A clusterwise linear regression model of alumni giving

Pablo L. Durango-Cohen*

Department of Civil and Environmental Engineering,
Northwestern University,
2145 Sheridan Road, A332,
Evanston, IL 60208-3109, USA
Fax: (847)-491-4011
E-mail: pdc@northwestern.edu
*Corresponding author

Elizabeth J. Durango-Cohen

Stuart School of Business,
Illinois Institute of Technology,
565 West Adams Street,
Chicago, IL 60661, USA
E-mail: durango-cohen@iit.edu

Weizeng Zhang

Department of Civil and Environmental Engineering,
Northwestern University,
2145 Sheridan Road, A310,
Evanston, IL 60208, USA
E-mail: WeizengZhang2009@u.northwestern.edu

Abstract: We present a clusterwise regression model to analyse alumni contributions to a private, PhD-granting university in the Midwestern USA. The model provides a framework to simultaneously segment a population, and to explain the effect of various factors on the mean annual value of donations. We contribute a different approach to marketing studies in the university fundraising context, where segmentation is often based on intuitive, albeit possibly biased criteria. Instead, in clusterwise regression, individuals are assigned to segments with the objective of maximising the within-segment variation explained by a set of regression models. Our main finding is that individuals in different segments display systematic, but unobserved differences in their responses, i.e., the coefficients in the segment-level regression models exhibit differences in their magnitude, sign and level of significance. We discuss how characterising such differences can support tailored solicitation strategies.

Keywords: fundraising; not-for-profit organisations; higher-education finance; educational economics; USA; segmentation; latent class modelling; group effects; cross-sectional heterogeneity; clusterwise linear regression; CLR; exchange algorithm.

Reference to this paper should be made as follows: Durango-Cohen, P.L., Durango-Cohen, E.J. and Zhang, W. (2012) 'A clusterwise linear regression model of alumni giving', *Int. J. Education Economics and Development*, Vol. 3, No. 4, pp.330–347.

Biographical notes: Pablo L. Durango-Cohen is an Associate Professor in the Transportation Systems Analysis and Planning Programme at Northwestern University. He received his PhD in Industrial Engineering and Operations Research from the University of California, Berkeley. His research interests include problems of segmentation, estimation and control with applications in transportation and marketing. His work has been recognised with multiple awards, including an NSF Early CAREER Development Award. He serves on the editorial board of the *Journal of Infrastructure Systems*.

Elizabeth J. Durango-Cohen is an Assistant Professor of Operations Management in the Stuart School of Business at the Illinois Institute of Technology. She received her PhD in Industrial Engineering and Operations Research from the University of California, Berkeley. Her current research efforts focus on the interface of marketing and operations. In addition to her work on direct-marketing optimisation for non-profit institutions, she is also interested in modelling the effect of capacity on pricing decisions for supply chains with competing national and store-brand products.

Weizeng Zhang is a PhD Candidate in the Transportation Systems Analysis and Planning Programme at Northwestern University. He received his BS in Transportation Engineering from Southeast University in Nanjing, China, as well as an MS in Transportation Systems from Northwestern University. His research involves development and application of econometric methods for market segmentation with applications in transportation and fundraising for non-profit organisations.

1 Introduction

Of the $30.30 billion raised by colleges and universities in 2011, 44.3% ($13.45 billion) came from contributions by individuals (Voluntary Support of Education Survey, 2012). Given the magnitude of support individual alumni provide, understanding the variables that drive contributions plays a vital role in the success of university fundraising efforts. This is particularly important in the current economic climate. In 2009 and 2010, colleges and universities in the USA saw record decreases in alumni contributions in 50 years of record-keeping, with drops of 6.2% and 11.9%, respectively (Hall and Joslyn, 2011). Motivated, in part, by the slow economic recovery from the global recession of 2008, and to make up for shortfalls in government funding, universities and other not-for-profit organisations have increased the volume and sophistication of their fundraising efforts.

Quantitative marketing studies of university fundraising/alumni giving can be described as either segmentation studies aimed at grouping individuals in a relevant population, e.g., alumni, active donors, etc., to support tailored solicitation strategies; or econometric studies aimed at describing and predicting donations across a population, as well as identifying and estimating the effect of characteristics associated with giving, such as demographic and socio-economic variables (e.g., Okunade, 1996), or the effect of factors that can influence an individual's propensity/motivation to donate, such as awareness of the need for financial support, academic/athletic reputation, individual recognition, or tax benefits (Gaier, 2005; Holmes, 2009). Numerous studies address both aims simultaneously. Often, econometric models are used to describe the behaviour of previously established segments, e.g., annual fund contributors vs. major donors as in the seminal study of Lindahl and Winship (1992). Common model specifications are employed to identify differences contributing to segment membership.

In this paper, we present a clusterwise linear regression (CLR) model to describe and predict the monetary value of alumni donations at a private, PhD-granting university in the Midwestern USA. Instead of relying on pre-established segments, segmentation is an output of the model's estimation. In addition, the CLR model yields estimates of the effect of explanatory variables on the monetary value of donations for individuals in each of the population segments. To the best of our knowledge, the work constitutes one of the few, possibly the only application of CLR in the university fundraising context. As other studies in the same context employing different approaches: fixed or random effects models in particular, the results show that significant unobserved heterogeneity exists among the individuals in the population.[1] In such models, heterogeneity is captured by different intercepts for each individual in the population, but the coefficients describing the effect of the explanatory variables are assumed constant and homogeneous across the population. In contrast, we find radically different coefficients, i.e., level of significance, magnitude and sign, describing the effect of explanatory variables for individuals in different segments. This is our most important finding. It reinforces the choice of a methodology that provides sufficient flexibility to capture such behavioural differences, and, as discussed in the context of our empirical study, can have significant implications on the design of solicitation strategies.

The remainder of the paper is organised as follows: in Section 2, we position our work with respect to segmentation and econometric models in the university fundraising context. An empirical study of alumni giving behaviour at a PhD-granting university in the Midwestern USA is presented in Section 3. The section includes a detailed description of the data used for analysis (§3.1), the model formulation and an overview of the estimation approach (§3.2), as well as the results and analysis (§3.3). Conclusions, discussion and directions for future work appear in Section 4.

2 Related work

To describe the features that distinguish the proposed approach from others in the literature, we build on Green (1977) and Wedel and Kamakura (2000) who provide an overall taxonomy of quantitative market segmentation models that includes the following dimensions:[2]

1 underlying segmentation approach/method

2 segmentation bases

3 capabilities of econometric model used to explain responses.

We proceed to explain these characteristics further, and use them to describe and organise work appearing in the university fundraising literature, as well as to contrast other approaches to CLR. With respect to the underlying segmentation approach, models are characterised as either *a priori* or *post-hoc*. In *a priori* segmentation, the number and types of segments are determined in advance of the analysis, whereas in *post-hoc* segmentation they are determined as a result of the analysis based on goodness-of-fit or other criteria. The conditions used to assign individuals to segments are referred to as segmentation bases. Segmentation bases are classified as either *observable*, i.e., relying on observed/measured trait, response or institutional variables, or *unobservable*, frequently attributed to latent (psychographic) variables.[3] The capabilities of econometric models depend on factors such as type of model, explicit and implicit assumptions in the formulation and estimation, variables included in the model, data characteristics, etc. Green (1977) describes econometric models as either *predictive* or *descriptive*. Predictive models relate a set of explanatory or independent variables to the outcomes of a set of dependent variables. Descriptive models represent the (joint) distribution of the variables or other characteristics without distinction between outcome and explanatory variables.

The majority of segmentation studies in the university-fundraising context use *a priori* methods relying on observable bases consisting of variables representing demographic or socioeconomic traits (e.g., age, gender, marital status, income), donor patronage [e.g., major donors vs. annual fund contributors (Lindahl and Winship, 1992); donors vs. non-donors (Sun et al., 2007)], and/or contribution frequency [e.g., consistent vs. occasional donors (Wunnava and Lauze, 2001)]. The use of *a priori* methods facilitates interpretation of the results, and the implementation of solicitation strategies that are consistent with managerial practices/intuition that, presumably, determine/influence the number and types of segments. In addition to facilitating interpretation of the results, observable bases are appealing because they support assignment of individuals entering the population into the segments. That is, if a segmentation basis is unobservable, new entrants cannot (easily) be assigned to segments. Predictive models, e.g., regression, are often estimated to describe the contribution behaviour of individuals in each segment. A common specification is used for all segments to make judgments about the significance of different variables in explaining the behaviour of individuals assigned to the different segments. Studies focused on specific subsets, often one segment, of an alumni population, e.g., alumni who graduated between 1988 and 1990 (Marr et al., 2005), alumni with a known valid phone number (Belfield and Beney, 2000), or graduate degree alumni (Okunade, 1996) follow the same logic.

CLR is similar to the aforementioned models in that a set of commonly-specified regression models, one for each segment, are used to explain within-segment variation (in an outcome variable). The main difference is that instead of relying on managerial intuition, the segmentation is obtained during the model estimation, i.e., it is a *post-hoc* method. The criterion used to assign individuals to segments equates to maximising the within-segment variation explained by the segment-level regression models, which

means that the segmentation basis is unobservable. As we explain below, the fact that the segmentation basis is unobservable allows the framework to account for unobserved heterogeneity, which can be significant in the university fundraising context.

In the university-fundraising literature, few models employ *post-hoc* segmentation procedures. As with CLR, the motivation for such procedures is to identify segments in an objective fashion, and, in turn, improve within segment homogeneity and between segment heterogeneity. In contrast to CLR, models in this category are generally descriptive. Recent studies in the university-fundraising literature include Weerts and Ronca (2009), who employ the classification and regression tree (CART) methodology to distinguish between donors and non-donors, and then to predict characteristics tied to alumni giving. They find that giving is correlated with household income, religious background, alumni involvement level, alumni beliefs about institutional needs, and number of institutions competing for alumni gift dollars. As described in Wedel and Kamakura (2000), CART and other tree search procedures are used to set thresholds that are used to assign individuals to segments when there are multiple dependent variables of interest. Le Blanc and Rucks (2009) is a recent study in the university-fundraising context. The former uses a cluster analysis algorithm, based on the nearest centroid sorting method, to segment an alumni population on the basis of average donation. The latter proposes a finite-mixture model to segment individuals on the basis of their contribution sequences, i.e., longitudinal data. The analysis is based on the same data used in this paper, and the results highlight the potential pitfalls of using aggregate statistics to summarise response data.

To position CLR within the context of predictive university-fundraising models, we proceed to describe how such econometric models have been formulated to account of unobserved heterogeneity.

As is the case in the selection of segmentation bases, studies generally assume that within-segment variation can be explained by observed trait, response or institutional variables: personal characteristics (marital status, gender, age, race), socio-economic variables (income and education, past giving, sector of employment, type of financial aid received), behavioural factors (membership in fraternities, volunteering for the college, membership in alumni chapters), and institutional characteristics (size, endowment value). Lindahl and Winship (1992) is a seminal example in the context of university fundraising. They present logit models for both major gifts and annual fund prospects, using both trait/socio-economic and response data. In their analysis, past giving turns out to be the strongest single factor in predicting future giving in both segments, but other factors are also statistically significant.

Recognising that latent (psychographic or institutional) variables can contribute to systematic variation in contribution behaviour, significant effort has been devoted to study the effect of factors that might influence or serve as proxies for such variables, i.e., to "find variables that are observable to development offices" (Monks, 2003). Such factors are typically identified through extensive surveys. Example variables tied to alumni giving that are commonly cited in the literature include: student quality (Baade and Sundberg, 1996), awareness financial support need (Taylor and Martin, 1995; Bekkers and Wiepking, 2007; Pearson, 1999; Weerts and Ronca, 2008), satisfaction with undergraduate experience (Monks, 2003; Gaier, 2005; Clotfelter, 2001, 2003), altruism (Becker, 1974; Andreoni, 1989; Kim et al., 2011), and prestige/recognition (Yoo and Harrison, 1989).

In addition to identifying proxies for latent variables, formulation of fixed or random effects models have been used to capture the effect of unobserved heterogeneity. Such model specifications capture heterogeneity with different intercepts for each individual in the population. Netzer et al. (2008) and Holmes (2009) are relevant examples in the university fundraising context. The former presents a random effects logit model to predict the likelihood of a donation in a given year. The latter presents random effects probit and tobit models to predict the likelihood and monetary value of donations. In both studies, comparison of the results with benchmark models led to the conclusion that unobserved heterogeneity is significant.

Instead of including proxy variables, fixed or random effects, CLR models rely on unobservable segmentation bases to capture unobserved heterogeneity. The structure of the model allows for heterogeneity to manifest itself through different coefficients, i.e., level of significance, magnitude and sign, describing the effect of (observed) explanatory variables for individuals in different segments. This is radically different than the other approaches where the effect of explanatory variables is assumed constant and homogeneous across the donor population. Perhaps more importantly, as discussed in the empirical study, characterising differences in the segment-level responses supports the development of tailored solicitation strategies.

3 Empirical study

This section presents our analysis of the CLR model we estimated to study the monetary value of contributions at a private, PhD-granting university in the Midwestern USA. We begin in Section 3.1 by summarising the data used in the study. We then proceed to describe the formulation, as well as our implementation of the the estimation approach in Section 3.2. Our results and analysis, including extensive model diagnosis, are presented in Section 3.3.

3.1 Data

The data used in the analysis were provided by a major research university in the Midwestern USA. The data include 31,762 contributions by 8,910 individuals who donated to the university at least once during fiscal years 2000 through 2009.[4] Each contribution corresponds to the total amount that an individual donated in a given year. Major donors, with at least one donation exceeding $5,000, were excluded from the analysis. Table 1 summarises the contribution data for the ten-year period.

Overall, we observe that the mean amount per contribution displays an increasing trend over the 10-year period. The monetary values corresponding to the 75th percentile of the yearly distributions are significantly lower than the annual means, indicating skewness toward small values. The trend in the number of contributions per year is less clear, though we do observe sharp declines from peaks in 2000, 2001 and 2007. Relative to the size of the donor population, the number of contributions reflects intermittent patterns, which are an important feature of the data. To describe further, we note that of the 8,910 individuals in the dataset 33.8% contributed only once, 16.9% twice, 11.4%, 7.4%, 6.2%, 4.9%, 5.3%, 4.5%, 5.1%, 4.6% contributed 3, 4, 5, 6, 7, 8, 9, and 10 times, respectively.

Table 1 Data summary

Fiscal year	*Number of contributions*	*Mean amount per contribution*	*Std. deviation across contributions*	*Contribution percentiles*		
				25th	*50th*	*75th*
2000	4,037	$169	$603	$25	$50	$100
2001	3,558	$161	$565	$25	$50	$100
2002	2,753	$170	$525	$25	$50	$100
2003	2,996	$170	$576	$25	$50	$100
2004	2,874	$191	$574	$25	$54	$100
2005	2,901	$188	$496	$30	$75	$146
2006	3,147	$203	$707	$25	$75	$101
2007	3,409	$228	$731	$30	$80	$150
2008	3,165	$209	$635	$35	$80	$150
2009	2,922	$207	$551	$35	$100	$150

In addition to the contribution data summarised above, for each individual in who contributed, the dataset includes age, gender, field of study, degrees earned (i.e., BS, MS or PhD), graduation year, college (e.g., architecture, business, design, engineering), and residential zip/postal code. These data, as well as derived statistics, are summarised in Table 3.

3.2 Model formulation

In this section, we present a CLR model to explain the contribution behaviour of the donor population described earlier. We also provide an overview of the procedure to estimate the model, including details of our implementation. For pedagogical reasons, as well as to establish a benchmark for comparison, we begin, however, by presenting a population-level regression model. The model appearing in equation (1), describes the average monetary value of contributions during the decade FY00-FY09.

$$\begin{aligned} lll y_i = {} & \beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{AGESQ}_i + \beta_3 \text{UD}_i + \beta_4 \text{AD}_i \\ & + \beta_5 \text{GENDER}_i + \beta_6 \text{PD}_i + \beta_7 \text{INCOME}_i + \epsilon_i \end{aligned} \tag{1}$$

where y_i corresponds to the natural logarithm of individual i's average annual contribution over the analysis period; AGE and AGESQ, respectively correspond to the individual's age and the square of his/her age; UD, AD, and PD are indicator variables with values of 1, when the individual earned undergraduate, i.e., BA/BS, graduate, i.e., MA/MS or PhD, or professional, i.e., business or law, degrees at the institution; GENDER is an indicator variable with value of 1 for males and 0 for females; INCOME corresponds to th natural logarithm of the median household income in the zip/postal code of individual i's residence; ϵ_i is the error term; finally, β_0 is the intercept, and $\beta_l,\ l = 1, \ldots, L$ are the coefficients associated with the L explanatory variables. For the model presented in (1), $L = 7$. For simplicity, we collect the explanatory variables associated with i in the vector, $\boldsymbol{x}_i \equiv [x_{1i}, \ldots, x_{Li}]$, and rewrite (1) as shown below.

$$y_i = \beta_0 + \sum_{l=1}^{L} \beta_l x_{li} + \epsilon_i \tag{2}$$

CLR is a statistical framework that allows generalisations of the above model. Rather than estimating a single set of coefficients, $\beta_l, l = 0, \ldots, L$, describing the effect of the explanatory variables across the entire population, the framework assumes that the population is comprised of K segments. The framework supports simultaneous assignment of individuals to segments, as well as estimation of a set of coefficients for each of the segments. Building on (2), a CLR model is shown in equation (3) below:

$$y_i = \sum_{k=1}^{K} a_i^k \left[\beta_0^k + \sum_{l=1}^{L} \beta_l^k x_{li} \right] + \epsilon_i \tag{3}$$

Following the original CLR model of Späth (1977, 1979, 1982), the indicator variable, $a_i^k = \{1, \text{ when individual } i \text{ is assigned to segment } k; \ 0, \text{ otherwise}\}$, describes the assignment of individual i, and the coefficients, $\beta_l^k, l = 0, \ldots, L$, capture the effect of the explanatory variables for individuals assigned to segment k. Also, each individual is assigned to exactly one segment, i.e., $\sum_{k=1}^{K} a_i^k = 1$.[5]

The coefficients in the model, including the assignment variables, are obtained by solving the problem of minimising the residual sum of squares (RSS), presented in equation (4) below:

Minimise:

$$\text{RSS} \equiv \sum_{i=1}^{I} \left(y_i - \sum_{k=1}^{K} a_i^k \left[\beta_0^k + \sum_{l=1}^{L} \beta_l^k x_{li} \right] \right)^2 \tag{4}$$

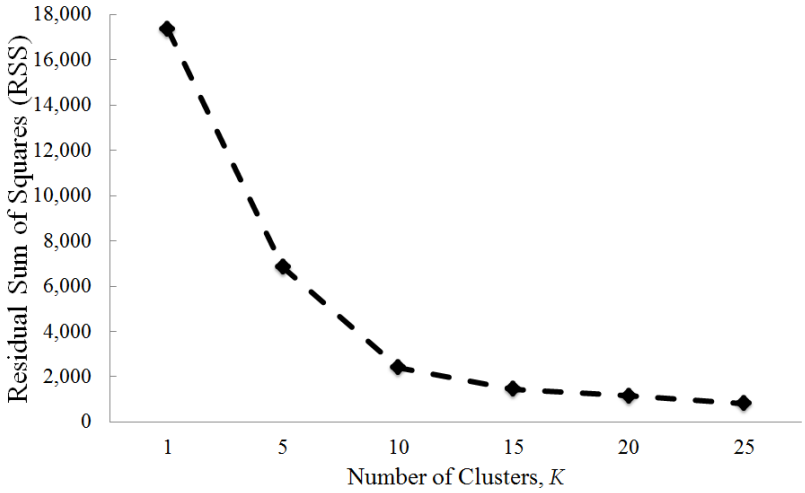
We use the exchange algorithm of Späth (1977, 1979, 1982) to solve the above optimisation problem. For a given number of segments, K, an initial, random assignment of individuals to segments is provided, and the algorithm is used to iteratively adjust the (binary) assignment (variables) to reduce the RSS criterion. This operation, in turn, depends on estimation of the coefficients in the model, capturing the effect of the explanatory variables for the individuals assigned to each of the segments. In our implementation, the coefficients for each of the segments are computed using the ordinary least squares (OLS) closed-form expressions.[6] We set a 5% RSS reduction as the algorithm's convergence/termination criterion. To reduce the likelihood of convergence to a local minimum, for a given K, we ran ten instances of the exchange algorithm: each instance with a different initial assignment. The results reported are for the instance with the smallest RSS.

3.3 Results and discussion

Figure 1 displays the RSS as a function of the number of segments, K, for K between 1 and 25. As expected, the rate at which RSS improves is a decreasing function of K. In terms of selecting a preferred K for analysis, depending on the specific application, in addition to trading off a parsimonious model with achieving high goodness-of-fit, the selection ought to account for the increased complexity of managing larger numbers of segments. In the remainder of the section, we analyse a 10-segment model, which

provides satisfactory fit. The discussion is focused on generating insights stemming from comparison of the 10-segment model with the population-level regression model, as well as describing managerial/behavioural implications of the results. To some extent, our observations are independent of the number of segments in the model selected for analysis.

Figure 1 RSS vs. number of segments



The estimation results for the population-level regression model and for the 10-segment CLR model are presented in Table 2. The table includes the coefficients and p-values, as well as the ensuing (within-segment) R^2. R^2 can be interpreted as the percentage of the (within-segment) variation that is explained by a (segment-level) regression model. p-values below 0.01 (E-02) are not reported.[7] Coefficients with associated p-values greater than 0.05, presented in bold-face type, can be interpreted as statistically-insignificant at the 95% confidence level. Table 3, used for analysis and interpretation of the results, summarises average trait and response data for the individuals in each segment, as well as the averages across the population.

The main observation that follows from results presented in Table 2 is that the segment-level coefficients exhibit appreciable differences in terms of magnitude, sign, and level of significance. They are also different than the coefficients obtained from the population-level regression model. This observation, i.e., the fact that certain variables can exhibit radically-different effects for (subsets of) individuals in the same alumni/donor population, validates the use of CLR as a methodology to represent the contribution data, and, as described in the following examples, can have implications on the design of fundraising strategies.

- The coefficient associated with the variable GENDER captures the effect of gender on donations. The results for the population-level model, indicate that the effect is positive, which means that a (randomly selected) male donor is more likely to donate more than a (randomly selected) female donor (*ceteris paribus*). The fact that the sign of the coefficient varies across the segments indicates that the average result for the population does not apply to all segments. University

development offices might be well-served, as a result, to pursue males or females in different segments more or less aggressively.

- Analysing the coefficients obtained for AGE and AGESQ yields insight about the effect of age for individuals in each segment. The signs obtained for the Segments 1, 2, 3 and 7 coefficients mean that the dependent variable is a convex function of age. This is not an unexpected result considering usual career/income trajectories. An implication is that donations for individuals in Segment 2 are expected to decrease up age 38.5, and increase thereafter. For all other segments donations are increasing as a function of age (over the relevant range). As reported in the literature, the function is concave for the other segments, as well as for the population as a whole. This result suggests that there could be value in adjusting the timing/messaging of specific solicitations for each segment to reflect different potential. For example, a common alumni solicitation strategy is to encourage participation in the first few years after graduation. Based on the results, it appears that there could be value in extending this period for certain population segments. Also, particularly for segments exhibiting a convex relationship between age and monetary value, it makes sense to (continue to solicit) older individuals.
- Median family income for an individual's residence zip/postal code, the variable INCOME, is used as a proxy for socio-economic factors. The positive coefficient of the corresponding variable in the population-level model indicates positive correlation between income and the average monetary value of donations across the entire sample. Statistically-insignificant or (relatively large) negative coefficients in three out of ten segments, mean that this relationship does not hold across the population. Our conjecture is that psychographic variables play a role in this result. Such variables could be influenced by factors such as correspondingly large living/housing expenses, leading to smaller disposable incomes, in locations (zip codes) with higher average family incomes; or keeping in mind that various (charitable/non-profit) organisations solicit high-income areas more aggressively, by decisions to allocate resources to multiple/additional organisations. Ultimately, the factors elicit different responses from the different populations segments. Importantly, the result suggests different solicitation/messaging strategies, even for individuals who live within the same zip code.
- The coefficient of the undergraduate dummy variable, UD, reveals the effect of an individual obtaining an undergraduate degree from the institution on the observed donations. In segments with positive coefficients, donors who attended the university as undergraduates are expected to donate more, which may motivate the university to put more effort to reach its undergraduate alumni in the segment. Interpretation of the coefficients for AD and PD is similar. The signs and (relative) magnitudes of the coefficients might influence segment-level marketing strategies. For example, segments with large, positive coefficients for AD, might be sent solicitations emphasising faculty/department/college-level initiatives, as opposed to university-wide programmes; or networking/professional opportunities/events might reinforce behaviour of segments with large, positive PD coefficients.

Table 2 Estimation results: coefficients, (non-zero) p-values, and R^2

Variable	*Seg 1*	*Seg 2*	*Seg 3*	*Seg 4*	*Seg 5*	*Seg 6*	*Seg 7*	*Seg 8*	*Seg 9*	*Seg 10*	*Pop.*
Constant	–2.24	–2.79	–2.31	–3.21	25.15	–8.66	9.86	–8.99	–3.80	–34.12	–3.8009
	–1.11E–02	–6.29E–02	*–1.18E–02*	5.44E–02	1.02E–01	1.35E–01	–1.09E–02	1.19E–01	2.53E–01	1.26E–01	7.774E–02
AGE	*0.11*		*0.23*								
AGESQ	2.30E–04	8.17E–04	1.67E–04	–2.10E–04	–8.83E–04	–1.10E–03	8.37E–04	–9.87E–04	–2.00E–03	–5.23E–04	–4.08E–04
	0.06	–1.34	–0.39	2.31	–1.26	1.60	–0.40	1.13	1.65	–1.03	0.28
UD	*0.15*										
	1.23	0.38	–1.96	0.59	0.12	1.05	2.67	0.65	0.47	0.53	0.53
AD					*0.02*						
	0.19	–0.34	–0.14	–0.41	–0.18	0.29	–1.15	3.76	–1.15	–0.88	0.096
PD											*0.03*
	0.22	–0.39	0.08	–0.55	2.10	0.54	–0.80	0.19	–0.50	0.14	0.093
GENDER			*0.02*							*0.01*	*0.02*
	0.27	0.57	0.51	0.24	–2.31	0.61	–0.79	0.58	*0.03*	2.97	0.29
INCOME									*0.53*		
R^2	0.55	0.78	0.83	0.90	0.85	0.77	0.89	0.88	0.79	0.86	0.12

Table 3 Segment characteristics

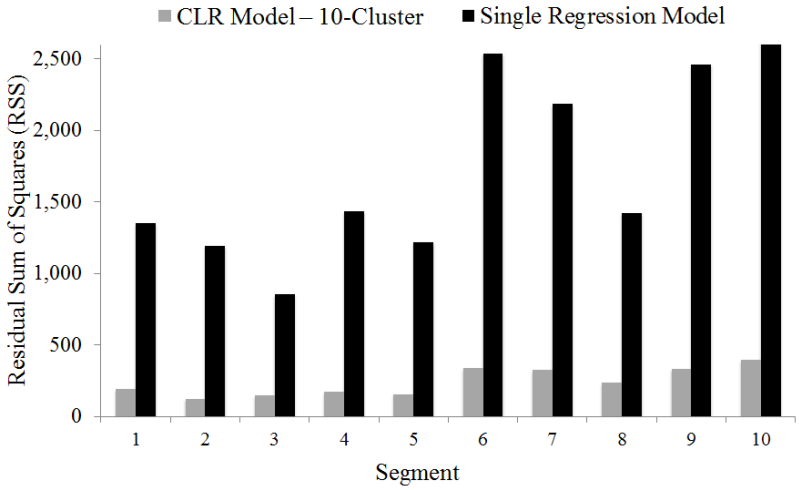
Variable	*Seg 1*	*Seg 2*	*Seg 3*	*Seg 4*	*Seg 5*	*Seg 6*	*Seg 7*	*Seg 8*	*Seg 9*	*Seg 10*	*Pop.*
Av. AGE	59.78	58.41	57.15	56.49	58.06	57.06	57.50	57.55	58.79	57.48	57.86
UD = 1	53.12%	51.61%	51.11%	43.34%	52.73%	49.57%	53.73%	63.16%	38.95%	54.42%	51.16%
AD = 1	30.06%	33.43%	35.48%	39.54%	35.24%	33.39%	29.32%	28.1%	42.45%	35.72%	34.31%
PD = 1	30.62%	32.26%	32.93%	40.41%	34.74%	34.23%	32.44%	25.39%	46.76%	34.49%	34.47%
GENDER = 1	80.81%	80.79%	76.94%	74.21%	74.94%	77.67%	80.71%	80.12%	77.8%	81.19%	78.54%
Av. median income ($)	75,718	77,013	76,257	78,032	75,243	76,502	76,688	75,247	77,662	74,172	76,244
Av. donation ($)	7.56	10.63	16.46	26.97	43.39	49.02	81.31	102.23	146.89	174.57	67.42
Std. dev. donation ($)	7.43	16.00	14.31	35.99	120.47	45.36	218.61	297.08	317.10	383.08	211.74
Coef. of variation	0.98	1.50	0.87	1.33	2.78	0.93	2.69	2.91	2.16	2.19	3.14
Max donation ($)	90	235	100	330	1,929	320	2,600	3,310	3,585	2,785	3,585
No. donors	1,058	682	902	923	806	815	897	961	973	893	1,866

The statistics presented in Table 3 suggest that the segments are homogeneous in terms of the characteristics captured by the explanatory and response variables. For example, the range of the average age across the ten segments is 56.49 (Seg. 4) to 59.78 (Seg. 1). It appears that the mean annual donation per donor,[8] which we use to order the segments, may partially explain the between segment variation, i.e., it could play a role in the segmentation basis. The large coefficients of variation, however, reflect high within segment variation, and consequently, significant overlap between segments. In short, the results emphasise that segmentation bases in CLR are related to the regression models' capability to explain within-segment variation, i.e., unobservable bases. Based on the within-segment R^2 statistics reported in Table 2, we note that the regression models explain large percentages of the within-segment variation. Further, we note that the overall R^2 for the CLR model is 0.88, which is significantly greater than the 0.12 for the population-level regression model. Thorough analysis of the variation captured by the CLR model is presented in Section 3.3.1.

Another observation that follows from Table 3 is that the ensuing segments are relatively balanced in terms of the number of individuals. This result is likely related to data characteristics, the variables in the model, the implementation of the exchange algorithm, etc. As discussed in Simester et al. (2006), this result can be attractive from a managerial perspective. Also, sometimes unbalanced segments are indicative of 'overfitting' the data.

To conclude this section, we use Figure 2 to illustrate the pitfalls of using an 'average' model to make predictions about a population that exhibits (unobserved) heterogeneity. In particular, the figure compares predictions of the dependent variable generated by the population-level model (1) vs. those generated by the clusterwise regression model (3). From a marketing perspective, the fact that individuals do not respond in a homogeneous fashion, as predicted by population-level models, can have negative consequences. In addition to more accurate predictions, CLR models provide segment-level descriptions of the effect of explanatory variables.

Figure 2 Model predictions



3.3.1 Model diagnosis

In this section, we follow Brusco et al. (2008) to diagnose the adequacy of the 10-segment CLR model presented earlier. In particular, they explain that CLR models have significant potential to overfit the data, and lead to meaningless segment-level regression models. The problem arises because the estimation process, i.e., minimisation of the RSS, is carried out without regard to the relative proportion of variation explained by the segmentation vs. variation explained by the segment-level regression models. In technical terms and borrowing the notation from Brusco et al. (2008), the total variation of the dependent variable, denoted TSY, a constant for a given dataset, can be written as follows:

$$\text{TSY} = \sum_{i=1}^{I} (y_i - \bar{y})^2 \tag{5}$$

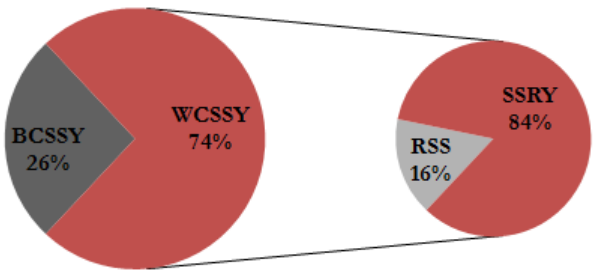
where $\bar{y}$ is the mean value of the dependent variable, i.e., $\bar{y} \equiv \frac{\sum_{i=1}^{I} y_i}{I}$. In turn, TSY can be decomposed into the variation explained by the segmentation, BCSSY, and the within-segment variation, WCSSY. WCSSY can be further decomposed into the variation explained by the segment-level regression models, SSRY, and RSS. In summary,

$$\text{TSY} = \text{BCSSY} + \text{SSRY} + \text{RSS} \tag{6}$$

Because TSY is constant, minimisation of RSS maximises BCSSY+SSRY. If BCSSY is much greater than SSRY, the segment-level regression models have less value because most variation is explained by the segmentation. That is, except for the constant/intercept terms, individuals respond in a relatively homogeneous fashion.

To validate the 10-segment model, we calculate the TSY, BCSSY, WCSSY, SSRY and RSS and present the results in Figure 3. As shown in the figure, BCSSY accounts for 26% of TSY; SSRY accounts for 84% of WCSSY, which is 74% of TSY. The results show that the variation of the dependent variable is mostly explained by the segment-level regression models rather than the segmentation.

Figure 3 Percentage of total variation, TSY, explained by 10-segment CLR model components (see online version for colours)



4 Conclusions

We present a CLR model to analyse contribution behaviour at a private, PhD-granting university in the Midwestern USA. The model provides a framework to simultaneously segment a population of donors, and to explain and predict the effect of various factors that play a role in their contribution behaviour, i.e., the mean annual value of their donations. The work herein contributes a different approach to the university fundraising literature, which has focused on studying segments established in advance of the analysis based on intuitive, albeit possibly biased, criteria. Instead, in CLR individuals are assigned to segments with the objective of maximising the within-segment variation explained by a set of regression models, i.e., an objective criterion.

The dataset used in our study included over 31,000 aggregate annual contributions by approximately 9,000 individuals who donated to the university between Fiscal years 2000 and 2009. We use a 10-segment model to elaborate on the salient features of CLR. The model yields a reasonable number of segments, and was selected because it provides adequate overall fit-to-data. Moreover, rigorous diagnosis led to the conclusion that the regression models provide excellent explanations of the within-segment variation, i.e., the models are meaningful and the application of CLR is valid.

Our most significant finding is that individuals in different segments display systematic but unobserved differences in their behavioural responses, i.e., the coefficients, describing the effect of age, gender, income and degrees earned, in the segment-level regression models exhibit differences in their magnitude, sign and level of significance. We provide an extensive discussion on how such differences could be exploited by university development offices to devise tailored segment-level solicitation strategies. This result highlights the strength of CLR compared to other econometric approaches to capture unobserved heterogeneity, fixed and random effects models in particular, which rely on the assumption that the coefficients describing the effect of the explanatory variables are constant and homogeneous across the population.

Possible extensions motivated by limitations of the CLR model include the development of clusterwise tobit models (Jedidi et al., 1993) to account for the bias that is introduced by ignoring alumni who decided not to donate over the analysis period. Regression is not well-suited to address this technical problem, which has been dealt with by recent work in the university fundraising literature, e.g., Holmes (2009). A related direction would be to formulate models capable of processing the raw contribution sequences. Finally, and in spite of obtaining satisfactory and managerially-interesting results, we note that there is a vast literature addressing limitations of the CLR model and exchange algorithm, introduced by Späth (1977, 1979, 1982), and used herein. Examples include the work of Lau et al. (1999), who proposes rigorous solution procedures for the estimation problem, and the work of DeSarbo and Cron (1988), Wedel and SteenKamp (1989), DeSarbo et al. (1989) and Brusco et al. (2003), who propose generalisations of the original model or allow for estimation under more general assumptions. Building on such work could improve the rigor of the model, and lead to additional insights.

References

Andreoni, J. (1989) 'Giving with impure altruism: applications to charity and Ricardian equivalence', *Economics of Education Review*, Vol. 97, No. 6, pp.1447–1458.

Baade, R. and Sundberg, J. (1996) 'What determines alumni generosity?', *Economics of Education Review*, Vol. 15, No. 1, pp.75–81.

Becker, G.S. (1974) 'The theory of social interactions', *Journal of Political Economy*, Vol. 82, No. 6, pp.1063–1093.

Bekkers, R. and Wiepking, P. (2007) 'Generosity and philanthropy: a literature review', Working paper.

Belfield, C. and Beney, A. (2000) 'What determines alumni generosity? Evidence from the UK', *Education Economics*, Vol. 8, No. 1, pp.65–80.

Brusco, M., Cradit, J. and Tashchian, A. (2003) 'Multicriterion clusterwise regression for joint segmentation settings: An application to customer value', *Journal of Marketing Research*, Vol. 40, No. 2, pp.225–234.

Brusco, M., Cradit, J.D, Steinley, D. and Fox, L. (2008) 'Cautionary remarks on the use of clusterwise regression', *Multivariate Behavioral Research*, Vol. 43, No. 1, pp.29–49.

Clotfelter, C. (2001) 'Who are the alumni donors? Giving by two generations of alumni from selective colleges', *Nonprofit Management & Leadership*, Vol. 12, No. 2, pp.119–138.

Clotfelter, C. (2003) 'Alumni giving to elite private colleges and universities', *Economics of Education Review*, Vol. 22, No. 2, pp.109–120.

DeSarbo, W. and Cron, W. (1988) 'A maximum likelihood methodology for clusterwise linear regression', *Journal of Classification*, Vol. 5, No. 1, pp.249–282.

DeSarbo, W., Oliver, R. and Rangaswamy, A. (1989) 'A simulated annealing methodology for clusterwise linear regression', *Psychometrika*, Vol. 54, No. 4, pp.707–736.

Gaier, S. (2005) 'Alumni satisfaction with their undergraduate academic experience and the impact of alumni giving and participation', *International Journal of Educational Advancement*, Vol. 5, No. 4, pp.279–288.

Green, P. (1977) 'A new approach to market segmentation', *Business Horizons*, Vol. 20, No. 1, pp.61–73.

Hall, H. and Joslyn, H. (2011) *Giving Rose by 2.1% Last Year, New Estimate Shows*, The Chronicle of Higher Education, 20 June [online] http://chronicle.com/article/Giving-Rose-by-21-Last-Year/127954/ (accessed 14 December 2012).

Holmes, J. (2009) 'Prestige, charitable deductions and other determinants of alumni giving: evidence from a highly selective liberal arts college', *Economics of Education Review*, Vol. 28, No. 1, pp.18–28.

Jedidi, K., Ramaswamy, V. and Desarbo, W. (1993) 'A maximum likelihood method for latent class regression involving a censored dependent variable', *Psychometrika*, Vol. 58, No. 3, pp.375–394.

Kim, M., Gibson, H. and Ko, Y. (2011) 'Understanding donors to university performing arts programs: who are they and why do they contribute?', *Managing Leisure*, Vol. 16, No. 1, pp.17–35.

Lau, K., Leung, P. and Tse, K. (1999) 'A mathematical programming approach to clusterwise regression model and its extension', *European Journal of Operational Research*, Vol. 116, No. 3, pp.640–652.

Le Blanc, L. and Rucks, C. (2009) 'Data mining of university philanthropic giving: cluster-discriminant analysis and Pareto effects', *International Journal of Education Advancement*, Vol. 9, No. 2, pp.64–82.

Lindahl, W. and Winship, C. (1992) 'Predictive models for annual fundraising and major gift fundraising', *Nonprofit Management and Leadership*, Vol. 3, No. 1, pp.43–63.

Marr, K., Mullin, C. and Siegfried, J. (2005) 'Undergraduate financial aid and subsequent alumni giving behavior', *Quarterly Review of Economics*, Vol. 45, No. 1, pp.123–143.

Monks, J. (2003) 'Patterns of giving to one's alma mater among young graduates from selective institutions', *Economics of Education Review*, Vol. 22, No. 2, pp.121–130.

Netzer, O., Lattin, J. and Srinivasan, V. (2008) 'A hidden Markov model of customer relationship dynamics', *Marketing Science*, Vol. 27, No. 2, pp.185–204.

Okunade, A. (1996) 'Graduate school alumni donations to academic funds: micro-data evidence', *American Journal of Economics and Sociology*, Vol. 55, No. 2, pp.213–229.

Pearson, J. (1999) 'Comprehensive research on alumni relationships: four years of market research at Stanford', *New Directions for Institutional Research*, Vol. 101, No. 1, pp.5–21.

Simester, D., Sun, P. and Tsitsiklis, J. (2006) 'Dynamic catalog mailing policies', *Management Science*, Vol. 52, No. 5, pp.683–696.

Späth, H. (1977) 'Computational experience with the exchange method', *European Journal of Operational Research*, Vol. 1, No. 1, pp.23–31.

Späth, H. (1979) 'Algorithm 39: clusterwise linear regression', *Computing*, Vol. 22, No. 4, pp.367–373.

Späth, H. (1982) 'Algorithm 48: a fast algorithm for clusterwise linear regression', *Computing*, Vol. 29, No. 2, pp.175–181.

Sun, X., Hoffman, S. and Grady, M. (2007) 'A multivariate causal model of alumni giving: implications for alumni fundraisers', *International Journal of Educational Advancement*, Vol. 7, No. 4, pp.307–332.

Taylor, A. and Martin, J. (1995) 'Characteristics of alumni donors and nondonors at a research I, public university.', *Research in Higher Education*, Vol. 36, No. 3, pp.283–302.

Voluntary Support of Education Survey (2012) Council for Aid to Education, New York, NY.

Wedel, M. and Kamakura, W. (2000) *Market Segmentation: Conceptual and Methodological Foundations*, Kluwer Academic Publishers, Norwell, MA.

Wedel, M. and SteenKamp, J. (1989) 'Fuzzy clusterwise regression approach to benefit segmentation', *International Journal of Research in Marketing*, Vol. 6, No. 4, pp.241–258.

Weerts, D. and Ronca, J. (2008) 'Characteristics of alumni donors who volunteer at their alma mater', *Research in Higher Education*, Vol. 49, No. 3, pp.274–292.

Weerts, D. and Ronca, J. (2009) 'Using classification trees to predict alumni giving for higher education', *Education Economics*, Vol. 17, No. 1, pp.95–122.

Wunnava, P. and Lauze, M. (2001) 'Alumni giving at a small liberal arts college: evidence from consistent and occasional donors', *Economics of Education Review*, Vol. 20, No. 6, pp.533–543.

Yoo, Y. and Harrison, W. (1989) 'Altruism in the market for giving and receiving: a case of higher education', *Economics of Education Review*, Vol. 8, No. 4, pp.367–379.

Notes

1 Unobserved heterogeneity refers to individual-specific, but unobserved factors that contribute to systematic differences between individuals. These differences are unobserved either because they are not observable/measurable, e.g., personality traits that might explain an individual's propensity to donate in response to solicitations, or because data are (inadvertently) missing.

2 *Clustering* is often used instead of *segmentation* in other disciplines.

3 We note that segmentation bases can be unobserved due to missing data.

4 Fiscal years begin June 1 and end May 31, e.g., fiscal year 2001, hereafter labelled 'FY01', runs from June 1, 2000 through May 31, 2001.

5 Other versions of the model presented in equation (3), e.g., the linear relaxation where $0 \leq a_i^k \leq 1$ and $\sum_{k=1}^{K} a_i^k = 1$, have also appeared in the literature.

6 The estimation approach, therefore, relies on the assumptions such as strict exogeneity, independence among the explanatory variables, i.e., absence of multicollinearity, and spherical error variance. Other, more general approaches, have appeared in the literature.

7 The computation of the p-values relies on the assumption that residuals are independent and identically-distributed normal random variables.

8 We note that the mean annual donation per donor includes \$0 transactions in years of inactivity, which explains why the figures differ from the mean amount per contribution reported in Table 1.