NON-PARAMETRIC ESTIMATION OF MEAN CUSTOMER LIFETIME VALUE

PHILLIP E. PFEIFER AND HEEJUNG BANG

PHILLIP E. PFEIFER

is Alumni Research Professor at the Darden Graduate School of Business, University of Virginia, Charlottesville; e-mail: Pfeiferp@virginia.edu

HEEJUNG BANG

is an Assistant Professor, Division of Biostatistics & Epidemiology, Weill Medical College of Cornell University, New York, NY.

The authors thank Professors D. Y. Lin and A. A. Tsiatis for their very helpful comments on earlier drafts of this paper.

This paper is about how to use data from a random sample of customer relationships to calculate an appropriate average customer lifetime value (CLV). When the sample contains only completed relationships, the simple unweighted average is appropriate. When the sample contains a mix of active and completed relationships, the lifetimes of the active relationships are said to be right censored because the observed lifetime to date is but a lower bound on the eventual lifetime. Because of this censoring, a simple average of the sample CLVs to date will be a biased estimate of the mean CLV. This paper presents and explores several non-parametric estimation methods for correcting for this bias.

JOURNAL OF INTERACTIVE MARKETING VOLUME 19 / NUMBER 4 / AUTUMN 2005
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dir.20049

INTRODUCTION

This paper is about how to calculate an appropriate average customer lifetime value (CLV) for a sample of n customer relationships. Although there are a host of articles describing how to calculate CLV as a function of retention rates, gross margin per transaction or per period, remarketing costs, and a hurdle rate (see, for example Berger & Nasr, 1998; Blattberg & Deighton, 1996; Dwyer, 1989; Hughes, 1997; and Pfeifer & Carraway, 2000), very few articles offer advice to managers on how to estimate the necessary inputs (Berger, Bolton, Bowman, Briggs, Eleten, Kumar, Parsuraman, & Terry, 2002 is a notable exception). As suggested by Jain and Singh (2002, p. 44), "research is needed into estimation methods that provide stable, consistent, and unbiased estimates (of CLV)."

The purpose of this paper is to begin to do just that by addressing the question of how to use customer-level data to estimate an average or mean CLV.

To be more specific, this paper addresses how to use data from a random sample of n customer relationships to estimate the mean CLV of the population of relationships from which the random sample was drawn. For example, if the n relationships all came from a particular source, then the mean of the population represents the long-run average CLV of relationships from this same source. Although the random-sample assumption might seem unduly restrictive, it is the usual assumption invoked whenever one uses data to draw inferences about a population. The construction of confidence intervals, for example, requires this random-sample assumption.

The mean CLV is of particular interest to interactive marketers because it represents the long-run average (dollar) value of a relationship from the population. The mean CLV is the metric used to make informed prospecting decisions in that it represents a limit on spending designed to initiate new relationships from the population. So, for example, after a cruise line estimated a surprisingly high mean CLV of $4,581 for its population of new customers, it decided to adjust its acquisition spending upwards (Berger, Weinberg, & Hanna, 2003).

The mean CLV is usually sufficient for making marketing decisions affecting groups of customers (prospects) when it is either infeasible or uneconomical to capitalize on individual differences. Prospecting decisions about media, copy, promotions, and sales force can often be made based on how each of these affects the average CLV of the customers acquired without regard for individual differences.

Even decisions about how to treat existing customers are sometimes made based on group averages. Consider a firm conducting an experiment to decide how to treat a segment of customers reaching a specified milestone (e.g., their first called-in complaint, the graduation of their oldest child, or their first missed payment). The firm might apply treatments A and B randomly to the next several customers reaching the milestone, track the subsequent performance of two groups, and calculate the average CLV (discounting all future cash flows back to the point at which the treatment was administered) of the two groups in order to identify which treatment worked better.

By CLV, we mean the present value of the future cash flows attributable to the customer relationship over the lifetime of that relationship (see Pfeifer, Haskins, & Conroy, 2005). Thus, the specification of CLV requires a careful delineation of a starting time point so as to be clear about what cash flows are in the future (and included in CLV) and what are in the past (and not included in CLV). For the purposes of this paper, it is important that the CLVs calculated for the sample customer relationships all use a common starting time point in the firm's relationship with the customer. Often that start point is right before the receipt of the initial revenue from the customer. Customer lifetimes (or survival times) are usually measured relative to the time the firm acquired the customer. An implication of the random-sample assumption is that the calendar time of customer acquisition can be ignored. It is also important that all sample CLVs be calculated in a consistent manner across customers and throughout the study period.

Finally, we will assume that CLV is defined using a finite horizon, H. The cash flows over at most H periods will be included in CLV. This means that customers active for more than H periods will be treated as if the relationship ended at H. The finite horizon assumption is one of convenience and statistical necessity.

The construction of an appropriate average CLV depends on the nature of the sample relationships. In the second section, we consider the straightforward case in which the sample consists of n completed relationships. The major contribution of the paper is made in the third section, where we consider samples consisting of a mix of active and completed relationships with some of the sample relationships ending prior to H. The fourth section considers samples containing a mix of active and completed relationships but with none of the relationships ending prior to H. The situation in that section applies to firms in customer-migration situations (in the sense of Dwyer, 1989) in which customers, once acquired, remain customers for the duration. In contrast, the situation in the third section applies to firms in customer-retention situations (in the sense of Dwyer, 1989) in which customers can be lost for good. The paper concludes with a brief summary.

The paper will consider only simple, non-parametric estimation methods. By non-parametric we mean that we will not assume a particular probability distribution (such as the lognormal, gamma, or exponential) for either the customer lifetime or the CLV. Instead, we will explore estimation methods that work regardless of the underlying probability distribution. (These methods are also known as "distribution-free" methods.) By simple, we mean methods that do not use predictive models of any sort. Methods such as Lin's regression analysis (Lin, 2000a), Carides's regression model (Carides, Heyse, & Iglewicz, 2000), and Lin's proportional means model (Lin, 2000b) will not be considered in this paper. The paper will also not consider the use of individual-level stochastic models of customer relationships such as the one used by Venkatesan and Kumar (2004).

SAMPLES CONTAINING ONLY COMPLETE RELATIONSHIPS

The finite-horizon assumption means that if the n sample customers were acquired more than H periods ago, their lifetimes will be complete (by definition), and their CLVs known. In situations such as this (where the n customer relationships have incubated for the full H periods), the calculation of an appropriate average CLV is straightforward. The simple unweighted average of the n sample observed CLVs is the appropriate estimate of the mean CLV. Confidence intervals for the mean can be constructed in the usual way as plus or minus the appropriate number of standard errors, where the standard error is the sample standard deviation calculated from the n sample CLVs divided by the square root of n.

Both Berger, Weinberg, & Hanna (2003) and Reinartz and Kumar (2000) illustrate a "cohort and incubate" approach to estimating mean CLV. A cohort of n customers acquired within a narrow time window is tracked over the subsequent H periods. The resulting n CLVs are all complete and known and the simple unweighted average is the appropriate estimator of mean CLV.

An equivalent way to calculate the unweighted average CLV is to calculate the present value of the total cash flows for the cohort and divide by n. While this aggregate approach may be a simpler way to calculate the average, it makes it difficult to construct a confidence interval.

SAMPLES CONTAINING A MIX OF ACTIVE AND COMPLETE RELATIONSHIPS WITH SOME RELATIONSHIPS ENDING PRIOR TO *H*

In the last section, we considered samples containing only completed relationships. By necessity, those samples will be restricted to relationships initiated more than H periods ago. This consideration of only "older" relationships will limit the usefulness of the calculated average in informing decisions the firm makes today. If the firm wants to include information from "newer" relationships in the calculation of an average CLV, the sample will contain a mix of active and completed relationships.

Both Reichheld and Sasser (1990) and Gupta and Lehman (2003) are examples where an average CLV was calculated using data from both active and completed customer relationships.

The estimation of mean CLV will no longer be straightforward, due to the fact that some of the n observed relationships will be active at the time of the study. (For our purposes, we will refer to the time at which data collection stops and the average CLV is to be calculated as the time of the study.) The n sample relationships will consist of a mix of complete and active relationships. The presence of active customers in the sample complicates the estimation of mean CLV because for the active relationships, the firm can calculate CLV to date but does not know the eventual CLV. In these situations, we say that the customer lifetime is subject to right censoring. A simple averaging of the CLVs to date of the active customers with the complete CLVs for the complete relationships is not appropriate as it will usually underestimate the mean CLV. One of the major purposes of this paper is to show how to correct for this bias.

To meet the challenges posed by right censoring requires some assumptions about the censoring process. In this paper, we will assume the censoring occurs completely at random. This means that the censoring time for a customer (the difference between the time of the study and the time of acquisition) is independent of both the customer lifetime and eventual CLV. One practical implication of the random censoring assumption is that "newer" relationships (those with short censoring times) must be no better or worse than "older" relationships. (Although this may appear restrictive, it is an assumption implicit in the calculation of an average for the entire sample.)

In this section, we also assume that some of the sample relationships ended at H and others ended prior to H. This will be the case, for example, if the firm is in a "customer retention" situation as defined by Dwyer (1989). In customer-retention situations, customers who do not renew are considered "lost for good." In our context, this is equivalent to some relationships ending (being complete) at lifetimes shorter than H. We also assume that the end of the customer relationship is observed and recorded by the firm. For a brief discussion of the additional measurement challenges faced by firms uncertain of the status of their customer relationships, see Mulhern (1999).

The problem of estimating CLV when the sample contains active relationships is very similar to the problem of estimating the medical costs associated with treatments designed to improve the health and survival times of patients. The similarities will allow us to take advantage of a host of recent research on the problem of estimating mean medical costs in the presence of censoring. For example, Young and Buxton (2004) compare several methods for estimating mean total medical costs in the presence of censoring. Our paper will illustrate the appropriate use of these methods towards estimating mean CLV.

The remainder of this section is organized around the types of data available for use in estimating mean CLV when the sample contains a mix of active and complete relationships. We begin by considering situations in which CLV and censored lifetimes are available. Next,we consider situations in which CLV and censoring times are available, and then situations in which CLV histories and censored lifetimes are available. Finally, we present the results of a simulation illustrating the statistical properties of the estimators presented in this section.

Estimating Mean CLV Using CLV-To-Date and Censored Lifetime Data

Suppose the firm knows $(CLV_i, X_i, \Delta_i; i = 1, 2, \ldots n)$ for each of n randomly sampled customer relationships, where CLV_i is the CLV-to-date for customer i, X_i is the lifetime to date of the firm's relationship with customer i, and Δ_i is an indicator variable reflecting whether or not the firm's relationship with customer i is still ongoing. To be more specific, if customer i's relationship with the firm has ended (because the customer declined to renew her subscription, for example), $\Delta_i = 1$ and CLV_i is the complete CLV for the customer and X_i is the complete (or uncensored) lifetime for the customer. If customer i is still active at the time of the study, $\Delta_i = 0$ and CLV_i is the incomplete CLV-to-date and X_i is the observed lifetime to date of customer i. If $\Delta_i = 0$, we say that the customer's eventual lifetime has been right-censored at X_i. CLVs are measured in dollars, and lifetimes are measured in units of time.

The numerical example of Table 1 will be used to illustrate how to use right-censored lifetime data to estimate mean CLV.

The 30 cases in Table 1 are a random sample of customers of a firm providing services on a subscription basis. The time unit is months, and the firm uses a three-year horizon ($H = 36$). Consequently, the two sample customers with lifetimes-to-date of 36 months are considered complete ($\Delta_i = 1$), even though they may still be active at the time of the study.

TABLE 1 Status ($\Delta = 0$ for active at time of study, $\Delta = 1$ for complete), Lifetimes (months), Cash Flows (dollars per month), Customer Lifetime Values, and Replaced CLVs for 30 Example Customers

i	Δ_i	X_i	CF_i	CLV_i	$CLV_i^{replaced}$
1	0	2	$20.99	$41.67	$430.74
2	1	2	$16.99	$33.73	$33.73
3	0	4	$19.99	$78.97	$444.92
4	1	4	$24.99	$98.72	$98.72
5	0	5	$19.00	$93.58	$458.24
6	1	6	$18.99	$111.96	$111.96
7	0	7	$19.99	$137.16	$472.67
8	0	7	$24.99	$171.47	$472.67
9	1	7	$18.99	$130.30	$130.30
10	0	8	$23.00	$179.91	$488.97
11	0	8	$24.99	$195.47	$488.97
12	0	9	$24.99	$219.36	$488.97
13	0	10	$19.95	$194.10	$488.97
14	0	10	$19.99	$194.48	$488.97
15	1	10	$24.99	$243.13	$243.13
16	0	11	$19.99	$213.40	$505.36
17	1	11	$24.99	$266.78	$266.78
18	1	11	$24.99	$266.78	$266.78
19	1	13	$26.95	$338.33	$338.33
20	0	15	$23.99	$345.79	$563.92
21	1	15	$18.99	$273.72	$273.72
22	0	16	$21.50	$329.74	$596.17
23	0	18	$20.99	$360.37	$596.17
24	0	20	$24.99	$474.37	$596.17
25	0	21	$25.99	$516.75	$596.17
26	0	23	$25.99	$563.19	$596.17
27	0	26	$21.50	$522.79	$596.17
28	1	26	$20.99	$510.39	$510.39
29	1	36	$15.95	$524.06	$524.06
30	1	36	$22.95	$754.05	$754.05
Average	0.40	13.23	$22.12	$282.51	$430.74

The example customers pay a fee (at the end of each month) that varies across customers but not across time for a given customer. The monthly cash flow reported in the fourth column of the table accounts for the monthly fee and a small monthly variable cost but does not include acquisition spending, installation costs, or company fixed costs. The CLV reported in column 5 is the present value (at a monthly discount ratio of 0.995) of the monthly cash flows over X months. Thus, for example, the CLV of customer 2 is calculated as $0.995 \times \$16.99 + 0.995^2 \times \$16.99 = \$33.73$.

The firm's relationship with customer 2 is complete ($\Delta_i = 1$), which means the firm will receive no additional cash flows from this customer. The CLV for customer 2 is complete at the $33.73 value as the lifetime for customer 2 is complete at two months.

In contrast, the firm's relationship with customer 1 is active ($\Delta_i = 0$) at the two-month point. For our example, this means the firm will receive at least two monthly payments and the customer's eventual lifetime will be at least two months. It also implies that the firm's relationship with customer 2 could end at $X = 2$ if the customer decides not to renew for the third month. Our convention for handling this customer is to say her lifetime is right-censored at $X = 2$. What is important to remember going forward with this example is that a customer lifetime censored at X is at risk of terminating at X. Said another way, the eventual lifetime of an active customer will be some number greater than or equal to X, the right-censored lifetime to date.

We order the 30 customers in this example based on X from lowest (customer 1) to highest (customer 30). In the case of ties (i.e., multiple cases with the same X value), we put the censored cases first. Censored cases preceding complete cases is consistent with the notion that relationships censored at X are at risk of ending at X in our example. (If the situation had been such that relationships censored at X were not at risk of ending at X, one would put the complete cases first in the case of ties.)

The firm's relationships with 12 of the example customers are complete at the time of the study ($\Delta_i = 1$). Consequently, the lifetimes and CLVs for these 12 customers are known. In contrast, the example firm's relationships with the other 18 customers are still active. The lifetimes for these 18 active customers are right-censored at the values given in Table 1, and their eventual CLVs are uncertain.

Before we introduce a correct way to use the example data in Table 1 to estimate mean CLV, let us illustrate a pair of incorrect methods. One biased method would be to simply ignore the censoring status of the relationships and average the entire available sample. For the numbers in Table 1, this available sample method gives an estimate of ($41.67 + $33.73 + · · · + $754.05)/30 = $282.51. Clearly, this method underestimates the mean. The actual CLVs of the active customers will be greater than or equal to their observed CLVs to date, and using their CLVs to date in our average produces an average that is too low.

Another naive method would be to ignore the censored relationships and average only the complete cases. This produces an average of ($33.76 + $98.72 + · · · + $754.05)/12 = $295.99. This complete cases method also underestimates the mean because it includes only those relationships that are complete at the time of the study. Completed relationships tend to have lower than average CLVs (in situations where CLV increases over time).

There are at least two equivalent methods that yield a consistent (i.e., asymptotically unbiased) estimate of mean CLV under right censorship. We find the first to be very intuitive and easy to explain. The advantages of the second are that it is well established in the field of survival analysis in biostatistics and it facilitates the calculation of the variance of the estimated mean. A variance estimate will allow the construction of confidence intervals for the mean CLV.

We call the first method replace from the right (RR). Starting from the right (the last row of Table 1), the customer with the largest lifetime is customer 30. Since the CLV for customer 30 is known, we can use the $754.05 directly in our calculated average. The same is true of customers 29 and 28. Their CLVs are known to be $524.06 and $510.39, and we can use these numbers directly in our RR average.

Moving to customer 27, we note that the eventual lifetime for customer 27 will be some time greater than or equal to 26 months. To estimate the eventual CLV for customer 27, the RR method uses the average CLV for customers 28 through 30—customers in the sample with lifetimes greater than or equal to 26. The replaced CLV value for customer 27 becomes ($510.39 + $524.06 + $754.05)/3 = $596.17. The logic here is that the average CLV of the three sample customers with lifetimes greater than or equal to 26 is a good estimate of the eventual CLV for customer 27.

Continuing to the left, we see that the lifetime of customer 26 is right censored at 23 months. The RR estimate of the eventual CLV for customer 26 is the average of the CLVs (uncensored or replaced) for the customers "to the right" of customer 26. That average is ($596.17 + $510.39 + $524.06 + $754.05)/4 = $596.17. Notice that the $596.17 number in this average is the replaced value for customer 27. Thus, the RR algorithm is recursive in that later replacement values are calculated using earlier replaced values. Notice also that the replaced value for customer 26 is identical to the replaced value for customer 27. Consecutive censored cases always receive the same replaced value in the RR method.

The algorithm continues moving to the left "replacing" uncensored CLVs with the actual CLV and replacing censored CLVs with the average of "downstream" replaced CLVs. The RR estimate of the mean CLV is then the average of all n replaced CLVs:

$$\hat{\mu}_{RR} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CLV}_i^{\mathrm{replaced}} \qquad [1]$$

where

$$\mathrm{CLV}_i^{\mathrm{replaced}} = \Delta_i \mathrm{CLV}_i + (1 - \Delta_i)\frac{\sum_{j=i+1}^{n} \mathrm{CLV}_j^{\mathrm{replaced}}}{n - i} \qquad [2]$$

and it is understood that the cases are sorted in ascending order based on lifetimes. Cases tied at a common value of X are sorted in ascending order based on Δ if relationships censored at X are at risk of ending at X and in descending order based on Δ if relationships censored at X are not at risk of ending at X.

Table 1 shows the replaced values for the example customers. The average of these 30 replaced values is \$430.74, an asymptotically unbiased estimate of the mean CLV.

As mentioned earlier, the RR method is easy to understand and implement. Essentially, it "corrects" the CLV of each right-censored relationship by replacing it with the average of the replaced CLVs of the downstream relationships. This method is non-parametric.

The RR estimate as calculated using equations 1 and 2 can be shown to be equivalent to a weighted complete case (WCC) estimate calculated as follows:

$$\hat{\mu}_{\mathrm{WCC}} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \frac{\mathrm{CLV}_i}{K_i} \qquad [3]$$

where

$$K_i = \prod_{j=1}^{i}\left(1 - \frac{[1 - \Delta_j]}{n + 1 - j}\right). \qquad [4]$$

This estimator is called the weighted complete case estimator because it uses only the complete case CLVs (those with $\Delta_i = 1$) with each case properly weighted. (Because the RR estimator replaces each censored case with an average of downstream replaced values, it too "ignores" the CLVs of censored cases and uses only the complete case CLVs. In other words, the numbers in the next to last column of Table 1 for the censored cases do not affect any of the numbers calculated in the last column.)

While the RR estimator might be easier to understand and implement in a spreadsheet, the WCC estimator has the advantage of showing in a mathematically convenient form how the complete-case CLVs are combined to compute the estimate. In the absence of censoring, these estimators reduce to the sample average.

A formal proof of the equivalency between RR and WCC might use the self-consistency algorithm of Efron (1967) together with an algebraic argument similar to the one found in Huang and Louis (1998) and Bang and Tsiatis (2000, Appendix 3). This will not be attempted here. In the remainder of this section, we will refer to the WCC estimator with the understanding that everything we say also applies to the RR estimator.

The WCC estimator represented by equations 3 and 4 is almost identical to the "simple weighted estimator" studied by Bang and Tsiatis (2000). If there are no ties, the two estimators are equivalent. But where the Bang and Tsiatis version of the estimator treats ties as "truly" tied, the WCC estimator reacts to the ordering of the tied cases. As discussed earlier, the ordering of the censored and complete cases for relationships with identical X values matters when using the WCC estimator. If the censored cases come first, the WCC estimator considers the censored cases as at risk of ending at X. If the complete cases come first, the WCC estimator considers the censored cases to be not at risk of ending at X. Thus, the WCC estimator has the flexibility to handle either situation. By convention, however, the estimator studied by Bang and Tsiatis considers the censored cases at risk of ending at X regardless of how the cases are sorted.

It may be instructive to point out that, with WCC, the X values are important only to the extent that they are used to sort the cases. Once the cases have been sorted, the magnitudes of the Xs are no longer relevant to the implementation of the WCC estimator. This highlights the importance of getting the sorting correct for tied cases when using WCC.

From Bang and Tsiatis (2000), we know that the WCC estimator is consistent and asymptotically normal with variance consistently estimated by

$$\hat{V}(\hat{\mu}_{\mathrm{WCC}}) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(\mathrm{CLV}_i - \hat{\mu}_{\mathrm{WCC}})^2}{K_i} + \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\Delta_i)}{K_i^2}\{G_i(\mathrm{CLV}^2) - G_i^2(\mathrm{CLV})\}\right] \qquad [5]$$

where

$$G_i(\mathrm{CLV}) = \frac{K_i}{(n - i + \Delta_i)}\sum_{j=i}^{n}\frac{\Delta_j\,\mathrm{CLV}_j}{K_j}.$$

For our 30 example customers, $\hat{V}(\hat{\mu}_{\mathrm{WCC}}) = 3{,}455.83$. This number will be useful in the construction of confidence intervals for the mean CLV.

Instead of applying WCC or RR to the data sorted based on lifetimes, one might consider applying these methods to the data sorted based on CLVs. In situations where CLV increases over time, it is sensible to think of the CLVs themselves as being right-censored. Then why not simply ignore the lifetimes and apply WCC/RR directly to the right-censored CLVs?

It turns out that the naive WCC/RR applied to sorted, censored CLVs will (in general) be biased even though the censoring process is independent of a customer's lifetime (Lin, Feuer, Etzioni, & Wax, 1997). The reason for this is subtle. Customers who accumulate CLV at a higher rate are likely to have larger CLVs at both the survival time and censoring time (Etzioni, Feuer, Sullivan, Lin, Hu, & Ramsey, 1999; Lin, 2003). In this way, the censoring process becomes informative with respect to CLV even though it is not informative with respect to lifetimes. Uninformative censoring is a key assumption for the standard survival analysis techniques to be valid. This finding helped motivate the recent developments in the analysis of censored medical cost data referenced throughout this paper.

Estimating Mean CLV Using CLV-To-Date and Time of Censoring

In most marketing situations, the firm will know the date at which each customer relationship started. Indeed, it is this birth date (or acquisition date) that defines time zero in the construction of CLV. The birth date of a relationship also defines the censoring time for that relationship in that the complete customer lifetime will be observed only if the relationship ends sometime between the birth date and the time of the study—otherwise the customer lifetime is censored. The gap between the time of the study and a customer birth date is thus the censoring time for the customer. Customers acquired a long time ago have long censoring times. Customers acquired more recently have short censoring times. There is a censoring time associated with each customer regardless of whether the relationship ends up being censored.

If the firm knows the birth dates of its customer relationships, then it also knows the censoring times of the n sample relationships. With this knowledge comes the opportunity to estimate mean CLV using a simple unweighted average CLV of the relationships with censoring times greater than or equal to H. In other words, the firm can simply average the CLVs of the incubated cases (the customers acquired more than H periods ago). None of these customer lifetimes will be censored, and the resulting average of the incubated cases (IC) is an unbiased estimate of the mean CLV. The variance of this estimate is the familiar s^2/n', where n' is the number of sample relationships with censoring times greater than or equal to H.

The IC method "solves" the censoring problem by ignoring the information in all the cases that had the potential to be censored. Although this estimator is unbiased, it is inefficient because it uses only n' of the available n data points. In addition, it is particularly vulnerable to departures from the random-sample assumption. Perhaps more importantly, if the quality of relationships does change over time, this method suffers because it uses only the oldest cases.

Estimating Mean CLV Using CLV Histories and Censored Lifetime Data

We now consider situations in which the firm knows not only the CLVs to date, but also the CLV histories of the n customer relationships. With the availability of CLV histories comes the opportunity to improve upon the estimators presented in *Estimating Mean CLV Using CLV-To-Date and Censored Lifetime Data*.

A "weakness" of the WCC estimator is that it completely ignores the CLVs-to-date of the censored cases. The weighted available sample (WAS) method achieves improved efficiency (i.e., decreased variance of the estimate) by adjusting the WCC based on comparing the CLVs-to-date of the censored cases to the historical average CLV (at the same time point) taken over all cases with equal or longer lifetimes. Thus, if the censored cases have particularly high (low) CLVs-to-date, the WAS estimate will be higher (lower) than the WCC estimate

In equation form, the WAS estimator is as follows:

$$\hat{\mu}_{\text{WAS}} = \frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i \frac{\text{CLV}_i}{K_i} + (1-\Delta_i)\frac{(\text{CLV}_i - \text{CLV}_i^*)}{K_i}\right] \quad [6]$$

where CLV_i^* is the average CLV at X_i of cases i through n and it is understood that the cases have been sorted in ascending order based on X_i. An equation for CLV_i^* is

$$\text{CLV}_i^* = \frac{1}{n+1-i}\sum_{j=i}^{n}\text{CLV}_j\,(X_i)$$

where $\text{CLV}_j(X)$ is the historical CLV observed at time X of customer j.

The calculation of $\hat{\mu}_{\text{WAS}}$ for the example customers is illustrated in Table 2. The sixth column lists the K_i values calculated using equation 4. The seventh column lists the CLV_i^* values for the censored cases. In general, the calculation of a CLV_i^* value will involve querying

TABLE 2 The Calculation of the WAS Estimator for the Example Customers

i	Δ_i	X_i	CF_i	CLV_i	K_i	CLV_i^*	$\Delta_i\frac{\text{CLV}_i}{K_i} + (1-\Delta_i)\frac{(\text{CLV}_i - \text{CLV}_i^*)}{K_i}$
1	0	2	$20.99	$41.67	0.9667	$43.91	−$2.32
2	1	2	$16.99	$33.73	0.9667		$34.89
3	0	4	$19.99	$78.97	0.9321	$88.26	−$9.97
4	1	4	$24.99	$98.72	0.9321		$105.90
5	0	5	$19.00	$93.58	0.8963	$109.99	−$18.31
6	1	6	$18.99	$111.96	0.8963		$124.92
7	0	7	$19.99	$137.16	0.8589	$155.13	−$20.93
8	0	7	$24.99	$171.47	0.8216	$155.91	$18.93
9	1	7	$18.99	$130.30	0.8216		$158.59
10	0	8	$23.00	$179.91	0.7825	$178.29	$2.07
11	0	8	$24.99	$195.47	0.7434	$178.21	$23.22
12	0	9	$24.99	$219.36	0.7042	$198.97	$28.96
13	0	10	$19.95	$194.10	0.6651	$219.27	−$37.85
14	0	10	$19.99	$194.48	0.6260	$220.75	−$41.96
15	1	10	$24.99	$243.13	0.6260		$388.40
16	0	11	$19.99	$213.40	0.5842	$242.51	−$49.82
17	1	11	$24.99	$266.78	0.5842		$456.62
18	1	11	$24.99	$266.78	0.5842		$456.62
19	1	13	$26.95	$338.33	0.5842		$579.08
20	0	15	$23.99	$345.79	0.5311	$319.50	$49.49
21	1	15	$18.99	$273.72	0.5311		$515.34
22	0	16	$21.50	$329.74	0.4721	$342.26	−$26.53
23	0	18	$20.99	$360.37	0.4131	$384.90	−$59.38
24	0	20	$24.99	$474.37	0.3541	$429.44	$126.90
25	0	21	$25.99	$516.75	0.2951	$441.96	$253.47
26	0	23	$25.99	$563.19	0.2361	$465.37	$414.37
27	0	26	$21.50	$522.79	0.1770	$494.77	$158.29
28	1	26	$20.99	$510.39	0.1770		$2,882.84
29	1	36	$15.95	$524.06	0.1770		$2,960.02
30	1	36	$22.95	$754.05	0.1770		$4,259.09
Average	0.40	13.23	$22.12	$282.51			$457.70

the CLV histories of cases i through n, recording the $\text{CLV}_j(X_i)$ values, and averaging them together.

The calculation of CLV_i^* for the example situation was made easier by the fact that cash flows are constant across time for each customer. Thus, a customer's historical CLV-to-date is a known function of the customer cash flow (CF) and the number of periods:

$$\text{CLV}_j(X) = \frac{\beta}{1-\beta}(1-\beta^X) \times CF_j$$

where $\frac{\beta}{1-\beta}(1-\beta^X)$ is the present value (at discount ratio β) of X equal unit cash flows.

The last column of Table 2 gives the weighted CLVs for complete cases and the weighted adjustments for censored cases. The average of these 30 numbers, \$457.70, is the WAS estimate of the mean. The WAS estimate is about \$27 higher than the WCC estimate, primarily because the historical CLVs-to-date of censored cases 24 through 27 were higher than average. The WAS estimator adjusts for this, whereas the WCC estimator did not.

The WAS estimator is a case-based version of the "improved" estimator suggested by Zhao and Tian (2001). If there are no ties, $\hat{\mu}_{\text{WAS}}$ and the Zhao and Tian improved estimator are equivalent. If there are ties, one can apply a tie-breaking technique of adding and subtracting epsilons to/from the X values in such a way so as to make the Zhao and Tian improved estimator applied to the epsilon-adjusted Xs equivalent to WAS. Under this equivalency, the properties of the Zhao and Tian improved estimator are shared by $\hat{\mu}_{\text{WAS}}$, and we can say that $\hat{\mu}_{\text{WAS}}$ is consistent and asymptotically normal with the variance estimator presented in equation 7. Equation 7 is the variance equation originally given by Zhao and Tian (2001) (using counting-process integrals) expressed using the convenient case-based notation of this paper.

$$\begin{aligned} Var(\hat{\mu}_{\text{WAS}}) &= Var(\hat{\mu}_{\text{WCC}}) \\ &\quad - \frac{2}{n^2}\sum_{i=1}^{n}\frac{(1-\Delta_i)}{(n+1-i)K_i}\sum_{j=i}^{n}\frac{\Delta_j}{K_j}[\text{CLV}_j \\ &\quad - G_i(\text{CLV})][\text{CLV}_j(X_i) - \text{CLV}_i^*] \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\frac{(1-\Delta_i)}{(n+1-i)K_i^2}\sum_{j=i}^{n}[\text{CLV}_j(X_i) - \text{CLV}_i^*]^2 \end{aligned} \quad [7]$$

From equation 7, Zhao and Tian (2001, Appendix A) argued that the WAS estimator will have a lower variance than the WCC estimator whenever partial CLV histories are highly positively correlated with eventual CLV (i.e., whenever CLV-to-date is a leading indictor of the eventual CLV). Since this will usually be true for most firms, WAS will have lower variance than WCC. This is born out by our example, where the variance of $\hat{\mu}_{\text{WAS}}$ is estimated using equation 7 to be 3,178.32—a number lower than the 3,455.83 variance calculated previously for the WCC estimator. In terms of standard errors, the extra computational effort of the WAS estimator lowers the standard error from \$58.8 to \$56.4.

We look upon the estimators presented so far as case-based estimators. By that, we mean their basic approach to estimating the mean is to calculate a sample average "CLV" over the n cases. For the RR estimator, the CLVs of the censored cases are first replaced with downstream averages before taking the overall average. For the WCC estimator, only the complete cases are averaged but using weights determined by the sample-censoring pattern. For the WAS estimator, the complete cases and adjustments for the censored cases are weighted, then averaged.

We expect that interactive marketers will be quite comfortable with these case-based estimators. After all, interactive marketers are famous for their customer databases that allow them to track the profitability and lifetime value of each of their customer relationships.

In situations where a customer database is not available and/or the firm does not track CLV for each individual, the firm might attempt to estimate the mean CLV using a more traditional time-interval-based approach. A well-known example of this approach is the Reichheld and Sasser (1990) study in which they combined average cash flows by year of customer lifetime with average defection rates by year of customer lifetime into an overall average net present value per customer. Although the Reichheld and Sasser study is a good example of a time-based-approach, their paper makes no mention of how active (censored) customers were treated in the calculation of annual average profits and annual defection rates.

The final estimator we present uses a time-interval-based approach originally proposed by Lin, Feuer, Etzioni, & Wax (1997) and specifically addresses the treatment of censored cases. The weighted partition averages (WPA) method begins by dividing the entire time period $(0, H)$ into K subintervals, or partitions. Although the partitions are usually of equal lengths for administrative or operational convenience, this is not a requirement. Next, the WPA method calculates the average CLV accumulated within each of the K partitions for cases active at the start of the partition that were not censored mid-partition. (Relationships that ended within the partition are included in the average as are relationships that were censored at the end of the partition.) Lin, Feuer, Etzioni, & Wax (1997) explored the estimator with and without observations censored mid-partition and reported that the one excluding cases censored mid-partition performs better in most practical situations. Finally, these K partition average CLVs are multiplied by an estimate of the probability a relationship will survive to start each partition and summed. Of course, when calculating the survival probability estimates, we will account for censoring.

Let the K partitions be $(a_1, a_2]$ through $(a_K, a_{K+1}]$, where $a_1 = 0$ and $a_{K+1} = H$, and let $\overline{\text{PCLV}}_k$ represent the average partition CLV for cases active at a_k that were not censored mid-partition. By partition CLV, we mean the present value of the cash flows received from a customer during the kth partition, which is also the difference between the CLV-to-date at a_{k+1} and the CLV-to-date at a_k,

$$\text{PCLV}_{i,k} = \text{CLV}_i(a_{k+1}) - \text{CLV}_i(a_k).$$

If $\hat{S}(t)$ is our estimate of the probability a relationship will survive to time t, our WPA estimate of the mean CLV becomes

$$\hat{\mu}_{\text{WPA}} = \sum_{k=1}^{K} \hat{S}(a_k) \times \overline{\text{PCLV}}_k. \qquad [8]$$

What remains to be explained is the procedure for constructing $\hat{S}(t)$, the estimated probability a case will survive to at least time t.

This $\hat{S}(t)$ function, called the sample survivor function, is usually constructed using the Kaplan-Meier method (Kaplan & Meier, 1958). Recently, Drye, Wetherill, and Pinnock (2001) demonstrated the use of this function to model the distribution of customer lifetimes.

Developing the equation for $\hat{S}(t)$ requires additional notation. Let $t_1 < t_2 < \ldots < t_D$ be the (ordered, distinct) sample survival times. (For the example customers, there are $D = 10$ distinct sample survival times of 2, 4, 6, 7, 10, 11, 13, 15, 26, and 36.) Let d_j be the number of sample relationships ending at t_j, and let Y_j be the number of sample relationships at risk of ending at t_j. (Since sample relationships ended or censored prior to t_j are not at risk of ending at t_j in the sample, they are not counted in Y_j.) The Kaplan-Meier estimator for the survivor function can now be written as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \Pi_{t_j \le t}\left[1 - \dfrac{d_j}{Y_j}\right] & \text{if } t \ge t_1. \end{cases}$$

Interactive marketers might want to interpret this as the sample cumulative retention rate, where the sample retention rate at each t_j is calculated based only on the Y_j customers whose retention is at risk in the sample. Table 3 shows the construction of the sample survivor function for the example customers. The resulting $\hat{S}(t)$ function is a step function that steps down at each of the D sample survival times. The magnitude of each step is a function of the ratio of the number of relationships ending to the number at risk of ending, where relationships censored or terminated at an earlier time are not considered at risk of ending.

For the purposes of illustration and comparison to Reichheld and Sasser (1990), we will apply this partition method to the example customer data using annual partitions. We use annual partitions with full knowledge that the resulting estimator will be biased and less efficient than if we use monthly partitions. After illustrating the calculations, we will discuss the properties of the partition estimator in more detail.

Table 4 gives the results of the calculations necessary to apply equation 8. The three $\hat{S}(t)$ values are taken directly from Table 3. The three average PCLV values are calculated by first separating each CLV_i into three annual components and then calculating an average for each year using only cases that were active at the start of the year and did not get censored mid-year. There were 19 such cases for year 1, six such cases for year 2 (cases 19, 21, 27, 28, 29, and 30) and three such cases for year 3 (cases 28, 29, and 30). Since relationship 28 ended two months into the third year, its PCLV value for year 3 was only $(0.995^{25} + 0.995^{26})(\$20.99) =$

TABLE 3	The Construction of the Kaplan-Meier Sample Survival Function for the Example Lifetime Data

j	MONTH t_j	NUMBER OF RELATIONSHIPS ENDING d_j	NUMBER AT RISK OF ENDING Y_j	$\hat{S}(t_j)$
1	2	1	29	$[1 - 1/29] = 0.9655$
2	4	1	27	$[0.9655][1 - 1/27] = 0.9298$
3	6	1	25	$[0.9298][1 - 1/25] = 0.8926$
4	7	1	22	$[0.8926][1 - 1/22] = 0.8520$
5	10	1	16	$[0.8520][1 - 1/16] = 0.7987$
6	11	2	14	$[0.7987][1 - 2/14] = 0.6846$
7	13	1	12	$[0.6846][1 - 1/12] = 0.6276$
8	15	1	10	$[0.6276][1 - 1/10] = 0.5648$
9	26	1	3	$[0.5648][1 - 1/3] = 0.3766$
10	36	2	2	$[0.3766][1 - 2/2] = 0.0000$

TABLE 4	The Construction of the WPA Estimate for the Example Customers Using Annual Partitions

k	a_k	a_{k+1}	$\hat{S}(a_k)$	$\overline{\text{PCLV}_k}$
1	0	12	1.0000	\$226.16
2	12	24	0.6846	\$161.45
3	24	36	0.5648	\$145.88

\$36.94, whereas the PCLV values for year 3 for customers 29 and 30 were \$164.29 and \$236.39, respectively, reflecting the full 12 monthly cash flows the firm received from these last two customers during their third year. The average PCLV for year 3 was calculated as (\$36.94 + \$164.29 + \$236.39)/3 = \$145.88. On average, the firm received \$145.88 of customer present value in year 3 from customers who survived to reach year 3. The estimated probability a randomly selected customer will survive to reach year 3 is 0.5648, calculated based on the survival and censoring histories of all 30 sample customers.

Therefore, the final WPA estimate is calculated as

$$\hat{\mu}_{\text{WPA}} = (1.000)(\$226.16) + (0.6846)(\$161.45) + (0.5648)(\$145.88) = \$419.09.$$

By way of summarizing, the WPA method is biased (in general) because it simply ignores the censored cases when it calculates the average partition CLVs. It does a good job, however, in calculating estimated survival probabilities because it goes to the trouble to account for censoring on a case-by-case basis using the Kaplan-Meier method.

To remove the bias of the WPA method, one has several options, all of which involve additional calculation effort. One approach would be to use RR/WCC to estimate the mean PCLV within each partition (see Bang & Tsiatis, 2000; Jiang & Zhou, 2004). A second approach would be to partition finely enough to eliminate mid-partition censoring. Or one might use the sample censoring times as the boundaries of the partitions, thereby eliminating mid-partition censoring. The first approach, using RR/WCC within each partition, is going to remove the bias but will still ignore the PCLV of the censored cases. The later two approaches will eliminate the bias and not ignore any of the partial

TABLE 5 The Construction of the WPA Estimate for the Example Customers Using Monthly Partitions

k	a_k	a_{k+1}	$\hat{S}(a_k)$	$\overline{\text{PCLV}_k}$	$\hat{S}(a_k) \times \overline{\text{PCLV}_k}$
1	0	1	1.0000	$22.01	$22.01
2	1	2	1.0000	$21.90	$21.90
3	2	3	0.9655	$22.01	$21.25
4	3	4	0.9655	$21.90	$21.14
5	4	5	0.9298	$21.78	$20.25
6	5	6	0.9298	$21.80	$20.27
7	6	7	0.8926	$21.83	$19.48
8	7	8	0.8520	$21.90	$18.66
9	8	9	0.8520	$21.67	$18.46
10	9	10	0.8520	$21.44	$18.26
11	10	11	0.7987	$21.50	$17.17
12	11	12	0.6846	$21.25	$14.55
13	12	13	0.6846	$21.14	$14.47
14	13	14	0.6276	$20.66	$12.97
15	14	15	0.6276	$20.56	$12.90
16	15	16	0.5648	$20.60	$11.63
17	16	17	0.5648	$20.59	$11.63
18	17	18	0.5648	$20.48	$11.57
19	18	19	0.5648	$20.57	$11.62
20	19	20	0.5648	$20.46	$11.56
21	20	21	0.5648	$20.01	$11.30
22	21	22	0.5648	$19.23	$10.86
23	22	23	0.5648	$19.14	$10.81
24	23	24	0.5648	$18.04	$10.19
25	24	25	0.5648	$17.95	$10.14
26	25	26	0.5648	$17.86	$10.09
27	26	27	0.3766	$16.99	$6.40
28	27	28	0.3766	$16.90	$6.36
29	28	29	0.3766	$16.82	$6.33
30	29	30	0.3766	$16.73	$6.30
31	30	31	0.3766	$16.65	$6.27
32	31	32	0.3766	$16.57	$6.24
33	32	33	0.3766	$16.48	$6.21
34	33	34	0.3766	$16.40	$6.18
35	34	35	0.3766	$16.32	$6.15
36	35	36	0.3766	$16.24	$6.11
				Total	$457.70

CLV information. In this way, these last two approaches behave a lot like WAS.

As it turns out, the WPA method with partitions that eliminate mid-partition censoring is equivalent to WAS. The implication for us is that one should never go to the trouble to select partitions in the WPA method to match the sample censoring times. If one were going to do that, it seems simpler to use the equivalent WAS estimator. In situations where it is convenient to select a fine enough partition to eliminate mid-partition censoring, then WPA and WAS will involve a similar amount of effort.

For example, monthly partitions of the example cases are fine enough so as to eliminate mid-partition censoring. The implementation of WPA using monthly partitions is illustrated in Table 5, where the resulting estimate of $457.70 matches exactly the WAS estimate calculated earlier.

TABLE 6 Summary of Estimators and Their Properties

ESTIMATOR	DESCRIPTION	ABBREVIATION	PROPERTIES
Available Sample	The average of all CLVs regardless of censoring status.	AS	Biased downward[1]
Complete Cases	The average of the complete case CLVs.	CC	Biased downward[2]
Incubated Cases	The average of the CLVs of the incubated cases—those acquired *H* or more periods ago.	IC	Unbiased with variance σ^2/n'
Weighted Complete Cases or Replace from the Right	The average after replacing the CLVs of censored cases with the average of downstream replaced values.	WCC or RR	More efficient[3] than IC whenever some sample relationships ended prior to *H*.
Weighted Available Sample	WCC adjusted based on comparing CLV to date of censored cases to the average CLV at that date of all cases.	WAS	More efficient than WCC whenever CLV is highly positively correlated with CLV-to-date.
Weighted Partition Average	Average partition CLVs combined using KM survival probabilities.	WPA	Biased if there is mid-partition censoring. Conjectured to be equivalent to WAS if no mid-partition censoring.

[1] Whenever CLV-to-date increases over time.

[2] Whenever CLV and lifetime are positively correlated and some sample cases end prior to *H*.

[3] When comparing two unbiased estimators, the one with the lower variance is said to be more efficient.

It may be instructive to view the Table 5 calculations as a discrete-time estimate of mean CLV. With lifetimes measured in months and cash flows assumed to occur at the end of each month, it is indeed appropriate (and perhaps simpler) to perform our analysis in discrete time. We see from Table 5 that the WPA estimate reduces to the sum of the products of Kaplan-Meier cumulative retention rates and average present values of cash flows for each of 36 discrete time points. Because this discrete-time estimate is equivalent to $\hat{\mu}_{WAS}$, we now can say that this discrete time estimate will be consistent and asymptotically normal with a variance estimate given by equation 7.

Illustrative Simulations

We performed a series of simulations to illustrate the biases and relative efficiencies of the estimators considered in this paper. Let us emphasize that the purpose of these simulations is not to learn more about the properties of these estimators. These estimators have been well studied (both in theory and using simulation) by others in the field of biostatistics and health economics. Table 6 lists the estimators and their statistical properties. The purpose of the simulations is to demonstrate (using simulated examples) these properties.

Customer Relationships. We consider a customer-retention situation with annual cash flows and annual retention events over at most five periods ($H = 5$). For convenience, we do not include a time-zero initial cash flow in the construction of CLV and assume that cash flows occur at the end of each of the first five years (subject to the customer being retained). We assume that a single set of retention probabilities (p1, p2, p2, p4, p5) applies to all relationships, and use a discount ratio of 0.9 when calculating CLV.

Our simulations consider three separate populations of customer relationships. These populations share a common set of retention probabilities (0.6, 0.8, 0.9, 0.95, 0.95) chosen to reflect relationships in which renewals signal increased "loyalty" but will differ with respect to the annual cash flows. Letting $CF_{i,t}$ be a random variable representing the cash flow to the firm from customer i in period t in the event that customer i is retained in period t, our three populations are constructed as follows:

RANDOM	$CF_{i,t} \sim N(100,40^2)$
TREND	$CF_{i,t} \sim N(77.5237 \times 1.1^t,23.5902^2)$
INDIVIDUAL DIFFERENCES	$CF_i \sim N(100,19.1523^2)$

In the RANDOM population, cash flows for each customer in each period are generated independently from a normal distribution with mean \$100 and standard deviation \$40. The RANDOM population has a mean CLV of \$174.32 and standard deviation of \$174.20.

In the TREND population, expected cash flows grow over time at a rate of 10% per year. Each customer faces an identical trend, and cash flows are distributed randomly about the trend. The "base line" trend of \$77.5237 and the standard deviation of \$23.5902 about that trend were carefully chosen so that both the mean and standard deviation of CLV for the TREND population are identical to the mean and standard deviation for the RANDOM population.

In the INDIVIDUAL DIFFERENCES population, cash flows are constant across time for each customer, but this constant annual cash flow varies across customers following a normal distribution with mean \$100 and standard deviation \$19.1523. Once again, this mean and standard deviation were carefully selected so that the INDIVIDUAL DIFFERENCES population is identical in mean and standard deviation of CLV to the other two.

Censoring. We consider two levels of censoring, heavy and light. With heavy censoring, the sample is divided into equal fifths associated with censoring times of 1+, 2+, 3+, 4+, and 5+. In other words, one-fifth of the customers in each sample were acquired a little more than five years prior to the end of the study, one-fifth a little more then four years prior, . . . and one-fifth a little more than one year prior. Under heavy censoring, one-fifth of the sample will have been incubated.

Light censoring means samples equally divided into three groups with censoring times 3+, 4+, and 5+. With light censoring, exactly one-third of the sample has been incubated. Light censoring means the customers are of "older" vintages and we expect fewer to be active at the end of the study. Heavy censoring means the customers are of "younger" vintages and we expect more to be active at the end of the study.

Estimators. The simulations will demonstrate the performance of four estimators on each of the three populations under both heavy and light censoring when used with a sample of size $n = 300$. We expect the first two estimators, AS and CC, to be biased downward. We expect WCC/RR to show very little bias. For the discrete-time relationships considered here, we conjecture that WAS is equivalent to WPA applied using five annual partitions. Thus, our fourth estimator is WAS/WPA, which theory predicts will show very little bias and be slightly more efficient than WCC/RR. Because the efficiency gains from using WAS relative to WCC depend on the amount of correlation between CLV to date and total CLV, we expect higher gains from using WAS relative to WCC for the INDIVIDUAL DIFFERENCES population. (Because cash flows are constant across time for customers in the INDIVIDUAL DIFFERENCES population, CLV to date should be more positively correlated with CLV for this population than for the other two.)

The performance of a fifth estimator IC will also be considered. We know from theory that this estimator will be unbiased with variance equal to the population variance divided by the sample size. Under heavy sampling, the sample size is 60. Under light censoring, the sample size is 100. Thus, we know that IC will have a mean of \$174.32 and standard deviation of \$22.5 (\$174.2/$\sqrt{60}$) under heavy censoring and \$17.4 (\$174.2/$\sqrt{100}$) under light censoring. Because we know the properties of the IC estimator, there is no reason to include it in the simulation.

Performance Metrics. Our simulations will keep track of two performance metrics, BIAS and RMSE (root mean squared error) defined as follows:

$$\text{BIAS} = \frac{\sum_{i=1}^{N} (\hat{\mu}_i - \mu)}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\mu}_i - \mu)^2}{N}}$$

where the i in these formulas refers to the application of the estimator in each of the N trials of the simulation. We used $N = 10{,}000$ in this study. The μ in these formulas is the population mean of $174.32—the same for all three populations.

BIAS measures how far off the estimator is "on average." We know from theory that both AS and CC will be biased downward—meaning BIAS will be a negative number. We expect the bias to be greater (BIAS will be a more negative number) for the TREND population. The other two simulated estimators are consistent—meaning that expected BIAS will approach zero for large n. The IC estimator is unbiased even for small n and will have an expected BIAS of 0.

The RMSE is an overall measure of estimator "accuracy" in that it accounts for both bias and variance. From theory, the RMSE for IC will be $22.5 and $17.4, respectively, for heavy and light censoring. We expect the RMSE to decrease as we move from IC to WCC/RR to WAS/WPA—in keeping with the theory that predicts efficiency gains by using all the cases (WCC compared to IC) and by accounting for the CLV-to-date of the censored cases in situations in which the CLV-to-date is positively correlated with CLV (WAS compared to WCC). This last idea suggests we will see greater improvement in the RMSE of WAS compared to WCC for the INDIVIDUAL DIFFERENCES population than for the other two. Finally, we expect the RMSEs to all be higher under heavy censoring, but that the relative improvements in RMSE from using the more sophisticated estimators to be greater under heavy censoring.

Table 7 shows the simulation results. For the most part, the results in this table match our predictions. Performance improves as we move from IC to WCC/RR to WAS/WPA. The BIAS of AS and CC is worse for the TREND population. Although performance is always worse under heavy censoring, the

TABLE 7 BIAS and RMSE of Five Estimators of Mean CLV for Three Populations under Heavy and Light Censoring

		LIGHT CENSORING			HEAVY CENSORING		
ESTIMATOR		**RANDOM**	**TREND**	**INDIVIDUAL DIFFERENCES**	**RANDOM**	**TREND**	**INDIVIDUAL DIFFERENCES**
AS	BIAS	−$24.5	−$29.5	−$24.2	−$55.0	−$61.7	−$55.0
	RMSE	$26.0	$30.6	$25.7	$55.5	$62.1	$55.5
CC	BIAS	−$73.6	−$76.0	−$73.3	−$100.0	−$102.3	−$100.1
	RMSE	$74.2	$76.5	$73.9	$100.4	$102.6	$100.5
IC	BIAS		$0			$0	
	RMSE		$17.4			$22.5	
WCC/RR	BIAS	−$0.2	−$0.2	$0.1	$0.0	$0.0	−$0.1
	RMSE	$10.5	$10.3	$10.8	$11.6	$11.0	$11.9
WAS/WPA	BIAS	−$0.2	−$0.2	$0.1	$0.0	$0.0	−$0.1
	RMSE	$10.0	$10.1	$10.2	$10.7	$10.7	$10.7

Note: Results for all but the IC estimator are sample averages based on $N = 10{,}000$ trials. The four estimators were applied to a common set of 10,000 simulated $n = 300$ sample customer relationships for each of the six combinations of population and censoring level. Consequently, comparisons within a column are more reliable than comparisons across columns.

degradation in performance is relatively small for the last two estimators—those that account for censoring in a sophisticated way.

The only prediction not clearly validated by these results is the prediction that the efficiency gains of WAS/WPA relative to WCC/RR would be greater for the INDIVIDUAL DIFFERENCES population. While there is some evidence for this (the heavy-censoring improvement from \$11.9 to \$10.7 is a little bit greater than the improvement from \$11.6 to \$10.7), it is well within the margin of the simulation's sampling error.

SAMPLES CONTAINING A MIX OF ACTIVE AND COMPLETE RELATIONSHIPS WITH NO RELATIONSHIPS ENDING PRIOR TO *H*

We close with comments on the use of the estimation methods of this paper in customer-migration situations (see Dwyer, 1989). In customer migration situations, all customers remain active until H. The sample of n customer-migration relationships can include a mix of active and complete relationships, but with the lifetimes of the active relationships all less than H and the lifetimes of the complete relationships all equal to H. If depicted as in Table 2, the zeros will come first and the ones last.

In this situation, RR, WCC, IC, and CC will be equivalent and reduce to the unweighted average of the CLVs of the complete cases. The AS method will again be biased whenever CLV increases over time. Finally, WAS/WPA will again be more efficient than WCC because it incorporates the partial CLV information from the active relationships.

SUMMARY

This paper addresses the question of how to use data from a random sample of n customer relationships to estimate the mean CLV of the population of relationships from which the random sample was drawn. When the sample contains only completed relationships, a simple unweighted average is appropriate. When some of the n sample relationships are active at the time of the study, we say that the customer lifetime is subject to right censoring. Because of this censoring, the simple unweighted average of the n observed CLVs-to-date will underestimate the mean CLV. This paper describes and compares several estimation methods that correct for this bias.

The simplest method for dealing with censoring, described in *Estimating Mean* CLV *Using* CLV*-to-Date and Time of Censoring*, is to "throw away" all observations that had the opportunity to be censored. In the context of this paper, this means "throwing away" the data from customers acquired less than H periods ago. (Recall that H is the horizon over which CLV is defined.) While the average CLV of the remaining customers (those acquired H or more periods ago) is an unbiased estimate of mean CLV, it is inefficient.

Estimating Mean CLV *Using* CLV*-to-Date and Censored Lifetime Data* describes two equivalent estimation methods (RR and WCC) that are more efficient because they make use of the information available in all the complete cases. The complete cases are those relationships that were acquired more than H periods ago together with the relationships acquired less than H periods ago that have since ended. To the best of our knowledge, our paper is the first to describe RR, a simple recursive algorithm for estimating mean CLV (or any marked variable as defined by Huang & Louis, 1998) in the presence of censoring. We also demonstrate numerically the equivalency between RR and WCC, which allows us to draw on the work of Huang and Louis (1998) and Bang and Tsiatis (2000) to describe the statistical properties of RR/WCC. The case-based WCC is also something of a contribution in that it adds a measure of flexibility for handling ties to the simple weighted complete-case estimator of Bang and Tsiatis (2000). In addition, WCC will be easier to implement than the Bang and Tsiatis estimator because the case-based K_i weights can be calculated directly from the sorted cases using equation 4.

Of note methodologically, the concept of a weighted estimator similar to our WCC first appeared in the survey sampling literature (Horvitz & Thompson, 1952). Estimators similar to WCC and WAS were later introduced to the biostatistics field via the pioneering work on semi-parametric efficiency in missing data by Robins, Rotnitzky, and Zhao (1994).

To take the next step in improved efficiency requires the use of CLV histories. *Estimating Mean* CLV *Using* CLV *Histories and Censored Lifetime Data* describes the WAS estimator. Like RR/WCC, WAS is a case-based version of an estimator studied previously (Zhao & Tian, 2001) in biostatistics. The WAS estimator achieves improved efficiency by adjusting the WCC estimator based on comparing the CLVs-to-date of the censored cases to the historical average CLV (at the same time point) taken over all cases with equal or longer lifetimes. In this way, the WAS estimator makes use of the information available in all the cases. For the 30-case example considered in the paper, the added computational effort of the WAS reduced the standard error of the estimate of the mean from $58.8 to $56.4. This relatively small reduction in standard error suggests that the improved efficiency of WAS may not be worth the extra computational effort required in some situations. Consequently, we expect that RR/WCC will be more widely used especially among those who want a convenient procedure.

The paper also examines the WPA estimator, a time-interval based approach. Although this approach will (in general) be biased, it fits well with traditional time-based accounting practices. From the work of Lin, Feuer, Etzioni, & Wax (1997), we know that WPA will be biased whenever there is mid-partition censoring-in which case it is better not to include the mid-partition censored cases when calculating the partition averages. When mid-partition censoring is avoided (either by using fine-enough partitions or by using the sample censoring times as the partition boundaries), our paper conjectures that WPA is equivalent to WAS. The conclusion here is that WPA and WAS are both ways to improve upon RR/WCC by incorporating information available in the censored cases. The WAS does this directly through equation 6, whereas the WPA does it indirectly through the use of partitions.

REFERENCES

Bang, H., & Tsiatis, A.A. (2000). Estimating Medical Costs with Censored Data. Biometrika, 87, 329–343.

Berger, P.D., Bolton, R.N., Bowman, D., Briggs, E., Eleten, V., Kumar, V., Parsuraman, A., & Terry, C. (2002). Marketing Actions and the Value of Customer Assets: A Framework for Customer Asset Management. Journal of Service Research, 5(1), 39–54.

Berger, P.D., & Nasr, N.I. (1998). Customer Lifetime Value: Marketing Models and Applications. Journal of Interactive Marketing, 12(1), 17–29.

Berger, P.D., Weinberg, B., & Hanna, R.C. (2003). Customer Lifetime Value Determination and Strategic Implications for a Cruise-Ship Company. Journal of Database Marketing & Customer Strategy Management, 11(1), 40–52.

Blattberg, R., & Deighton, J. (1996, July–August). Manage Marketing by the Customer Equity Test. Harvard Business Review, 136–144.

Carides, G.W., Heyse, J.F., & Iglewicz, B. (2000). A Regression-Based Model for Estimating Mean Treatment Cost in the Presence of Right-Censoring. Biostatistics, 1(3), 299–313.

Drye, T., Wetherill, G., & Pinnock, A. (2001). When are Customers in the Market? Applying Survival Analysis to Marketing Challenges. Journal of Targeting, Measurement, and Analysis for Marketing, 10(2), 179–188.

Dwyer, F.R. (1989). Customer Lifetime Valuation to Support Marketing Decision Making. Journal of Direct Marketing, 3(4), 8–15, reprinted in 11(4), 6–13.

Efron, B. (1967). The Two Sample Problem with Censored Data. In Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability (Vol. 4, pp. 831–853). New York: Prentice-Hall.

Etzioni, R.D., Feuer, E.J., Sullivan, S.D., Lin, D., Hu, C., & Ramsey, S.D. (1999). On the Use of Survival Analysis Techniques to Estimate Medical Care Costs. Journal of Health Economics, 18, 365–380.

Gupta, S., & Lehmann, D.R. (2003). Customers as Assets, Journal of Interactive Marketing, 17(1), 9–24.

Horvitz, D.G., & Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Huang, Y., & Louis, T.A. (1998). Nonparametric Estimation of the Joint Distribution of Survival Time and Mark Variables. Biometrika, 85, 785–798.

Hughes, A.M. (1997). Customer Retention: Integrating Lifetime Value into Marketing Strategies. Journal of Database Marketing, 5(2), 171–178.

Jain, D., & Singh, S.S. (2002). Customer Lifetime Value Research in Marketing: A Review and Further Directions. Journal of Interactive Marketing, 16(2), 34–46.

Jiang, H., & Zhou, X. (2004). Bootstrap Confidence Intervals for Medical Costs with Censored Data. Statistics in Medicine, 23, 3365–3376.

Kaplan, E.L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457–481.

Lin, D.Y. (2000a). Linear Regression Analysis of Censored Medical Costs. Biostatistics, 1(1), 35–47.

Lin, D.Y. (2000b). Proportional Means Regression for Censored Medical Costs. Biometrics, 56, 775–778.

Lin, D.Y. (2003). Regression Analysis of Incomplete Medical Cost Data. Statistics in Medicine, 22, 1181–1200.

Lin, D.Y., Feuer, E.J., Etzioni, R., & Wax, Y. (1997). Estimating Medical Costs from Incomplete Follow-Up Data. Biometrics, 53, 419–434.

Mulhern, F.J. (1999). Customer Profitability Analysis: Measurement, Concentration, and Research Directions. Journal of Interactive Marketing, 13(1), 25–40.

Pfeifer, P.E., & Carraway, R.L. (2000). Modeling Customer Relationships Using Markov Chains. Journal of Interactive Marketing, 14(2), 43–55.

Pfeifer, P.E., Haskins, M.R., & Conroy, R.M. (2005). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. Journal of Managerial Issues, 17(1), 11–25.

Reichheld, F.F., & Sasser, W.E., Jr. (1990, September–October). Zero Defections: Quality Comes to Services. Harvard Business Review, 105–111.

Reinartz, W.J., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. Journal of Marketing, 64, 17–35.

Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. Journal of the American Statistical Association, 89, 846–866.

Venkatesan, R., & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing, 68, 106–125.

Young, T., & Buxton, M. (2004). A Comparison of Methods to Adjust for Censored Cost Data Under Different Censoring Mechanisms. Working paper, Health and Economics Group, Brunel University, Uxbridge, UK.

Zhao, H., & Tian, L. (2001). On Estimating Medical Cost and Incremental Cost-Effectiveness Ratios with Censored Data. Biometrics, 57, 1002–1008.