(s)
29	ALUM_VOLUNTEER	Count of distinct number of activities (involved in/as department advisory board, gold council, …, 42 possibilities)
30	ALUM_CLUB	Count of distinct number of activities (Tech Old Timers, Polyclub, …)
31	ALUM_LEADER	Count of distinct number of activities (involved in/as class officer, trustee search committee, fund board, …, 30 possibilities)
32	FRAT	Name of fraternity/sorority, blank otherwise
33	SPORT_COUNT	Count of varsity sports listed
34	VARSITY_SPRTS	Concatenated list of varsity sports
35	WPI_AWD	Yes: constituent received this award at WPI
36	TAYLOR_AWD	Yes: constituent received this award at WPI
37	SCHWIEGER_AWD	Yes: constituent received this award at WPI
38	GODDARD_AWD	Yes: constituent received this award at WPI
39	GROGAN_AWD	Yes: constituent received this award at WPI
40	BOYNTON_AWD	Yes: constituent received this award at WPI
41	WASHBURN_AWD	Yes: constituent received this award at WPI
42	RES_CITY	Home city (permanent address)
43	RES_STATE	Home state code
44	RES_ZIP	Home zip code (5 or 9-digit format)
45	RES_COUNTRY	Home country
46	TITLE	Job title if known, blank if unknown
47	WORK_CITY	Work city (business address)
48	WORK_STATE	Work state code
49	WORK_ZIP	Work zip code (5 or 9-digit format)
50	WORK_COUNTRY	Work country
51	STU_CLUB	Count of clubs (Outing Club, Science Fiction, Sport Parachute, …)
52	STU_ARTS	Count of arts and literature organizations (Masque, Pathways, Peddler, …)
53	STU_INTL_CLUB	Count of international clubs (Indian Students Association, …)

54	STU_CLUB_SPORT	Count of club sports (scuba, bowling, autocross, …)
55	STU_PROF_SOC	Count of undergrad professional societies
56	STU_MUSIC	Count of music band: glee club, baker's dozen …
57	STU_CLS_OFF	Count of class officer (freshman, sophomore, …)
58	STU_SCH_INVOLVE	Count of school involvement (student activities board, resident advisor)
59	STU_SPEC_PROG	Count of special programs (undergraduate employment program, exchange, …)
60	STU_INTRAMURAL	Count of intramural sports (basketball, softball, table tennis, …)
61	STU_HONR_SOC	Count of honor societies (Pershing Rifles, Sigma Mu Epsilon, Skull, …)
62	STU_PROJECT_CTR	Project center info (from the student courses)
63	ALU_PROJECT_CTR	Project center info (from alumni activities)
64	GRAD_DISTINCTION	H: graduated with high distinction, D: graduated with distinction, and blank
65	ALUM_CONTACTS	Contacts made as an alumnus (phone calls, personal visits, …)
66-90	FISCAL_YEAR_X (X: 1983~2007)	Total gift and memo for the specific fiscal year**[3]**
91-102	GIFT_CLUB_X (X: 1996~2007)	gift club designation for the specific fiscal year

[1] WPI undergraduates do not have a "true" GPA. Standard "numerical equivalent for passed courses" approved by the faculty was used.

[2] PAC stands for President's Advisory Council.

[3] Note the 1983 number is a cumulative amount given up through 1983 as the values were loaded into "Banner".

Each of the constituents is assigned a best (primary) category. The supplied dictionary lists 37 distinct categories, but only 18 of them are present in the data. The four letter codes of these 18 categories and their definitions are given in Table 1.2 along with their frequencies and percentages in descending order of size.

Table 1.2 Constituent Category and Distribution

Code	Category	Count	Percentage
ALUM	Alumna/Alumnus	24,027	49.43%
PRNT	Parent	10,601	21.81%
GRAD	Graduate Alumnus	4,782	9.84%
FRND	Friend	3,435	7.07%
WIDO	Widow/Widower	1,867	3.84%
CERT	WPI Certificate Recipients	1,207	2.94%
GPAR	Grandparent	770	1.58%
ALND	Non-degreed Alumna/us	646	1.33%
FACT	Faculty/Staff	445	0.92%
NEIG	Neighbor	319	0.66%
MPAR	Mass Academy Parent	311	0.64%
HOND	Honorary Degree Recipient	85	0.17%
STDT	Student	44	0.09%
HONA	Honorary Alumna/us	32	0.07%
TRUS	Trustee	19	0.04%
OTHR	Other Organizations	12	0.02%
FFOU	Family Foundation	1	0.00%
TRNS	Pre-Banner Class Transfer	1	0.00%

1.2.3 Quality Concerns

One concern regarding data quality comes from the high percentage of missing (blank) values across the file. As an example, the variable about job title has 68.7% null cells. Most of these cases are due to the fact that these types of information were collected on a self-report basis -- the constituents have no obligation of responding to such inquiries. Another issue arises from the confounding of responses, primarily seen in those variables with values extracted from the database as either yes or null (blank). While yes assures us a confirmative response, blank in many cases does not necessarily mean no: it simply means no answer was given.

These problems along with the messy (i.e. literally impossible to categorize) values in variables like "Job Title" and "Non-WPI Degree" brought a challenge for variable recoding.

1.2.4 Modeling Data

For analysis and modeling purposes, the data were divided into two groups: current plus former WPI students, and all others. Furthermore, in the "student" group, undergraduate, graduate and non-degree alumni (of categories ALUM, GRAD and ALND) form an especially desirable subgroup characterized by the most complete information across variables, which leads to the expectation of highest predictive power. The remaining categories in this group, certificate recipients and current students, appear to be less attractive in terms of modeling since they lack certain information due to the nature of the categories. Table 1.3 shows a pre-analysis grouping of the 102 original variables based on the type of information they contain. Table 1.4 then displays the completeness of information for the subgroups.

Table 1.3 Pre-analysis Grouping of the Original Variable

Variable Group	Original Variables			Count
Identifier	PERSON_NUM			1
Biographical Information	CATEGORY GENDER BIRTH_YEAR MARRIED NUM_OF_CHILD RES_CITY	RES_STATE RES_ZIP RES_COUNTRY WORK_CITY WORK_STATE WORK_ZIP	WORK_COUNTRY TITLE LEGACY WPI_SPS TRUSTEE	16
Education History	GPA GRAD_DISTINCTION PREF_CLAS BS_YEAR BS_MAJOR MS_YEAR MS_MAJOR PHD_YEAR	PHD_MAJOR CERT_YEAR CERT_MAJOR HONOR_YEAR HONOR_DEG NON_WPI_DEG WPI_AWD TAYLOR_AWD	SCHWIEGER_AWD GODDARD_AWD GROGAN_AWD BOYNTON_AWD WASHBURN_AWD HAD_SCHOLARSHIP STU_PROJECT_CTR ALU_PROJECT_CTR	24
Extracurricular Activities	ADM_VOL CLS_AGENT FRAT SPORT_COUNT VARSITY_SPRTS STU_CLUB	STU_ARTS STU_INTL_CLUB STU_CLUB_SPORT STU_PROF_SOC STU_MUSIC STU_CLS_OFF	STU_SCH_INVOLVE STU_SPEC_PROG STU_INTRAMURAL STU_HONR_SOC	17
Alumni Activities	REUNION ALUM_VOLUNTEER	ALUM_CLUB ALUM_LEADER		4
Giving Records	ALUM_CONTACTS FISCAL _YEAR_X	GIFT_CLUB_X PRES_FND	LIFETIME_PAC	41

Table 1.4 Completeness of Information for the Subgroups

Variable Group	"Student"		"Non-student"
	ALUM + GRAD + ALND	CERT + STDT	
Identifier	complete	complete	complete
Biographical Information	complete	complete	complete
Education History	complete	incomplete	none
Extracurricular Activities	complete	incomplete	none
Alumni Activities	complete	incomplete	none
Giving Records	complete	complete	complete

Overall, 29,455 (60.6%) of the constituents fall in the "best" subgroup of ALUM + GRAD + ALND, and thus makes a sufficiently large sample for analysis. For this reason, we decided to start the analysis with these three categories combined in the hope of getting the "best possible" model.

1.3 Statistical Methodologies/Models

A two-stage modeling approach was used in the analysis. For the first stage, the goal was to estimate the probability (likelihood) that a constituent is a contributor, and to assess the ability of this estimation in predicting constituents as either contributors or prospects. A logistic regression approach was chosen to model the relation between predictor variables and giving behavior. The goal of the second stage was to locate factors that have a statistically significant impact on the amount of contribution for the contributors. Note the response here has values on a continuous scale and thus a linear regression model was a natural choice.

After the models were built on the "best" subgroup, multiple imputation was done on the entire "student" group in an effort to deal with the missing values and also evaluate the stability of the imputation.

1.4 Software Package

The statistical computing package SAS® was used throughout this project. The choice was partially due to the extensive availability of documentation and technical support for the software in addition to its analysis capability and programming flexibility. The version of the package used was 9.1 TM Level 1M2 on Microsoft Windows XP professional platform.

Chapter 2

Data Preparation

2.1 Quality Control and Data Cleaning

Quality control of the data started with duplicated observation detection on the identifier variable and subsequent de-duplication if necessary. Extreme values and ranges of individual variables were examined to identify problematic cells. Natural associations among variables (columns) for individual observation (row) were then used as a reference for data cleaning [10]. A nice example is constituent with identifier 762250336. The value under "B.S. Year" appears to be 19 (which translates to 1919). But after printing out the entire row, we see the person was born in 1971 and obtained her bachelor's degree from MIT, so there should be an empty cell rather than 19. For the same person however, the value 95 under "M.S. Year" (which will be converted into 1995 later) can now be trusted with more confidence.

Variables in the file with dates containing years were presented in both two-digit and four-digit formats. For the purpose of new variable creation and recoding at a later phase, two-digit years were converted into four digits by identifying a cut-off value based on the variable's distribution.

2.2 Univariate Summarization

Univariate statistical analysis was conducted on each variable. Histograms and boxplots were constructed to display the distributions (location, spread, symmetry, etc.) of numeric variables and to perform a quick graphical check for outlier. Then descriptive statistics were calculated and examined. For categorical variables, frequency tables were obtained and checked.

Out of the 48,604 constituents, 24,204 (49.8%) turned out to be contributors. Table 2.1 gives a basic summary of the contribution amount for the whole population as well as the contributor group.

Table 2.1 Descriptive Statistics of Contribution Amount

	All constituents	Contributor Group
Counts	48,604	24,204
Minimum	$0.00	$0.02
Maximum	$5,979,538.69	$5,979,538.69
Mean	$2,044.85	$4,106.25
Standard Deviation	44,824.35	63,453.40
25 Percentile	$0.00	$50.00
Median	$0.00	$170.00
75 Percentile	$170.00	$695.00
Inter-Quartile Range	$170.00	$645.00
Total	$99,387,742.10	$99,387,742.10

Not that due to the skewness of the contribution amount's distribution, median and inter-quartile range (IQR) are more appropriate than mean and standard deviation here as measures of location and spread for the variable.

2.3 Modeling Universe Creation

2.3.1 Initial Variable Selection

Some of the 102 original variables were not included in the modeling universe for various reasons. 12 variables of the gift club designations from fiscal year 1996 to 2007 were dropped because the club entry standards changed over the years. "Preferred Class Year" was also excluded because of the huge overlap with "B.S. Year". The later variable was retained because it was believed to be more accurate and objective since preferred class year was picked by constituents themselves and thus bears fair amount of subjectivity. For the geographical location variables, "State" was chosen for its advantage of having standard abbreviations and fewer categories (which means easier cleaning and recoding and a much more consistent format compared with the "City" and "Zip Code" variables). Note here though that these dropped variables were still valuable references when new erratic cells were uncovered [10].

2.3.2 Response Variable Creation

The 25 variables carrying information of constituents' yearly contribution amount were used to create the response variables for the two models. Summing values across rows gave the total amount contributed by each constituent and in turn led to the definition of *contributor* as those with positive values. The remaining constituents were then designated to the *prospect* group.

2.3.3 Variable Recoding and Transformation

Many variables in the data take values of either "yes" or blank. For the purpose of maximizing the final model's predictive power in light of the limited number of candidate predictors available, we decided to keep as many variables as possible in this stage and thus coded them to indicators with "yes" as one and blank as zero. Care had to be taken when making interpretations about these indicators as zero here means no information available rather than simply "no".

The recoding produced 59 variables, all appended with suffix "_MOD" to distinguish them from their original versions. They include 28 binary indictors and 7 class variables (CATEGORY_MOD, GENDER_MOD, MARRIAGE_MOD, BSMAJOR_MOD, HOME_MOD, BIZSTATE_MOD and DISTINCTION_MOD). Table 2.2 gives the categorization detail for the "B.S. Major" as well as the two geographical region variables (which shared the same recoding scheme).

Table 2.2 Detail of "B.S. MAJOR" and "HOME/BUSINESS STATE"

Variable	Class	Contents
Home & Biz State	Mass	MA
	Rest_NewEng	CT, NH, RI, ME, VT
	Northeast	NY, NJ, PA, DE, MD, WV, DC
	West	CA, AK, AZ, CO, HI, ID, MT, NV, NM, OR, UT, WA, WY
	South	FL, AL, AR, GA, KY, LA, MS, NC, OK, SC, TN, TX, VA
	Midwest	IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD, WI
	Other	AE, AP, GU, PR, VI
	NA	QC, ZZ, ON, M, other, blank

B.S. Major **[1]**	MechanicalEngr	ME, MEA, MEB, MEN, MFE, MTE, IE, AE
	Elec./Comp.Engr	EE, ECE, EEB, EEC, EEN
	CivilEngr	CE, CEI
	ComputerSci	CS, CA, CSB, CSC, CSM
	ChemicalEngr	CM, CMB, CMN
	Chemistry	CH, CHI
	Physics	PH, PHE
	Math	MA, MAC
	BizEconomcs	MGE, BU, MG, MGC, MGS, MGT, MIS, EC, ET
	Bio./LifeSci	BBT, BBI, BC, BE, BIO, BM, BS, BB, LS, LSI
	HumanitiesArts	HT, HTE, HTH, HU, SS, SST, ST, TC, TW, IN
	OtherEngr	EP, EV, PL, FPE, NE
	Other	GS, ID, ND, SD
	NA	blank

[1] See "Appendix A" for the major codes.

Two original variables were recoded to enhance their interpretability: values of "B.S. Year" were subtracted from 2006 to produce a new "B.S. Recency" variable (which turned out later to have very strong predictive power for both models) and "Year of Birth" was translated into "Age" in a similar way.

Some new variables were created by consolidating original variables that deliver the same type of information and whose values are fairly sparse. Two approaches were used:

1) Taking maximum of indicators.

"M.S. Major" and "M.S. Year" are two original variables with information about the field of the master's program and the year the degree was awarded. They were first coded to binary indicators of value zero (if the original cell was blank) and one (if the original cell was not blank). These two new variables indicate the availability of such information in the data set. Secondly, a new binary variable indicating enrollment in WPI's master's program at some time point was created by taking maximum of the two aforementioned indicators. As a result, as long as one of the two original columns had something recorded, "MASTER_MOD" will be one. Only if both original columns were blank will it be zero. New variables created in the same fashion include: "PHD_MOD", "CERT_MOD", "HONOR_MOD" and "VIP_MOD" (based on "PRES_FND", "LIFETIME_PAC" and "TRUSTEE"), "INTL_MOD" (based on "RES_COUNTRY" and "WORK_COUNTRY"), "PROJECT_MOD" (based on "STU_PROJECT_CTR" and "ALU_PROJECT_CTR").

2) Summing up indicators/counts.

An example is the new variable "AWARD_MOD", which counts types of a certain set of awards the constituent received. The file comes with seven original variables corresponding to various types of awards ("WPI_AWD", "Taylor_AWD", "Schwieger_AWD", "Goddard_AWD", "Grogan_AWD", "Washburn_AWD" and "Boynton_AWD") with values of either "yes" or blank. Similarly, "yes" became one and blank became zero. "AWARD_MOD" was then constructed by summing the seven binary indicators. The new variable "ALUM_MOD" was created in the same way and counts the number of a set of alumni activities the constituent participated in.

Two variables, "Job Title" and "Non-WPI Degree" (the "messy" ones mentioned in section 1.2.3), were infeasible to categorize. In such cases, indicators of whether or not the constituent reported this information were created instead.

Transformations were done on some variables. The variable "Number of Children", highly skewed right with maximum value 12, has 4 as its 99th percentile. So it was regrouped into five categories of 0, 1, 2, 3, and 4 or more children.

After the recoding and transformation, "GPA", "Age" and "B.S. Recency" were the three variables left with large numbers of missing values. The 14,047 observations having non-missing values for all these three predictors were then flagged as the "complete" set out of the "best" subgroup of 29,455 alumni, graduate alumni and non-degree alumni and became the base for initial modeling.

2.3.4 Learning/Validation File Split

The modeling set was split into approximately equal-sized learning and validation files. In order to make the two sets more comparable, the split was conducted using stratified random sampling [6] with 20 equally-sized strata based on contribution amount. The choice of 20, rather than more commonly used 10 [16], was due to the fact that approximately half of the constituents made no contributions. Comparison of univariate statistics of the two files assured us they were similar with respect to the number of contributors and amount of contribution.

2.4 Variable Removal

The file splitting and subsetting up to this point rendered three variables no longer suitable for modeling. Indicators for legacy and honorary degree holder both became constants (all zero) and VIP Indicator had only one non-zero cell.

Chapter 3

Model Fitting

3.1 Logistic Regression Model

A logistic model is useful for modeling binary responses as a function of a set of predictors, and the fitted response can be used to estimate the probability (likelihood) of a certain event of interest [2]. For a logistic model with *n* predictors, the model equation is:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad (3.1)$$

in which *P* is the probability of the event of interest, β_0 is the intercept and β_i is the coefficient for the *i*th predictor X_i (i = 1 … n). Here, we can utilize this model to predict the tendency of giving for each constituent.

3.1.1 Initial Logistic Fit

Using the logistic procedure from SAS [3] with stepwise selection and variable entry and stay significance parameters both set at 0.05, an initial model was built on the complete records of the "best" subgroup (ALUM+GRAD+ALND). The resulting significant predictors, their *p*-values and the estimated signs for numeric predictors are shown in Table 3.1. The set is presented in descending order of statistical significance.

Table 3.1 Initial Logistic Fit Result

Predictor	Estimated Sign	*p*-value
Years since B.S. awarded	+	<.0001
Biz geographical region	Class variable	<.0001
Alumni activities count	+	<.0001
Number of children	+	<.0001
School activities indicator	+	<.0001
Home geographical region	Class variable	<.0001
GPA	+	<.0001
Reunion indicator	+	<.0001
Gender	Class variable	<.0001
Indicator, non-WPI degree reported	+	<.0001
WPI spouse indicator	+	0.0011
Honor society count	+	0.0041
International club activities count	-	0.0044
Professional society count	+	0.0136
Area of B.S. major	Class variable	0.0145
Awards Count	-	0.0316
Age	-	0.0327

3.1.2 Reality Check

Some of the signs for the parameter estimates in Table 3.1 seem counterintuitive. For example, the model has a negative sign for "Awards Count", which counts the types of award the constituent has received. One would think that award recipients should be more, not less, likely to give back to the school. To investigate the consistency of the estimated coefficient signs with the data, we performed "reality checks" by looking more closely at the data. For numeric variables like indicators and counts whose values are on a discrete scale, a simple cross tabulation will help reveal what the estimated sign should be. This is illustrated in Tables 3.2 and 3.3. We can easily tell that both variables should end up with positive signs.

Table 3.2 Reunion Indicator Cross-Tab

Reunion Indicator	Contributor	
	No	Yes
0	7618 **58.48%**	5409 **41.52%**
1	249 **24.41%**	771 **75.59%**

Table 3.3 Award Counts Cross-Tab

Award Count	Contributor	
	No	Yes
0	7846 **56.09%**	6141 **43.91%**
1	21 **35.59%**	38 **64.41%**
2	0 **0.00%**	1 **100.00%**

For numeric variables with values on a continuous scale, a side-by-side box plots grouped by contributor/prospect can accomplish the same task. Two examples are given below in Figures 3.1 and 3.2 regarding the "Age" and "B.S. Recency" variables.

Figure 3.1 Side-by-side Boxplot for "Age"

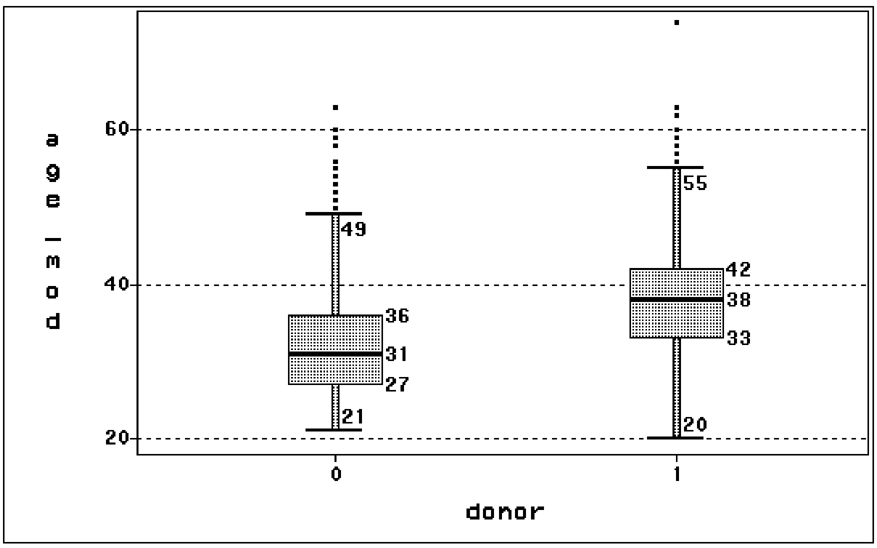
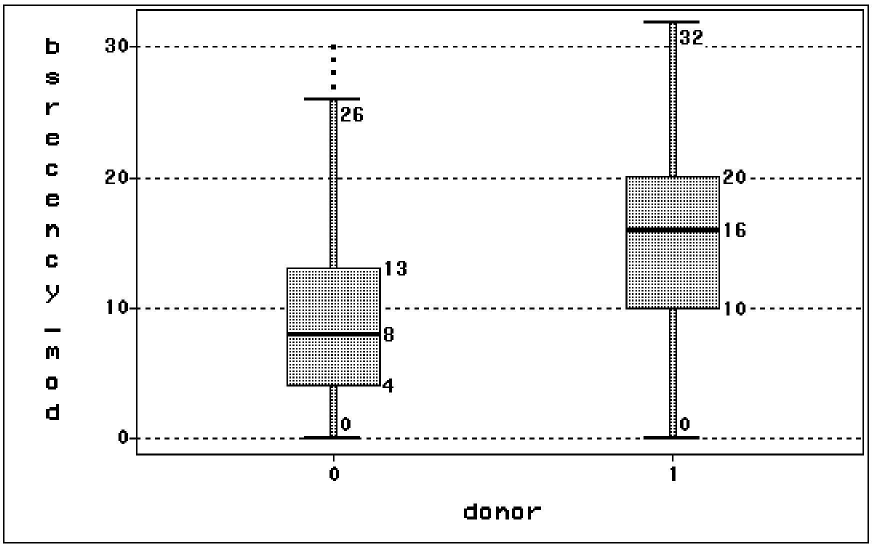


Figure 3.2 Side-by-side Boxplot for "B.S. Recency"

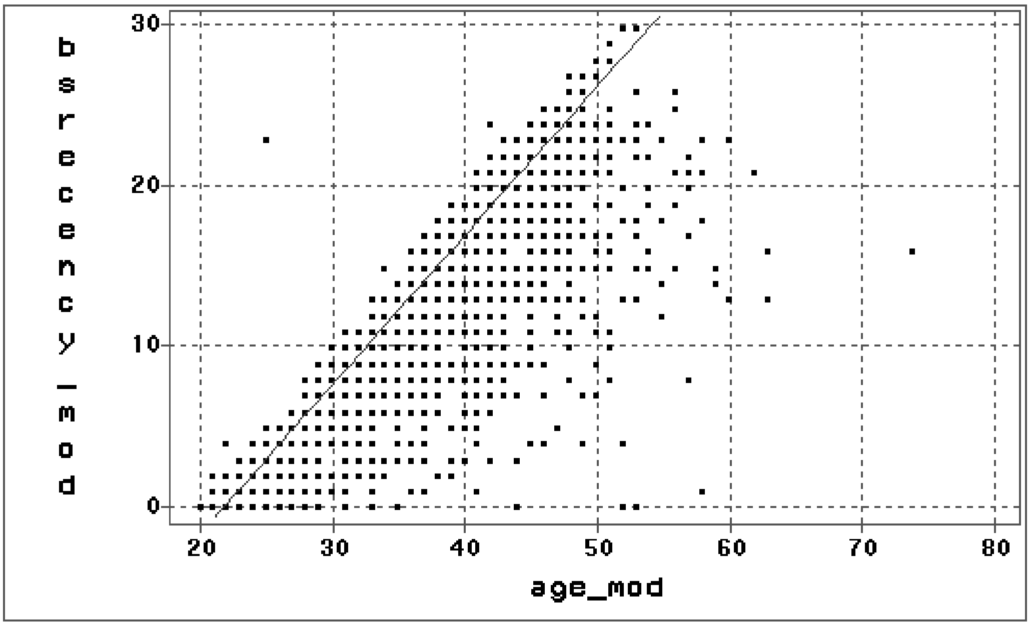


The two plots reveal that constituents graduated (with B.S. degree) earlier, thus of older age, are more likely to give. So we would conclude that the estimated sign for the age variable in the initial fit did not correspond to the marginal relation of the variable with the response. This is possibly caused by the existence of collinearity, because two highly correlated variables bring in redundant information, and compensation for the presence of the other might lead to a reversal of signs in their coefficient estimates [1].

3.1.3 Collinearity

Scatterplot matrices and correlation matrices constructed for the identified set of predictors were helpful in graphically displaying the existence of pairwise collinearity [1]. A simple scatterplot of "Age" and "B.S. Recency" along with a fitted linear regression line is shown in Figure 3.3.

Figure 3.3 Scatterplot of "Age" and "B.S. Recency"



A first glimpse might mask the true strong linear association. But the Pearson correlation is 0.9583, very high since the great majority of data points lie close to the fitted line which corresponds to the following equation:

$$B.S.\ Recency = -20.3408 + 0.9288 * Age$$

The estimated intercept and slope show an interesting fact that "B.S. Recency" is basically "Age" shifted 20 years.

3.1.4 Model Selection and Validation

The reality check and collinearity detection led to the idea of trying models with or without the "Age" and "Award Counts" variables. Also, "Home Region" and "Working Region" both stayed in the initial model, but values for these two could possibly overlap for many observations. A quick comparison showed a match rate of 52.86%. So over half of the pairs share the same values and it was then worth trying model fits with one of them excluded.

Table 3.4 gives the validation results of models with different candidate pools. Three measures were shown for comparison:

1) *Contributor prediction rate.* This is the percentage of contributors in the validation sample who have been correctly identified by the model as contributors.
2) *Prospect prediction rate.* Similarly, this is the percentage of prospects in the validation sample who have been correctly identified by the model as prospects.
3) *Prediction match rate.* This is the percentage of constituents in the validation sample who were correctly classified by the model.

Table 3.4 Performance of Logistic Models

Model No.	**1**	**2**	**3**	**4**	**5**
Model Detail	Initial Model	No Age	No Age & Award	No Age, Award, Home	No Age, Award, Biz
Contributor Pred. Rate	61.70%	60.11%	60.24%	59.47%	**61.80%**
Prospect Pred. Rate	79.64%	80.95%	**81.05%**	80.80%	79.76%
Prediction Match Rate	71.72%	71.74%	**71.86%**	71.37%	71.83%

We observe that all the five models are better at identifying prospects than contributors and the performances of the models have no considerable differences. For the purpose of identifying contributors, model 5 seems to outperform the others. If we want to identify prospects or achieve the highest overall classification accuracy instead, model 3 will produce the most desirable result.

The sets of significant predictors for models 3 and 5 along with their *p*-values and point estimates obtained using the maximum likelihood method are shown in Table B.1 and B.3 of Appendix B. The predictors are presented in descending order of statistical significance. Given the inputs, applying the model will give each constituent a predicted response, which is an estimate of the probability of giving (also known as "score" [16]).

An excerpt of the fitting and statistical details for model 3 can be found in Appendix C.

3.1.5 Odds and Odds Ratio

For a logistic model, in many cases the odds ratio is also of interest.

The *odds* of an event are calculated by dividing the probability of an event (P) by the probability of its complement, as $P/(1-P)$ [2]. For instance, if the probability a constituent is a contributor is 0.51, then the odds a constituent is a contributor are 0.51/0.49 = 1.04. An odds greater than one implies that the event is more likely to happen than not (the odds of an event that is certain to happen are infinite); if the odds are less than one the event is less likely to happen than not (the odds of an impossible event are zero). An event equally likely to happen or not has odds one.

An *odds ratio* is the ratio of the odds of one event to the odds of another event and is used to compare the odds of the two. In a logistic model, odds ratios are used to assess the effect of a predictor on the odds of the event being modeled (here the event a constituent is a contributor). Specifically, the coefficient of a numeric predictor is the proportional change in the odds for any one unit increase in that predictor. An odds ratio greater than one means that the event is more likely to happen when the predictor goes up one unit, given all other predictors remain unchanged [2].

In the logistic model equation (3.1), *P* is a function of $X_1,...,X_n$ and thus the

$$\log\left(\frac{P}{1-P}\right)=\beta_0+\sum_{i=1}^{n}\beta_i X_i \tag{3.1}$$

values of the odds $\frac{P}{1-P}$, denoted by $O(X_1,...,X_n)$, is also determined by levels of the predictors. The log odds of the event for a set of given predictor levels $x_1,...,x_n$, written as $\log[O(x_1,...,x_n)]$ is just

$$\log[O(x_1,...,x_n)]=\beta_0+\sum_{i=1}^{n}\beta_i x_i \tag{3.2}$$

Suppose the *j*th predictor has a one unit increase in its level (from x_j to x_j +1), then the log odds will correspondingly change to

$$\log[O(x_1,...,x_j+1,...,x_n)]=\beta_0+\sum_{i=1}^{n}\beta_i x_i+\beta_j \tag{3.3}$$

Subtracting (3.2) from (3.3) gives the difference between the two log odds

$$\log[O(x_1,...,x_j+1,...,x_n)] - \log[O(x_1,...,x_j,...,x_n)]=\beta_j \tag{3.4}$$

and this equals

$$\log\left(\frac{O(x_1,...,x_j+1,...,x_n)}{O(x_1,...,x_j,...,x_n)}\right)=\beta_j \tag{3.5}$$

which tells us the ratio between these two odds is

$$\frac{O(x_1,...,x_j+1,...,x_n)}{O(x_1,...,x_j,...,x_n)}=e^{\beta_j} \tag{3.6}$$

and this is just the odds ratio for the *j*th predictor.

For a categorical (class) predictor, its odds ratio is just the proportional change of the odds if the predictor changes from the baseline category (chosen in recoding) to the current category [2]. Appendix C gives details about the categorical variable recoding for model 3.

Table B.2 and B.4 of Appendix B show both point and interval estimates of the odds ratios for the significant numeric variables identified in model 3 and 5.

3.2 Linear Regression Model

A linear regression model is appropriate for modeling responses of continuous numeric type with one of the underlying assumptions being that the response comes from a normal distribution [1]. For a linear regression model with *n* predictors, the model equation is:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \tag{3.7}$$

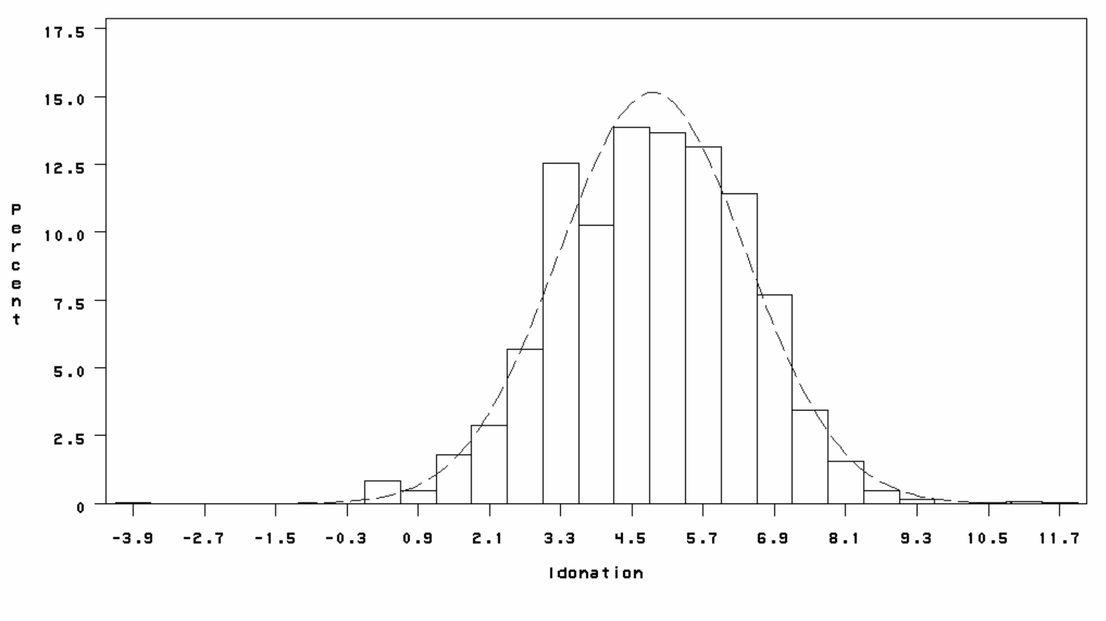
in which *Y* is the observed response, β_0 is the intercept, β_i is the coefficient for the *i*th predictor X_i (*i* = 1 … *n*) and ε is the random error term independently and identically distributed as $N(0,\sigma^2)$. Here, we will utilize this method to predict the amount of contribution for each of the known contributors.

3.2.1 Box-Cox Transformation

The response was highly skewed, so we chose a Box-Cox transformation [1] (See Appendix D for more information), which turned out to be a natural log.

Figure 3.4 shows a histogram of the transformed response with a fitted normal curve.

Figure 3.4 Histogram of the Transformed Contribution Amount



3.2.2 Model Fitting and Validation

Several linear regression models with slightly different groups of candidate predictors and significance levels for stepwise variable selection were tried and the two models in Table 3.5 ended up being the best two. As with the logistic fit, performance on the validation file was used as the criterion for comparison. The validation was done by first applying the respective model equation to the validation file, followed by grouping those constituents (in the validation file) into ten deciles based on their predicted giving amount. Percentages of the total real contribution amount for each decile were then calculated. The results are shown in Table 3.5.

Table 3.5 Performance of Linear Models

	Model1 (SLE=.01, SLS=.01)		**Model2** (SLE=.05, SLS=.01)	
Decile	**Amount**	**Percentage**	**Amount**	**Percentage**
1st	$443,515.89	**32.16%**	$433,850.53	31.46%
2nd	$224,542.17	**16.28%**	$225,325.17	16.34%
3rd	$121,517.23	8.81%	$122,212.23	8.86%
Top 20%	$668,058.06	**48.44%**	$659,175.70	47.80%
Top 30%	$789,575.29	57.25%	$781,387.93	56.66%

In an imaginary case where the constituents are randomly sliced into deciles, each decile is expected to account for roughly 10% of the contributions. But here, we see that the model-identified top 20% give almost half of the contribution amounts within the validation file. A direct marketing professional would thus recognize the model with over 300% lift [16] on the first decile and over 160% lift on the second one. Results between the models showed that model 1 performed better although the difference is relatively small.

The linear model based on the "complete" observations from the "student" contributors yields the following set of significant predictors, sorted in descending order of the magnitudes of their standardized coefficient estimates.

Table 3.6 Linear Model Results

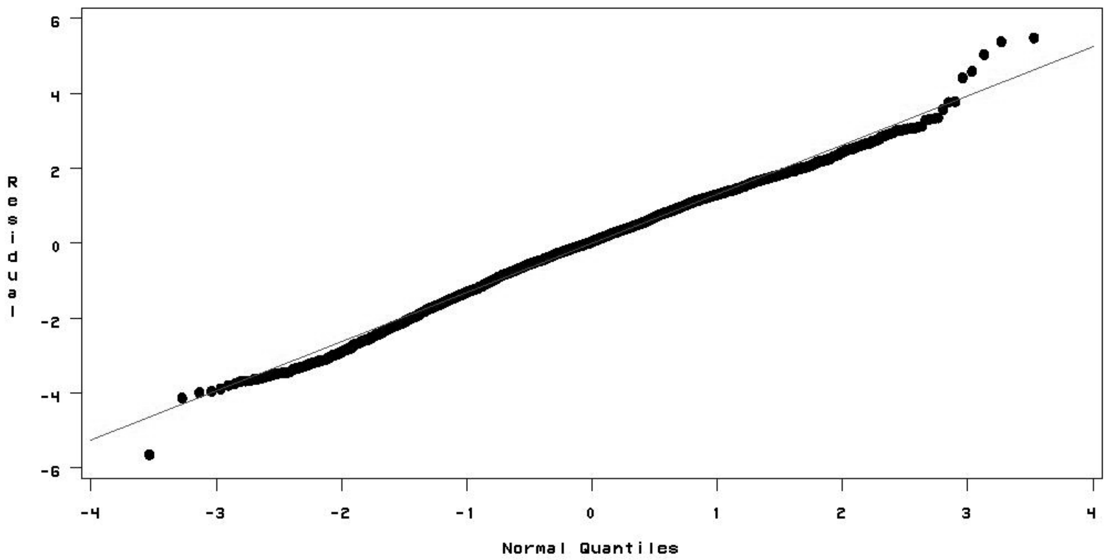
Predictor	**Coefficient Estimate**	**Standardized Estimate**
Years since B.S. awarded	0.10327	0.43306
Alumni activities count	0.30861	0.16025
Reunion indicator	0.47378	0.10083
GPA	0.43371	0.08514

School activities indicator		0.08125	0.05764
WPI spouse indicator		0.27620	0.05708
Count of intramural sports		0.07689	0.05672
Count of varsity sports		-0.10287	-0.04217
Contacts made as an alumnus		1.81278	0.04169
PhD Indicator		-1.75430	-0.04034
Biz Geographical Region	Mass	0.00049236	0.00024552
	Rest_NewEng	-0.03521	-0.01409
	Midwest	0.16820	0.05330
	Northeast	-0.03291	-0.01179
	South	0.10224	0.03518
	West	0.09509	0.03219
	Other	-0.06383	-0.01829

3.2.3 Model Diagnostics

Although predictive capability was the principal feature of interest in these models, residual plots were evaluated to check the usual assumptions of normality and homoscedasticity and appropriateness of fit [1]. The normal probability plot is given in Figure 3.5 as an example. No substantial deviations from these assumptions were detected.

Figure 3.5 Normal Probability Plot of the Residuals



3.3 Multiple Imputation for Missing Values

Missing values are an issue in a substantial number of statistical analyses. While analyzing only complete observations has its simplicity, the information contained in the incomplete ones is lost. Sometimes there are also systematic differences between the complete set and the incomplete set and this can make the resulting inference inapplicable to the population of all these observations, especially when the size of the complete set is relatively small.

For our case, the highest missing rate happened on the variable "GPA" (38.14%) followed by "B.S. Recency" (18.91%). So the size of the complete set is relatively large. Checking the data further we found out the categories of graduate and non-degree alumni have the "B.S. Rencency" cells all blank which is to be expected. Excluding these two categories reduced the missing rate to 0.90% for the single category of ALUM. This situation signals us it is not appropriate to impute values for all the three categories combined since it violates the important assumption of "missing at random" for imputation. So we decided to do the imputation by individual category.

The MI procedure from SAS is capable of creating multiply imputed data sets for incomplete data. It uses methods that incorporate appropriate variability across the imputations. Available methods include a parametric method (with multivariate normality assumption) like regression, a nonparametric method like propensity score and a Markov Chain Monte Carlo (MCMC) method [15].

Five imputations were run on the "student" group using the MCMC method. The multiply imputed data sets were then subjected to the same procedures for model selection, fit, and analysis used for the complete data. The five logistic models all produced the same set of 24 significant predictors with merely order of entering the model differing slightly. Table 3.7 lists the coefficient estimates from these five analyses with the predictors identified on the "complete" set bolded. We see the set includes all 17 variables from the model fitted on the "complete" fraction and the estimated values for the coefficients are fairly close across the models. This ensures us the stability and reliability of this imputation process.

Table 3.7 Modeling Results after Multiple Imputation

Predictor	**Coefficient Estimates for 5 Models**				
	1	**2**	**3**	**4**	**5**
Class agent	1.5638	1.5643	1.5630	1.5634	1.5642
Alumni activity indicator	0.6589	0.6591	0.6592	0.6594	0.6592
GPA	0.0608	0.0601	0.0613	0.0607	0.0600
B.S. Recency	0.0659	0.0658	0.0660	0.0659	0.0658
Non-WPI Degree	0.3153	0.3157	0.3154	0.3154	0.3155

Spouse Indicator		0.1783	0.1778	0.1784	0.1782	0.1780
Number of children		0.1503	0.1505	0.1503	0.1504	0.1505
Scholarship indicator		0.0925	0.0920	0.0928	0.0925	0.0921
Reunion indicator		0.7308	0.7311	0.7307	0.7308	0.7311
Greek house indicator		0.1322	0.1325	0.1322	0.1323	0.1325
Varsity sports		-0.1931	-0.1932	-0.1932	-0.1932	-0.1932
International Club		-0.2593	-0.2596	-0.2593	-0.2595	-0.2597
Club sport		0.0650	0.0651	0.0650	0.0651	0.0651
Professional Society		0.1531	0.1534	0.1530	0.1532	0.1535
Music indicator		0.1369	0.1370	0.1369	0.1369	0.1370
School Involvement		0.1963	0.1962	0.1962	0.1962	0.1962
Honor Society		0.1631	0.1628	0.1633	0.1631	0.1628
Project Center		0.1491	0.1487	0.1493	0.1490	0.1486
Marital Status	Divorced	0.2809	0.2827	0.2819	0.2820	0.2815
	Married	0.2949	0.2965	0.2961	0.2954	0.2956
	NA	-0.5442	-0.5536	-0.5507	-0.5475	-0.5480
	Other/Partner	0.1674	0.1689	0.1684	0.1678	0.1680
	Separated	-0.2425	-0.2420	-0.2410	-0.2425	-0.2430
	Single	-0.0785	-0.0775	-0.0773	-0.0781	-0.0783
B.S. Major	Biological/LifeSci	0.0904	0.0872	0.0868	0.0869	0.0844
	BizEconomcs	0.1913	0.1880	0.1875	0.1876	0.1853
	ChemicalEngr	0.1386	0.1357	0.1346	0.1351	0.1330
	Chemistry	-0.1025	-0.1051	-0.1065	-0.1059	-0.1078
	CivilEngr	0.3118	0.3089	0.3079	0.3083	0.3062
	ComputerSci	0.3307	0.3276	0.3269	0.3272	0.3249
	Electr./Comp.Engr	0.3361	0.3334	0.3320	0.3326	0.3307
	HumanitiesArts	0.3840	0.3808	0.3803	0.3804	0.3780
	Math	0.0872	0.0845	0.0832	0.0837	0.0818
	MechanicalEngr	0.2708	0.2678	0.2667	0.2672	0.2651

B.S. Major	NA	-2.4217	-2.3838	-2.3708	-2.3759	-2.3483
	Other	0.3389	0.3361	0.3349	0.3354	0.3334
	OtherEngr	0.00773	0.00460	0.00382	0.00413	0.00189
Biz region	Mass	0.0275	0.0273	0.0272	0.0274	0.0274
	Midwest	0.0546	0.0558	0.0558	0.0554	0.0553
	NA	-0.5169	-0.5170	-0.5171	-0.5170	-0.5170
	Northeast	0.0370	0.0371	0.0371	0.0371	0.0372
	Other	0.2301	0.2298	0.2299	0.2299	0.2299
	Rest_NewEng	0.1034	0.1031	0.1031	0.1031	0.1032
	South	0.1563	0.1562	0.1562	0.1563	0.1562
Home region	Mass	0.1693	0.1692	0.1696	0.1693	0.1691
	Midwest	0.2688	0.2675	0.2678	0.2681	0.2678
	NA	-0.9271	-0.9266	-0.9273	-0.9268	-0.9262
	Northeast	0.2461	0.2463	0.2463	0.2462	0.2462
	Other	0.2081	0.2082	0.2083	0.2082	0.2082
	Rest_NewEng	0.0489	0.0491	0.0493	0.0490	0.0488
	South	-0.0653	-0.0653	-0.0652	-0.0653	-0.0653
Gender	F	-1.8857	-1.9988	-1.6320	-1.9068	-2.1126
	M	-2.0699	-2.1827	-1.8159	-2.0907	-2.2964
	N	1.9281	2.5259	2.1341	2.2005	2.1573
Distinction	D	0.000270	0.000200	0.000351	0.000278	0.000187
	H	0.1050	0.1051	0.1049	0.1050	0.1051

Chapter 4

Conclusions

4.1 Summary

The logistic models discovered sets of variables bearing statistically significant impacts on the likelihood of giving for constituents in the student group. It also enabled us to assign a score [16] (i.e. predicted value for the response) to current and future individuals in the group so that efforts can be focused on the higher-scored fraction. To score the constituents with "complete" records inside the "student" group, the models built upon these observations shall be used. If scoring the remaining individuals is also desired, the average predicted value from models built after multiple imputation can be an option. But overall, the "complete" models are the ones to deliver and recommend for scoring future "student" constituents as we expect the incoming observations will all have complete information as a result of improved record keeping. The specific choice of model depends on what is to be achieved in a campaign and the performance of respective models.

The linear model gave a set of variables having statistical significance in driving the magnitude of giving for contributors. The relative importance of the predictors can be decided by comparing the absolute values of the standardized parameter coefficients (shown in Table 3.6). The larger they are, the higher contribution amount can be expected to receive for an increase of one standard deviation (which is comparable across the predictors after the standardization) in the predictor.

Comparing the sets of identified significant predictors from both models, there are seven common ones. So, regardless of the objective, whether to predict the possibility or the amount of giving, those who graduated earlier, work in particular geographical areas, participated in alumni activities and reunion activities in the past and had better academic performance and involved in school activities when attending WPI, and whose spouse is also a constituent are more likely to give and to give larger amounts on average.

4.2 Future Work

The modeling so far primarily focused on the "student" group. Profiles of the rest of the constituent categories (parents, neighbors, friends, etc.) can also be investigated to see whether with lesser amount of information, an effective predictive model can still be obtained.

Major contributors flagged by the VIP indicator (generated by consolidating "PRES_FND", "LIFETIME_PAC" and "TRUSTEE") were excluded in the modeling base. Although a fairly small group, they tend to account for a large portion of the total gifts and display distinctive behaviors, which makes examination of the group worthwhile.

Other approaches to analysis, such as classification and neural network methods, might be appropriate for analyzing this data set and could reveal other interesting findings as well.

Appendix A: Table of Major Codes

Table A.1 WPI Major Codes

Code	Description	Dept
AE	Aerospace Engineering	ME
AL	American Literature	HU
AM	Applied Mathematics	MA
AS	American Studies	ND
ASC	Assumption College	ND
ASD	Actuarial Science	ND
B1	Cellular and Molecular Biology	BB
B2	Biomaterials	BE
BB	Biology/Biotechnology	BB
BBI	Biology	BB
BBT	Biotechnology	BB
BC	Biochemistry	CH
BE	Biomedical Engineering	BE
BIO	Biology and Biotechnology	BB
BIOC	Computational Biology	BB
BIOE	Ecology & Environmental Bio	BB
BIOG	Cell & Molecular Bio/Genetics	BB
BIOM	Biomedical Interests	BE
BIOO	Organismal Biology	BB
BIOP	Bioprocess	BB
BIS	Biological Information Systems	BB
BM	Biomedical	BE
BMP	Biomedical Eng/Medical Physics	BE
BS	Biomedical Sciences	BB
BSMB	BS/MBA PROGRAM	ND
BSMS	BS/MS PROGRAM	ND
BU	Business	ND
BUSA	Business Administration	ND
CA	Computers with Applications	CS
CC	Customized Certificate	ND
CCN	Computers & Comm. Networks	ND
CE	Civil Engineering	CE
CEEV	Environmental	CE
CEI	Civil Engineering-Interdiscipl	CE

CET	Civil Engineering-Traffic	CE
CH	Chemistry	CH
CHB	Chemistry:Bio-organic Emphasis	CH
CHI	Chemistry-Interdisciplinary	CH
CHMC	Medicinal Chemistry	CH
CL	Clinical Engineering	BE
CM	Chemical Engineering	CM
CMB	Chem. Eng w/Biomedical Int.	CM
CMBC	Biochemical	CM
CMBM	Biomedical	CM
CMEV	Environmental	CM
CMMT	Materials	CM
CMN	Chem. Engr. w/Nuclear Int.	CM
CNE	Central New England College	ND
COMM	Commerce	ND
CPM	Construction Project Mgmt.	CE
CS	Computer Science	CS
CSB	Computer Sci w/Biomedical Int.	CS
CSC	Computers w/Commercial Appl.	CS
CSM	Computers w/Mathematical Appl.	CS
CV	Client / Server	DCS
DE	Differential Equations	MA
DENT	Dentistry	ND
DT	Drama/Theatre	HU
EC	Economics	SST
ECE	Electrical & Computer Eng.	EE
ECO	Ecology	BB
ED	Engineering - To Be Declared	ND
EE	Electrical Engineering	EE
EEB	Elect. Eng w/Biomedical Int.	EE
EEC	Elec. Eng. w/Comp. Eng. Spec.	EE
EECO	Computer Engineering	EE
EEN	Elec Engr w/ Nuclear Int	EE
EIT	Engineer in Training	ND
EL	English Literature	HU
EM	E-Commerce	DCS
EN	English	HU
EP	Environmental Policy & Develop	SST
ER	Entrepreneurship	MG
ES	Environmental Studies	ND
ET	Economics & Technology	SST

EV	Environmental Engineering	ID
EVS	Environmental Science	ND
FORS	Forestry	ND
FPE	Fire Protection Engineering	FPE
FPIN	Fire Protection Interests	FPE
FR	French	HU
GD	Geometric Dimens & Tolerance	DCS
GH	Global History	HU
GN	German	HU
GS	General Science (OldTimer)	ND
GWEP	Greater Worc Exec Prog	ND
HCC	Holy Cross College (32)	ND
HI	History	ND
HS	Hispanic Studies	HU
HT	Humanities Studies/Sci & Tech	HU
HTE	Humanities/Technology-English	HU
HTH	Humanities/Technology-History	HU
HTT	Humanities/Technology	HU
HU	Humanities and Arts	HU
HUAH	Art History	HU
HUAS	American Studies	HU
HUCW	Creative Writing	HU
HUDT	Drama/Theatre	HU
HUEV	Environmental Studies	HU
HUGN	German Studies	HU
HUHI	History	HU
HUHS	Hispanic Studies	HU
HULI	Literature	HU
HUMU	Music	HU
HUPY	Philosophy	HU
HURE	Religion	HU
HUST	HU Studies of Science & Tech	HU
HUWR	Writing and Rhetoric	HU
ID	Interdisciplinary	ID
IDM	Individually-Designed Minor	ND
IE	Industrial Engineering	MG
IME	Impact Engineering	ID
IMGD	Interactive Media & Game Dev	ID
IN	International Studies	ID
IS	Intersession	ND
ISCH	International Scholar	ND

ISCP	International Scholar Program	ND
ISM	Information Security - Mgmt	ND
IST	Information Security - Technic	DCS
IT	Information Technology	MG
LIT	Literature	HU
LS	Life Sciences	ND
LSI	Life Sciences-Interdisciplin	ND
LT	Law and Technology	ID
MA	Mathematical Sciences	MA
MAC	Actuarial Mathematics	MA
MAF	Financial Mathematics	MA
MAI	Industrial Mathematics	MA
MAS	Applied Statistics	MA
MAT	Mathematics	MA
MBA	Master of Business Admin.	MG
ME	Mechanical Engineering	ME
MEA	Mech. Eng. w/ Aerospace Int.	ME
MEAE	Aerospace	ME
MEB	Mech. Eng. w/ Biomedical Int.	ME
MEBM	Biomedical	ME
MEEM	Engineering Mechanics	ME
MEEV	Environmental	ME
MEMB	Biomechanical	ME
MEMD	Mechanical Design	ME
MEMF	Manufacturing	ME
MEMS	Materials Science	ME
MEN	Mech. Eng. w/ Nuclear Int.	ME
MENE	Nuclear	ME
METF	Thermal-Fluids	ME
MF	Manufacturing Systems Eng.	ME
MFA	Advanced Manufacturing Eng.	ME
MFE	Manufacturing Engineering	ME
MFM	Manufacturing Management	MG
MFS	Manufacturing Eng Mgmt	ID
MG	Management	MG
MGC	Management with Computer Appl.	MG
MGE	Management Engineering	MG
MGS	Management Science & Engr.	MG
MGT	Management	MG
MH	Mathematics	MA
MHS	Statistics	MA

MIS	Management Information Systems	MG
MM	Master of Mathematics	MA
MME	Master of Mathematics for Educ	MA
MN	Management Development	DCS
MNS	Master of Natural Sciences	BB
MPE	Materials Processing Eng	ME
MSM	Master of Science in Mgmt.	MG
MT	Management of Technology	MG
MTE	Materials Science and Eng.	ME
MTI	Marketing & Tech. Innovation	MG
MTL	Materials	ME
MU	Music	HU
MUSC	Music	HU
N1	Nanoscience	CM
NC	Non-Certificate (DCS/CPE)	DCS
ND	To Be Declared	ND
NE	Nuclear Engineering	ME
NURS	Nursing	ND
ODL	Operations Design & Leadership	MG
OIT	Operations & Information Tech.	MG
OL	Organizational Leadership	MG
OT	Special Topics	DCS
PDEN	Pre-Dental	ND
PH	Physics	PH
PHE	Engineering Physics	PH
PHL	Philosophy	HU
PHL1	Philosophy of Social Problems	HU
PHRM	Pharmacy	ND
PI	Process Improvement	DCS
PL	Urban & Environmental Planning	CE
PLE	Plant Eng. Certificate	ND
PM	Pre-Med	ND
PMED	Pre-Medical	ND
PO	Political Science & Law	SST
PR	Project Management	DCS
PS	Psychology	SST
PSM	Power Systems Management	ID
PSS	Psychological Science	SS
PVET	Pre-Veterinary	ND
PW	Professional Writing	HU
QI	Quality Improvement	DCS

RE	Religion	HU
RH	Rhetoric	HU
SC	Science (Freshmen Only)	ND
SD	System Dynamics	SST
SE	Structural Engineering	CE
SIM	School of Industrial Management	MG
SM	Systems Modeling	ID
SO	Sociology	SST
SP	Spanish	HU
SS	Social Science	SST
SST	Social Science & Technology	SST
ST	Society, Technology & Policy	SST
STA	Statistics	MA
TC	Tech, Sci & Prof Communication	ID
TEAC	Teaching	ND
TM	Technology Marketing	MG
TW	Technical Writing	ID
URB	Urban Planning	ND
URBN	Urban Studies	ND
WC	World Class Manufacturing	DCS
WD	Windows 2000	DCS
WH	World History	HU
WR	Writing and Rhetoric	HU
WT	Web Technologies	DCS

Appendix B: Logistic Modeling Results

Table B.1 Logistic Fit Results for Model 3

Predictor		Estimate	Standard Error	*p*-value
Years since B.S. awarded		0.0934	0.00502	<.0001
Biz geographical region	Mass	0.0900	0.1080	<.0001
	Midwest	0.00161	0.2098	
	NA	-0.4991	0.0988	
	Northeast	0.1569	0.1499	
	Other	-0.2025	0.5470	
	Rest_NewEng	0.1265	0.1257	
	South	0.2800	0.1580	
Alumni activities count		0.5564	0.0752	<.0001
Number of children		0.2573	0.0417	<.0001
School activities indicator		0.2142	0.0413	<.0001
Home geographical region	Mass	0.2199	0.0874	<.0001
	Midwest	-0.0531	0.1745	
	NA	-0.6728	0.1280	
	Northeast	0.2502	0.1236	
	Other	0.3195	0.4169	
	Rest_NewEng	0.0472	0.1002	
	South	-0.0598	0.1252	
GPA		0.6524	0.1231	<.0001
Reunion indicator		0.6021	0.1207	<.0001
Gender	F	2.9413	55.3900	<.0001
	M	2.6106	55.3900	

Indicator, non-WPI degree reported		0.3449	0.0789	<.0001
WPI spouse indicator		0.3413	0.1057	0.0011
International club activities count		-0.2896	0.0918	0.0044
Honor society count		0.1960	0.0792	0.0041
Professional society count		0.1626	0.0586	0.0136
Area of B.S. major	Biological/LifeSci	-0.0489	0.1158	0.0145
	BizEconomcs	-0.1438	0.1251	
	ChemicalEngr	-0.1830	0.1259	
	Chemistry	-0.2228	0.2456	
	CivilEngr	-0.0230	0.1068	
	ComputerSci	0.2337	0.1080	
	Electrical/ComputerEngr	0.2372	0.0894	
	HumanitiesArts	0.2935	0.2593	
	Math	0.0697	0.1843	
	MechanicalEngr	0.1094	0.0854	
	Other	-0.0625	0.5205	
	OtherEngr	0.0785	0.4296	
Greek house indicator		0.1412	0.0650	0.0354
Graduate with distinction	D	0.0162	0.0454	0.0457
	H	-0.1450	0.0670	

Table B.2 Odds Ratio Estimates for Model 3

Predictor	Point Estimate	95% Confidence Interval	
Alumni activities count	1.744	1.505	2.022
GPA	1.920	1.509	2.444
Years since B.S. awarded	1.098	1.087	1.109
Indicator, non-WPI degree reported	1.412	1.210	1.648

WPI spouse indicator				1.407	1.144	1.730
Number of children				1.293	1.192	1.404
Reunion indicator				1.826	1.441	2.313
Greek house indicator				1.152	1.014	1.308
International club activities count				0.749	0.625	0.896
Professional society count				1.177	1.049	1.320
School activities indicator				1.239	1.142	1.343
Honor society count				1.216	1.042	1.421
Gender	F	vs	N	>999.999	<0.001	>999.999
	M	vs		>999.999	<0.001	>999.999
Area of B.S. major	Biological/LifeSci	vs	Physics	1.335	0.801	2.225
	BizEconomcs	vs		1.214	0.721	2.044
	ChemicalEngr	vs		1.168	0.694	1.964
	Chemistry	vs		1.122	0.564	2.231
	CivilEngr	vs		1.370	0.830	2.262
	ComputerSci	vs		1.771	1.073	2.925
	Electr./Comp.Engr	vs		1.777	1.095	2.885
	HumanitiesArts	vs		1.880	0.924	3.828
	Math	vs		1.503	0.830	2.723
	MechanicalEngr	vs		1.564	0.967	2.531
	Other	vs		1.317	0.399	4.352
	OtherEngr	vs		1.517	0.549	4.190
Biz geographical region	Mass	vs	West	1.044	0.732	1.490
	Midwest	vs		0.956	0.562	1.627
	NA	vs		0.579	0.413	0.813
	Northeast	vs		1.117	0.728	1.713
	Other	vs		0.780	0.221	2.749
	Rest_NewEng	vs		1.083	0.737	1.591
	South	vs		1.263	0.813	1.963

Home geographical region	Mass	vs	West	1.311	0.981	1.752
	Midwest	vs		0.998	0.642	1.552
	NA	vs		0.537	0.375	0.769
	Northeast	vs		1.352	0.950	1.922
	Other	vs		1.449	0.553	3.792
	Rest_NewEng	vs		1.103	0.809	1.505
	South	vs		0.991	0.696	1.413
Graduate with distinction	D	vs	NA	0.894	0.774	1.031
	H	vs		0.760	0.610	0.947

Table B.3 Logistic Fit Results for Model 5

Predictor		**Estimate**	**Standard Error**	**P-value**
Years since B.S. awarded		0.0929	0.00512	<.0001
Alumni activities count		0.5599	0.0754	<.0001
Indicator, job title reported		0.4325	0.0587	<.0001
Home geographical region	Mass	0.2354	0.0776	<.0001
	Midwest	-0.1079	0.1424	
	NA	-0.5634	0.1433	
	Northeast	0.2786	0.1047	
	Other	0.1304	0.4045	
	Rest_NewEng	0.0700	0.0876	
	South	0.0140	0.1090	
School activities indicator		0.2055	0.0412	<.0001
Number of children		0.2072	0.0442	<.0001
GPA		0.4522	0.0914	<.0001
Reunion indicator		0.6336	0.1207	<.0001
Indicator, non-WPI degree reported		0.3311	0.0791	<.0001
WPI spouse indicator		0.2323	0.1118	<.0001

International student indicator		-0.6172	0.2221	0.0014
Gender	F	3.2056	91.3223	0.0022
	M	2.9052	91.3223	
Honor society count		0.1996	0.0790	0.0037
International club activities count		-0.2815	0.0922	0.0043
Professional society count		0.1580	0.0585	0.0073
Area of B.S. major	Biological/LifeSci	-0.0732	0.1153	0.0096
	BizEconomcs	-0.1607	0.1249	
	ChemicalEngr	-0.1584	0.1253	
	Chemistry	-0.2089	0.2456	
	CivilEngr	-0.0197	0.1062	
	ComputerSci	0.2204	0.1074	
	Electrical/ComputerEngr	0.2432	0.0891	
	HumanitiesArts	0.2724	0.2596	
	Math	0.0732	0.1831	
	MechanicalEngr	0.1162	0.0849	
	Other	-0.0267	0.5126	
	OtherEngr	0.0969	0.4295	
Greek house indicator		0.1422	0.0648	0.0266
Marriage	Divorced	-0.7714	39.2215	0.0374
	Married	-1.0593	39.2208	
	NA	-1.5833	39.2225	
	Other/Partner	-1.3808	39.2240	
	Separated	7.4173	235.3	
	Single	-1.2968	39.2208	

Table B.4 Odds Ratio Estimates for Model 5

Effect				Estimate	95% C.I.	
Alumni activities count				1.750	1.510	2.029
GPA				1.572	1.314	1.880
Years since B.S. awarded				1.097	1.086	1.108
Indicator, non-WPI degree reported				1.392	1.192	1.626
WPI spouse indicator				1.262	1.013	1.571
Number of children				1.230	1.128	1.342
Reunion indicator				1.884	1.487	2.387
Greek house indicator				1.153	1.015	1.309
Indicator, job title reported				1.541	1.374	1.729
International student indicator				0.539	0.349	0.834
International club activities count				0.755	0.630	0.904
Professional society count				1.171	1.044	1.314
School activities indicator				1.228	1.133	1.331
Honor society count				1.221	1.046	1.425
Home geographical region	Mass	vs	West	1.340	1.080	1.662
	Midwest	vs	West	0.950	0.673	1.342
	NA	vs	West	0.603	0.424	0.856
	Northeast	vs	West	1.399	1.070	1.829
	Other	vs	West	1.206	0.478	3.042
	Rest_NewEng	vs	West	1.135	0.898	1.435
	South	vs	West	1.074	0.814	1.416
Marriage	Divorced	vs	Widowed	1.741	0.102	29.789
	Married	vs	Widowed	1.305	0.080	21.297
	NA	vs	Widowed	0.773	0.042	14.269
	Other/Partner	vs	Widowed	0.946	0.046	19.429
	Separated	vs	Widowed	>999.999	<0.001	>999.999
	Single	vs	Widowed	1.029	0.063	16.750

Gender	F	vs	N	>999.999	<0.001	>999.999
	M	vs	N	>999.999	<0.001	>999.999
Area of B.S. major	Biological/LifeSci	vs	Physics	1.352	0.814	2.246
	BizEconomcs	vs	Physics	1.238	0.738	2.079
	ChemicalEngr	vs	Physics	1.241	0.741	2.081
	Chemistry	vs	Physics	1.180	0.595	2.341
	CivilEngr	vs	Physics	1.426	0.867	2.345
	ComputerSci	vs	Physics	1.813	1.102	2.983
	Electrical/ComputerEngr	vs	Physics	1.855	1.147	3.000
	HumanitiesArts	vs	Physics	1.910	0.939	3.883
	Math	vs	Physics	1.565	0.867	2.823
	MechanicalEngr	vs	Physics	1.634	1.013	2.634
	Other	vs	Physics	1.416	0.436	4.600
	OtherEngr	vs	Physics	1.602	0.580	4.423

Appendix C: Logistic Modeling Detail

Table C.1 Class Variable Recoding Detail

Class Var.	Categories	Design Variables											
Category	ALND	1											
	ALUM	-1											
Gender	F	1	0										
	M	0	1										
	N	-1	-1										
Marriage	Divorced	1	0	0	0	0	0						
	Married	0	1	0	0	0	0						
	NA	0	0	1	0	0	0						
	Other/Partner	0	0	0	1	0	0						
	Separated	0	0	0	0	1	0						
	Single	0	0	0	0	0	1						
	Widowed	-1	-1	-1	-1	-1	-1						
B.S. Major	Biological/LifeSci	1	0	0	0	0	0	0	0	0	0	0	0
	BizEconomcs	0	1	0	0	0	0	0	0	0	0	0	0
	ChemicalEngr	0	0	1	0	0	0	0	0	0	0	0	0
	Chemistry	0	0	0	1	0	0	0	0	0	0	0	0
	CivilEngr	0	0	0	0	1	0	0	0	0	0	0	0
	ComputerSci	0	0	0	0	0	1	0	0	0	0	0	0
	Elect./Comp.Engr	0	0	0	0	0	0	1	0	0	0	0	0
	HumanitiesArts	0	0	0	0	0	0	0	1	0	0	0	0
	Math	0	0	0	0	0	0	0	0	1	0	0	0
	MechanicalEngr	0	0	0	0	0	0	0	0	0	1	0	0
	Other	0	0	0	0	0	0	0	0	0	0	1	0

B.S. Major	OtherEngr	0	0	0	0	0	0	0	0	0	0	0	1
	Physics	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Bizstate	Mass	1	0	0	0	0	0	0					
	Midwest	0	1	0	0	0	0	0					
	NA	0	0	1	0	0	0	0					
	Northeast	0	0	0	1	0	0	0					
	Other	0	0	0	0	1	0	0					
	Rest_NewEng	0	0	0	0	0	1	0					
	South	0	0	0	0	0	0	1					
	West	-1	-1	-1	-1	-1	-1	-1					
Home	Mass	1	0	0	0	0	0	0					
	Midwest	0	1	0	0	0	0	0					
	NA	0	0	1	0	0	0	0					
	Northeast	0	0	0	1	0	0	0					
	Other	0	0	0	0	1	0	0					
	Rest_NewEng	0	0	0	0	0	1	0					
	South	0	0	0	0	0	0	1					
	West	-1	-1	-1	-1	-1	-1	-1					
Distinction	D	1	0										
	H	0	1										
	NA	-1	-1										

Table C.2 Summary of Stepwise Selection

Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	***p*-value**
	Entered	**Removed**				
1	bsrecency_mod		1	1	1049.2959	<.0001
2	bizstate_mod		7	2	249.6479	<.0001
3	alum_mod		1	3	149.7490	<.0001

4	child_mod		1	4	77.3250	<.0001
5	schinvolve_mod		1	5	60.0473	<.0001
6	home_mod		7	6	66.3851	<.0001
7	gpa_mod		1	7	41.0654	<.0001
8	reunion_mod		1	8	32.4448	<.0001
9	gender_mod		2	9	21.4961	<.0001
10	nonwpideg_mod		1	10	17.3076	<.0001
11	sps_mod		1	11	10.5696	0.0011
12	intlclub_mod		1	12	8.1032	0.0044
13	honorsoc_mod		1	13	8.2241	0.0041
14	profsoc_mod		1	14	6.0898	0.0136
15	bsmajor_mod		12	15	25.0615	0.0145
16	frat_mod		1	16	4.4279	0.0354
17	distinction_mod		2	17	6.1715	0.0457

Table C.3 Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.0	Somers' D	0.582
Percent Discordant	20.8	Gamma	0.583
Percent Tied	0.2	Tau-a	0.286
Pairs	11883776	c	0.791

Appendix D: Box-Cox Transformation

The second phase of analysis (linear regression model) starts with an initial check for the necessity of transformation on the response variable. Figure D.1 shows the histogram of the response variable with a fitted normal curve. Clearly there is no way to believe it comes from a normal distribution. So a transformation is necessary here. The technique of Box-Cox transformation [1] is then utilized to optimally locate the choice of transformation. Figure D.2 illustrate how the sum of squared errors changes with the choice of different λ, the order of the transformation. Both the software printout and the line plot led to the choice of $\lambda = 0$ which corresponds to a natural log transformation on the contribution amount. Figure 3.4 shows the histogram along with a fitted normal curve of the transformed responses which presents a much more plausible shape.

Figure D.1 Histogram of the Contribution Amount of Contributors

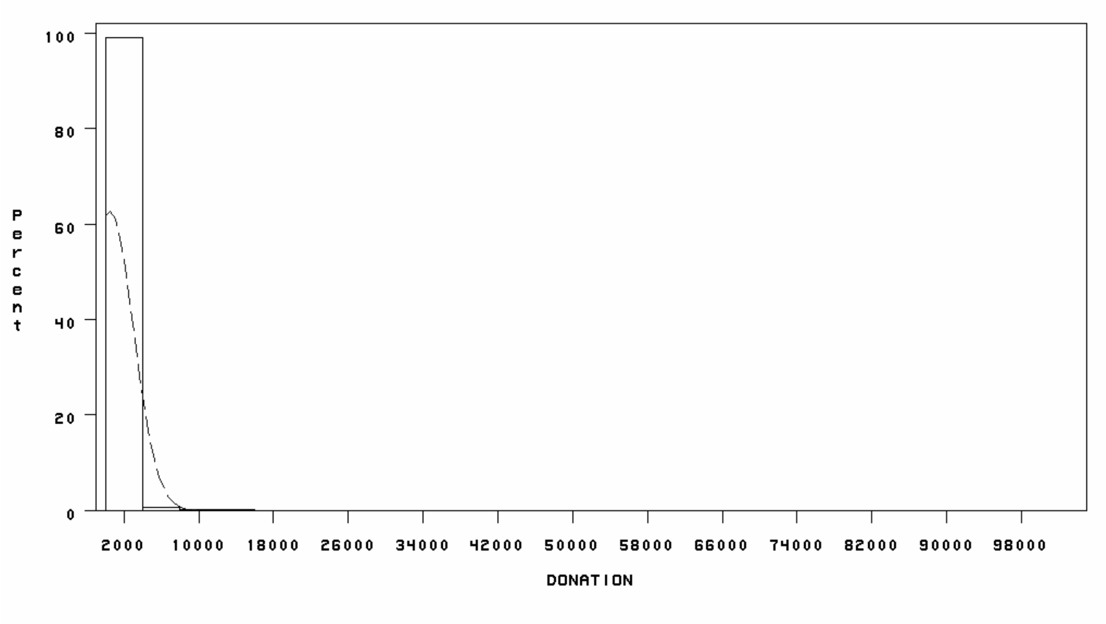
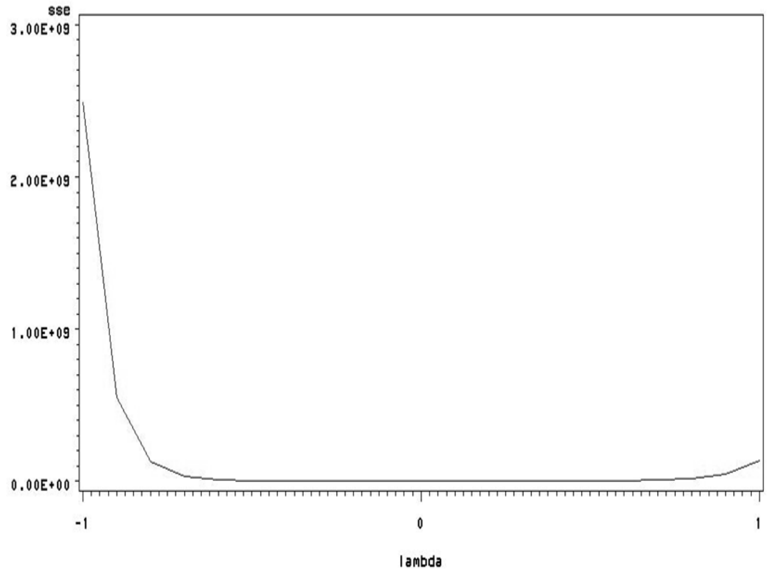


Figure D.2 Plot of Box-Cox Result



lambda	sse
-1.0	2490952332.55
-0.9	546645997.25
-0.8	126806153.67
-0.7	32390347.39
-0.6	9732712.10
-0.5	3699911.40
-0.4	1831143.62
-0.3	1138672.80
-0.2	839830.71
-0.1	703640.77
-0.0	653589.73
0.1	665557.49
0.2	741946.19
0.3	912577.11
0.4	1259187.92
0.5	1989959.76
0.6	3651205.39
0.7	7738583.61
0.8	18504667.41
0.9	48422112.93
1.0	135007693.53

Bibliography

[1] Michael H. Kutner, Christoper J. Nachtsheim, John Neter, William Li. *Applied Linear Statistical Models, fifth edition*. McGraw-Hill, 2005

[2] David W. Jr. Hosmer, Stanley Lemeshow, *Applied Logistic Regression, second edition*. Wiley-Interscience, 2000

[3] Paul D. Allison. *Logistic Regression Using the SAS System: Theory and Application, first edition*. SAS Publishing, 1999

[4] Alan Agresti, *Categorical Data Analysis, second edition*. Wiley-Interscience, 2002

[5] Stokes. *Categorical Data Analysis Using the SAS System, second edition*. WA (Wiley-SAS), 2006

[6] Sharon L. Lohr. *Sampling: Design and Analysis, first edition*. Duxbury Press, 1998

[7] Joseph D. Petruccelli, Balgobin Nandram, Minghui Chen. *Applied Statistics for Engineers and Scientists, first edition*. Prentice Hall, 1999

[8] Ron P. Cody, Jeffrey K. Smith. *Applied Statistics and the SAS Programming Language, fifth edition*. Prentice Hall, 2005

[9] Lora D. Delwiche, Susan J. Slaughter. *The Little SAS® Book: A Primer, third edition*. SAS Publishing, 2003

[10] Ronald P. Cody. *Cody's Data Cleaning Techniques Using SAS Software, first edition*. SAS Publishing, 1999

[11] Katherine Prairie. *The Essential PROC SQL Handbook for SAS Users, first edition*. SAS Publishing, 2005

[12] Kirk Paul Lafler, *Proc SQL: Beyond the Basics Using SAS, first edition*. SAS Publishing, 2004

[13] Ronald P. Cody, Ray Pass, SAS Institute. *SAS Programming by Example, first edition*. SAS Publishing, 1995

[14] Ronald P. Cody. *SAS Functions by Example, first edition*. SAS Publishing, 2004

[15] SAS Institute Inc. *SAS OnlineDoc 9.1.2*,
http://support.sas.com/onlinedoc/912/

[16] David Shepard Associates, Inc. *The New Direct Marketing: How to Implement A Profit-Driven Database Marketing Strategy, third edition*. McGraw-Hill, 1999