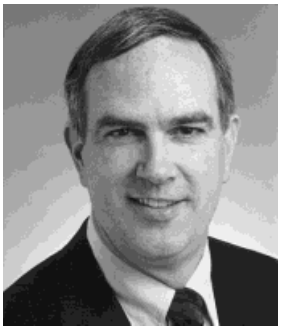
JAMES D. HESS
GLENN E. MAYHEW

Modeling Merchandise Returns in Direct Marketing

JAMES HESS is a professor of business administration at the University of Illinois at Urbana–Champaign. He received a BSE in electrical engineering and an AB in economics from Princeton and a PhD in economics from MIT. His research focuses on analytic models and empirical validation of theories of pricing decisions. GLENN MAYHEW is an assistant professor of marketing at Washington University's John M. Olin School of Business. He received a BA in Japanese from Brigham Young University, a MBA in Marketing from the University of Chicago and a PhD in Marketing from the University of California–Berkeley. His research interests include direct marketing, behavioral aspects of pricing, and the value of marketing.
The research proposal for this study was the winner of the Marketing Science Institute/Direct Marketing Educational Foundation proposal competition, "Challenges and Opportunities for Direct Marketing." We gratefully acknowledge the financial support of MSI and DMEF.



JAMES D. HESS

GLENN E. MAYHEW

ABSTRACT

Returns are a significant problem for many direct marketers. New models to more accurately explain and predict returns, as well as models that will allow accurate scoring of customers and merchandise for return propensity, would be useful in an industry where returns can exceed 20 percent of sales. We offer a split adjusted hazard model as an alternative to simple regression of return times. We explain why the hazard model is robust and offer an example of its estimation using data of actual returns from an apparel direct marketer.

CCC 0892-0591/97/02020-16

1. INTRODUCTION

Direct marketing exposes customers to merchandise through an impersonal medium such as a catalog, brochure, telephone conversation, or television commercial. The inability to inspect tangible merchandise makes the buyer's decision more risky. This is a serious problem for direct marketers in competition with stores, since many customers place great value on browsing through the merchandise they will buy and take home.

To reduce the customers' risk and to effectively compete against stores that have merchandise on display, many direct marketers offer very generous return policies. For example, L. L. Bean is famous for its early introduction of an unconditional, no-questions-asked warranty of its merchandise (Figure 1). Many direct marketers have followed suit. However, there are a wide variety of warranty policies in common use by sellers (3). Many direct marketers will exchange but not refund, charge restocking fees, or impose time limits on returns (7).

The result of liberal return policies can be a flood of returned merchandise. Fenvessy (4) states that returns of 4–25% of sales should be expected by direct marketers. It is reported that return rates at L.L. Bean which had historically been around 5% of sales jumped to 14% in the early 1990s (2: 42). We have been told by a major catalog marketer of women's apparel that over 30 percent of their sold items are returned, and in some merchandise categories the return rate is as high as 70 percent.

If direct marketers could create detailed statistical models of their returns, then they could learn more about this important cost of doing business. Direct marketing companies can collect detailed return data and use it to score customers or products on return propensity. The necessary level of detail in the data for such models, while extremely difficult for traditional retailers to collect, should be no problem for most direct marketers. The necessary data could include time between shipment and return, reason for return, etc. These data could be combined with data on the customer's past purchase history to form a very accurate picture of the return rate for an individual customer, a product category, or even an individual item.

What business decisions would be affected by the insights gleaned from a model of returns? This information could be used to score customers for either general propensity to return or to classify the customer as a quick returner or a late returner. The propensity to return should be a major factor in scoring customers for mailings, etc., because it will have a major effect on the customer's lifetime value. When scoring customers for returns, it is important to have a model that is sophisticated enough to judge subtle differences rather than relying solely on average return rates. Thus, intervention may be appropriate for a customer with a low overall average return rate if that customer buys primarily low-return-rate merchandise but returns it at an unprofitably high rate for that class of merchandise. Likewise, intervention may be inappropriate for the customer who seems to have a high return rate, but is found to primarily purchase merchandise with high return rates. Understanding the pattern of a customer's returns also will help to flag customers who are making particularly late returns or simply to better predict the operational flow of returns.

L. L. Bean Return Policy

Our Guarantee Our products are guaranteed to give 100% satisfaction in every way.
Return anything purchased from us at any time if it proves otherwise.
We will replace it, refund your purchase price or credit your credit card, as you wish.
We do not want you to have anything from L.L. Bean that is not completely satisfactory.

FIGURE 1
L.L. Bean return policy.

The marketer's merchandise can be scored in a way similar to the customer list. Just as customer return scores could be used as a basis for dropping excessive returners from mailing lists, return scores could be used to flag items to be dropped from future merchandising. This is important, since return rates should be taken into account when judging the profitability of each item or category.

A company might also be interested in understanding returns to project operational staffing or procurement demands or to develop operational standards. An early warning system could be developed that would warn of problems with an item in time to adjust ongoing orders from suppliers. Return

scores that exceed an operational standard can be a sign of any number of problems from mailings that do not accurately describe the merchandise to order-taking and picking and packing problems. Likewise, efforts to reduce such problems are hard to gauge without an objective criterion of measurement. A sophisticated return score is helpful here, as a simple long-term average return rate will not be able to match the variation in the customer base and item mix from week to week.

The first step in understanding returns is to find a way of modeling the phenomenon. This paper concentrates on developing a theoretically sound and practically estimable model of direct marketing returns. Such a model is described and then estimated using a small sample from an actual direct marketing database. We feel that this new model of direct marketing returns offers managers a new chance to understand their returns, and that the understanding of returns will lead to great gains in the practice of direct marketing.

2. MODELING RETURNS

Given the importance of understanding returns for customer and merchandise scoring and for operations, it is important to minimize error in return scoring. We propose a statistical approach to modeling returns (hazard rate models) that breaks out the effects of merchandise category, price, etc. to gain a more accurate view of the customer's and item's baseline return probability. Our approach also estimates the probability of return over time, giving the manager a predicted pattern of returns for operational control. The method is somewhat more complex than simple means or regressions, but we feel that the possibilities for savings are great enough to warrant further research by both practitioners and academics. We begin with the familiar regression model before spelling out the hazard rate model.

A. A Split Regression Model of Returns

Two key components of the return phenomenon must be modeled if returns are to be understood: the timing of return and the probability of return. The first return question, *when the return will occur,* may be modeled simply with a historic average time-to-return. This may be a time-to-return for the company as a whole, for the merchandise category, or for the individual item. Obviously, the narrower the scope of merchandise used in the calculation of a historic average time-to-return, the more closely it will match the item. On the other hand, a narrower scope may result in a very limited data set and little predictive power. One may gain more insight by moving from a simple average to a linear regression model that includes factors that may affect the time-to-return.

The time between sale and return may serve as the dependent variable to be regressed on independent explanatory variables that describe characteristics of the item and the customer. This model attempts to explain the variation in return times. It uses only data from items that have been returned, and thus cannot take advantage of the much larger database of nonreturn observations.

This brings up a serious drawback. In either the simple mean approach or regression model, one has the problem that not all of the items that will eventually be returned have already been returned (the data are "censored"). Thus, the return times will be biased downwards. One may attempt to correct for this by using only "old" sales where future returns are very unlikely. This calls the timeliness of the model into question.

If we do nothing, the regression results will be biased if the unusable "not yet returned" data differ from the "already returned" observations in the regression in some systematic way. Whether or not the results are biased, however, they will certainly be statistically inefficient, as available data are being ignored.

A final problem with regression models is that they are often supplemented by an arbitrary assumption of normally distributed random errors. This is inappropriate for modeling the time between sale and return because this variable must be positive. The normal distribution always has a negative tail, so the model is theoretically misspecified.

The second return question, *what is the probability that the product will be returned,* can be modeled by calculating the simple historic return rate. As with time-to-return, this may be a return rate for the company as a whole, for the merchandise category, or for the individual item. A more powerful approach is a discrete choice model, such as a logit or probit model, which can simultaneously estimate a baseline return rate and the influence of the various factors on that baseline rate. However, such a model

still faces the problem that return rate may confuse "not yet returned" with "never will be returned" observations due to data censorization.*

In summary, both the *when will it be returned?* and *will it be returned?* questions cannot be adequately answered with regression models, simple discrete choice models, or a combination of the two. A more unified approach to returns modeling is required.

B. Hazard Models

Another way to understand the timing of events is with hazard models (6,8,9). Hazard models are common in the measurement of reliability, and are often referred to as waiting time or failure time models. The basic idea is that the event of interest (the arrival of the next storm, the failure of a part, etc.) will eventually occur and the timing follows some statistical distribution. The hazard rate is the ratio of the probability that the event will occur in a short interval of time and the probability that it has not happened yet (see Technical Appendix, equation 1). This is a conditional probability: the probability that the event occurs "now" given that it has not occurred "yet." Therefore, it is important to think of hazard rates not as probabilities, but rather as ratios of probabilities. For example, while a probability density function must integrate to one, a hazard function need not. It need only be positive and asymptotically bounded above zero. Every probability distribution has an implied hazard function, and every hazard function has an implied probability distribution.

The hazard function is a pure function of time, but it can be adjusted by other parameters or covariates. In modeling returns, one might be interested in the influence of merchandise category, consumer characteristics, or other special purchase characteristics, such as whether the item was a gift. The adjustment is generally done by defining a baseline hazard that is a function only of time and multiplying by an adjustment factor that is a function of other variables that are thought to influence the timing of the event. Thus, one can gauge the importance of various merchandise or consumer characteristics in terms of their impact on return timing.

A concern in modeling returns is the chance that the event may never occur. Most items are never returned, while some come back after varying lengths of time. Does a nonreturn indicate that the item will not come back or simply that it has not come back yet? As discussed above, to leave out the nonreturn observations introduces inefficiency and possible bias into the model. This problem can be overcome, however, by using all of the observations in a split hazard model.

A split hazard model explains not only the returns, but also the nonreturns. The probability of seeing a return in a particular observation in a data set is the probability that the item would be returned multiplied by the probability that it would have been returned by that point in time. The probability of observing a nonreturn is the sum of two probabilities. The first is the probability that the item never will be returned. The second is the probability that the item is going to be returned multiplied by the probability that it would not have been returned by that time. Thus, all three possible situations are accounted for: the possibility that it will not be returned, the possibility that it has already been returned, and the possibility that it will be returned in the future. Accounting for all possibilities eliminates bias and allows all observations to be used, thus simultaneously eliminating inefficiency in estimation.

Modeling the split between returns and nonreturns also allows the direct marketer to study the impact of merchandise and consumer characteristics on return probability. As discussed above, merchandise or consumer characteristics can be included in an adjusted hazard rate model to gauge their impact on return timing. Including such variables in both the hazard adjustment function and the split function allows one to identify the characteristics that influence the probability or timing of return. Thus, the procedure becomes a valuable source of return "scoring."

C. Choosing a Split Adjusted Hazard Rate Model for Direct Marketing

Given that a split adjusted hazard rate model is the proper model of direct marketing returns, one must

* (Note that one may take a further step of jointly estimating the discrete choice model and time-to-return regression model in a two-step procedure, first estimating the discrete choice model and then including that model's results for the return observations in the regression function. This decreases the bias in the estimation of the regression parameters as correlations between the logit and regression parameters are explicitly modeled. However, this cannot eliminate the bias in either model that is caused by the inability to separate "never will be returned" and "not yet returned" observations. It also does away with one of the main benefits of the regression model, its great simplicity. The joint logit-regression equation is given in the Technical Appendix.)

next choose specific functional forms for the split, adjustment, and hazard rate. We will begin with a choice of functional form for the baseline hazard. One may observe the pattern of returns and then choose a functional form that fits that pattern, or choose a flexible functional form that can take on many different shapes. The first option has the advantage of simplicity. It may be possible to choose a simpler function with fewer parameters to be estimated. This option also has the disadvantage of being less general. As this is exploratory research, the first attempt to apply hazard models to direct marketing returns data, we see the loss of generality as a serious flaw.

We have chosen a functional form that is quadratic in time for nonnegative values that allows for an increasing or decreasing return rate over time, or any form that is first increasing and then decreasing, or vice versa. To guarantee that the hazard is strictly positive, we exponentiate a quadratic equation (see Technical Appendix, equation 2). We expect the parameters from the estimation of the model with return data to define a hazard rate that is bell-shaped, with negative values truncated. The truncation may be such that the hazard at time zero is very small and first rises and then falls over time, such as the first graph in Figure 2. Alternatively, the hazard could start high and then fall, as the second graph in Figure 2.

Next, one must choose a functional form for the hazard adjustment equation. Since hazard rates are ratios of probabilities, negative values are ruled out, so we exponentiate a simple linear function of covariates that describe attributes of the consumer, the item purchased, or the fulfillment process (see Technical Appendix, equation 3).

Finally, one must choose a functional form for the discrete split between returns and nonreturns. Many discrete choice models have been proposed in the marketing and econometrics literature, but the logit model has been by far the most popular because it is theoretically simple and its closed form probability equation lends itself well to maximum likelihood estimation (1). The probability of a return is theorized to depend on a number of covariates such as importance of fit or whether the item is a gift. The return/nonreturn probability does not depend on time. Time affects only the pattern of returns for items that are going to be returned. The logit return probability is based on a linear function of covariates of return. To form a return probability, this function is exponentiated and then divided by one plus the same exponentiated function (see Technical Appendix, equation 7).

3. AN APPLICATION OF THE MODELS TO ACTUAL RETURN DATA

A. Direct Marketing Return Data

To show the usefulness of the hazard model we obtained return data from a large direct marketer of apparel. The database is a small sample from the actual house list, but should be sufficient to estimate a simple returns model. The data are from a period of about four years for a random group of about 1,000 customers from the company's multimillion-name list. This group purchased 2,024 items over the time period ranging in price from a few dollars to about $400 (mean = $60, s.d. = $44), of which 242 items were returned. These data include the order date, return date, price, category of clothing or accessory (pants, shirt, etc.), and a code for the customer's stated reason for return.

The purchases and returns occur at various times, but the data had to be censored at the date they were sent to us. Therefore, while each observation has a purchase date, the lack of a return date does not mean the item never will be returned. The time between purchase and return varies from 2 to 104 days in the 242 return observations. The time from purchase to observation censorization varies from 1 to 1,308 days in the total set of 2,024 observations.

We are interested not only in the explanatory power of the estimated model, but also in its predictive accuracy. Therefore, in addition to estimating the model with the full data set, we will also re-estimate it using only a subset of the data. The complementary subset serves as a holdout sample for use in judging the fit of the model to new data. We will present fit statistics to compare the fit of the regression and hazard models for both the estimation and prediction samples. In creating the holdout sample, we assign a random number to each observation and then divide the data roughly evenly into four groups of observations. We then use each of these four samples in turn as a holdout sample, estimating the model on the remaining three quarters of the data. The results of estimating the model on the full sample and each of the four partial estima-

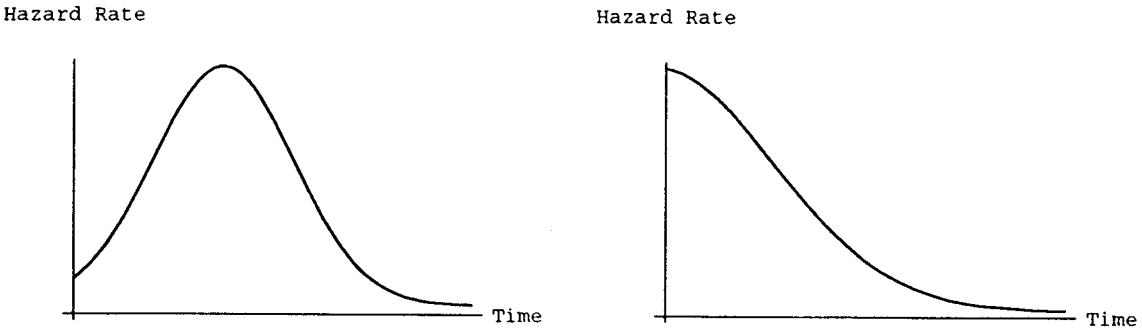


FIGURE 2
Bell-shaped hazard functions truncated at time zero.

tion samples should be similar, since they overlap so heavily. The fit statistics in the holdout samples, however, will not necessarily be similar, since the holdout samples are disjoint.

B. Results: Regression Model of Time-to-Return

We first estimate a regression model of time-to-return. This will serve as a baseline model for comparison to the more theoretically correct hazard model. Time-to-return is the dependent variable. An examination of the variables in the data reveals few that might be indicators of return time. Price is a possible indicator, with the hypothesis that a consumer who is going to return an item would be motivated to act more quickly if a larger amount of money is at stake. Thus, we hypothesize that more expensive items will be returned more quickly and that the price coefficient in a time-to-return regression will be negative. We do not include category of merchandise or reason for return, as we could find no suitable hypotheses for how these variables could affect the timing of return.

The results of the regression are shown in Table 1. The intercept suggests an average baseline time-to-return of 3.799 weeks. Price is measured in hundreds of dollars, and its coefficient is not significantly different from zero, suggesting that price has no systematic effect on returns. Thus, our hypothesis that price would have a negative impact on return time is not supported. The fit of the model is extremely poor (the R^2 of the model is 0.000). Thus, the model explains almost none of the variation in return time. Regression, while simple to execute, reveals no useful information regarding returns except the mean return time, 3.828 weeks. This information is of no help in scoring households or merchandise.

In operational problems such as forecasting returns, the pattern of returns over time is important. The regression model can offer only a bell-shaped normal curve centered at the mean return time. Also, with a normal curve, the mean and mode coincide. Thus, the model's modal return time is 3.828 weeks, while the true mode of the data is at 3.143 weeks. The regression's predicted pattern of returns, a bell-shaped normal curve with a mean and mode of 3.828 weeks and a standard deviation based on the standard error of the regression residuals of 2.625 weeks, is a poor match to the actual shape of the returns in the data, as we will show below.

C. Results: Logit Model of Return Rates

In addition to the regression model of return timing, we estimate a logit model of the rate of return. We include variables in the logit return/nonreturn model to capture the baseline return probability and the impact of price and the general importance of fit for the category. Price is hypothesized to positively affect probability of return. Our logic is that consumers will be less likely to accept a poor fit as the item becomes more expensive. For some items fit is simply less important than for other items (e.g., socks vs. suits) and for some items the fit is almost totally unimportant (scarves or ties). We define dummy variables that describe fit as somewhat important or very important (a zero for both variables

TABLE 1
Time-to-Return[1] Regressions on Return Observations

	Full Data	Sample 1	Sample 2	Sample 3	Sample 4
Constant	3.799	3.489	3.731	4.126	3.843
(Standard error)	(0.325)	(0.370)	(0.366)	(0.389)	(0.377)
Price[2]	0.041	0.258	0.234	−0.411	0.077
(Standard error)	(0.384)	(0.424)	(0.443)	(0.460)	(0.446)
Observations	242	175	189	176	186
R^2	0.000	0.002	0.001	0.005	0.000
Mean fitted return time	3.828	3.681	3.896	3.828	3.899
Standard error of residuals	2.625	2.530	2.654	2.657	2.649

[1] Time is measured in weeks.
[2] Price is measured in hundreds of dollars.

indicates the fit is unimportant). We expect the coefficients for both dummies to be positive and we expect the coefficient for the dummy representing categories where fit is "very important" to be larger than the coefficient for "somewhat important." The logit split and regression models are disjoint and estimated separately. The estimated coefficients of the logit model are shown in Table 2. The price coefficient is positive and significant in each of the estimated models, as we expected. The variables capturing the importance of fit, however, are uniformly insignificant. Thus, more expensive items are more likely to be returned, but differences in the importance of fit across categories have very little impact on the return rate.

The estimation of the logit model also allows us to compare the fits and predictive accuracy of the regression and split adjusted hazard models. Without the logit split, the regression model attempts to explain only the timing of returns, not the question of whether the item will be returned. To compare the models, therefore, we must either throw away the split portion of the hazard model, or add a split to the regression model. We augment the regression model with a logit split rather than handicap the hazard model.

D. Results: Split Adjusted Hazard Model

The baseline hazard function is an exponentiated quadratic function of time (referred to as a Box-Cox hazard function), as explained above. We expect the coefficients of the baseline hazard function to define a function that is increasing and then decreasing in time. The adjustment function is a simple ex-

TABLE 2
Logit Model of Return Rate

	Full Data	Sample 1	Sample 2	Sample 3	Sample 4
Constant—γ_1	−2.486	−2.774	−2.323	−2.599	−2.280
(Standard error)	(0.238)	(0.329)	(0.147)	(0.328)	(0.172)
Price[1]—γ_2	0.597	0.700	0.492	0.638	0.559
(Standard error)	(0.137)	(0.161)	(0.159)	(0.161)	(0.154)
Fit med import—γ_3	0.103	0.326	0.029	0.118	−0.033
(Standard error)	(0.240)	(0.336)	(0.069)	(0.351)	(0.105)
Fit high import—γ_4	0.136	0.262	0.174	0.220	−0.089
(Standard error)	(0.269)	(0.369)	(0.204)	(0.364)	(0.241)

[1] Price is measured in hundreds of dollars.

ponential function of price. We expect price to have a positive coefficient, our prior belief being that more expensive items would be returned more quickly (have a greater hazard of return). Finally, we include variables in the logit return/non-return model to capture the baseline return probability and the impact of price and the general importance of fit for the category. The logit model has the same variables as the logit model described above.†

The maximum likelihood results of the hazard model are shown in Table 3. First, let us examine the baseline hazard coefficients. All four coefficients are highly significant and they define a curve that rises from zero, peaks at 6.286 weeks, and then falls to approximately zero (0.01) at 14.501 weeks. It must be remembered that the time of peak hazard and the time of peak returns should not be expected to coincide: hazard is not the probability distribution of time-to-return. The time of peak returns is the mode or peak of the probability density of returns, not the peak of the hazard function. As explained above, however, a density function is implicit in each hazard function. The modal return time of 3.012 weeks that is implied by the estimated baseline hazard function is much closer to the true modal return time of 3.143 weeks than the regression model estimate of 3.828 weeks.

† Note that we do not show the estimated coefficient α_4 (see equations (2) and (3) in the Technical Appendix). The exponent of this coefficient defines a baseline hazard rate that is independent of time. The estimated coefficient of about −50 defines a baseline hazard of zero. However, α_4 also has a near infinite variance and covariances with all other coefficients. The software we use, GAUSS, is unable to estimate the standard errors if such a coefficient is included in the model.

TABLE 3
Split Adjusted Hazard Models of Return Time and Rate

	Full Data	Sample 1	Sample 2	Sample 3	Sample 4
Baseline Hazard[1]					
α_1	1.880	2.135	2.139	1.447	1.811
(Standard error)	(0.433)	(0.440)	(0.581)	(0.462)	(0.405)
α_2	0.216	0.224	0.213	0.213	0.217
(Standard error)	(0.018)	(0.021)	(0.020)	(0.021)	(0.020)
α_3	−1.323	−1.345	−1.338	−1.280	−1.334
(Standard error)	(0.094)	(0.109)	(0.105)	(0.114)	(0.107)
Hazard Adjustment					
Price[2]—β_1	0.059	−0.051	−0.109	0.339	0.075
(Standard error)	(0.223)	(0.137)	(0.290)	(0.235)	(0.172)
Logit Split					
Constant—γ_1	−2.450	−2.752	−2.310	−2.528	−2.234
(Standard error)	(0.226)	(0.331)	(0.171)	(0.469)	(0.145)
Price—γ_2	0.584	0.702	0.509	0.590	0.536
(Standard error)	(0.144)	(0.164)	(0.172)	(0.167)	(0.156)
Fit med import—γ_3	0.130	0.351	0.055	0.148	−0.006
(Standard error)	(0.222)	(0.340)	(0.131)	(0.483)	(0.022)
Fit high import—γ_4	0.161	0.279	0.200	0.246	−0.066
(Standard error)	(0.258)	(0.367)	(0.217)	(0.491)	(0.199)

[1] Time is measured in weeks.
[2] Price is measured in hundreds of dollars.

The price coefficient in the hazard adjustment equation is not statistically significant. This result matches the regression result that the price of the item is not having a significant impact on the timing of its return, given that it will be returned. Thus, our hypothesis that price would have a positive impact on hazard (a negative impact on time-to-return) is not substantiated in either the regression or hazard model.

The logit split coefficients are also shown in Table 3. The intercept of −2.450 implies a baseline return probability of 7.944%. The price coefficient is positive and significant, confirming our prior belief that more expensive items are more likely to be returned. The coefficients for the dummies describing the importance of fit are not significant, although the coefficient for "fit is very important" is larger than the coefficient for "fit is somewhat important" as we

hypothesized. The mean fitted return probability taking price and fit import into account is 12.497%, which is larger than the actual sample return rate of 11.957%. We expected such a difference in return probabilities, however, as the modeled return probability takes sample censorization into account, recognizing that some of the items that have been sold and not yet returned will be returned at some future time.

E. Comparing Fit and Predictive Accuracy

We must define a fit measure to compare the regression and hazard models. We have chosen the total absolute difference in cumulative distribution over the time of the longest observation. The regression function's cumulative distribution function is a normal cumulative distribution function with mean equal to the regression mean and standard deviation equal to the standard error of the regression residuals. Normal curves are defined from negative to positive infinity, but return times are only positive. Our choice to sum absolute deviation in cumulative distribution only over the space of sample days alleviates this problem, not penalizing the regression model for predicting returns in negative time. The actual cumulative flow of returns along with the flows that are predicted by applying the estimated regression and hazard models to the 2,024 observations are shown in Figure 3. It is clear from the figure that the hazard model lies much closer to the actual data than the regression model. Total absolute deviation should be smaller for the hazard model than for the regression model.

The actual fit statistics for the estimation samples are shown in Table 4. The hazard model clearly fits better than the regression model. Fit statistics based on a summation only over the time to the final observed return ($n = 104$ days) are also shown in Table 4. These fit statistics suggest that the hazard model also outperforms the regression model over the time of greatest interest, the time period where returns are most likely to occur.

Table 5 contains fit statistics for the four prediction samples. The hazard model is not uniformly superior in fitting the prediction sample as it was in estimation sample fit. It is interesting to note, however, that if only the return observations are examined, the hazard model is uniformly superior. This suggests that the difference in the logit return probabilities is the cause of the superior regression model fit. In fact, as seen in Table 5, the winning fit in each sample clearly belongs to the model with the predicted return probability closer to the actual return rate. This is to be expected, as summing the difference in return probability over the very large number of observation days (1,308 days is the longest censorization time and 104 days is the longest return time) leads to very large cumulative absolute errors. A possible explanation for the mixed prediction results described here is the bell shape of the distribution of returns. We investigate the importance of bell-shaped return patterns next.

4. AN APPLICATION OF THE MODELS TO SIMULATED RETURN DATA

As stated above, the regression model of return time has difficulty with non–bell-shaped return patterns due to its traditional assumption of normality in ran-

TABLE 4
Fit Statistics of Regression and Split Adjusted Hazard Models—Estimation Samples (Absolute Deviation from Actual Cumulative Return Proportion)

	Full Data	Sample 1	Sample 2	Sample 3	Sample 4
Cumulated over maximum observation time					
Cumulation Weeks	186.857	186.857	186.857	186.857	186.857
Regression model with logit split	3.718	3.446	3.768	3.805	3.856
Split adjusted hazard model	3.432	3.233	3.431	3.514	3.542
Cumulated over maximum return time					
Cumulation weeks	14.857	14.714	14.857	14.857	14.857
Regression model with logit split	0.687	0.659	0.725	0.645	0.729
Split adjusted hazard model	0.624	0.561	0.661	0.584	0.691

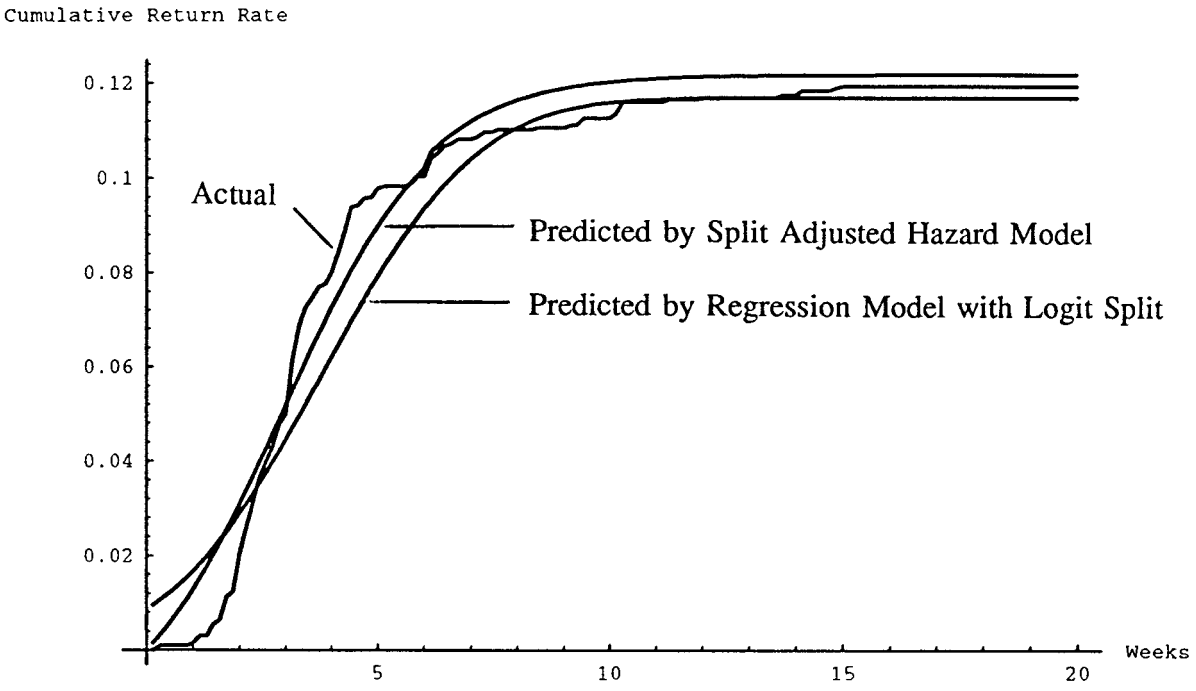


FIGURE 3
Cumulative return rates over time (in weeks)

TABLE 5
Fit Statistics of Regression and Split Adjusted Hazard Models—Prediction Samples (Absolute Deviation from Actual Cumulative Return Proportion)

	Sample 1	Sample 2	Sample 3	Sample 4
Cumulated over maximum *observation* time				
Cumulation weeks	186.857	186.857	186.857	186.857
Regression model with logit split	18.809	21.969	33.475	16.220
Split adjusted hazard model	12.688	28.818	28.038	22.622
Cumulated over maximum *Return* Time				
Cumulation weeks	14.857	13.571	14.714	14.000
Regression model with logit split	1.212	1.318	2.369	1.096
Split adjusted hazard model	0.691	1.505	1.852	1.256
Cumulated over maximum *Return* Time—*Return Observations Only*				
Cumulation weeks	14.714	13.571	14.714	14.000
Regression model (no split)	5.895	6.948	6.665	6.910
Adjusted hazard model (no split)	3.342	4.562	5.329	4.305
Mean return rates (percent)				
Fitted from regression model with logit split	11.701	12.446	11.568	12.189
Fitted from split adjusted hazard model	12.170	12.991	12.217	12.777
Actual	12.909	10.433	13.983	10.667

dom errors. The actual return data that we describe in the above section are reasonably bell-shaped. Therefore, to dramatize the difference in fit of the regression and hazard models as the return pattern becomes non–bell-shaped, we generated a data set of 2,000 observations whose time-to-return was exponentially distributed across only positive times, rather than normally distributed across the real number line. The simulated return data were designed to look similar to the actual data used in the previous section.

Specifically, for each item purchased a price was assigned from ($.50, $.75, . . . , $1.50) with equal frequency (the mean would be $1.00). With this single covariate, the item was "returned" with a probability computed as 1/(1 + exp(2.7 − 0.5*Price). As a result, 207 of the items were returned. For these, the time-to-return was determined by drawing randomly from an exponential distribution with a mean equal to exp(1.15 + 0.05*Price). The typical time-to-return would therefore be exp(1.20) = 3.3 weeks. The times were censored after 10 weeks.

The models used with the simulated return data are similar to those used with the actual return data described above, but are somewhat simpler. No importance-of-fit measures are included. Also, only the coefficient for the constant hazard is reported, as the other coefficients are zero in the theoretical model and in the estimation the coefficients were close to zero.‡

The results of the time-to-return regression and the logit model of return rate are presented in Table 6 and Table 7. As with the results of the models estimated with the actual return data, the regression gives a very poor fit, with only the constant being significant. In the logit model, both the constant and price are significant and have the correct signs. The results of the split adjusted hazard model are shown in Table 8. Baseline hazard and the logit split coefficients are significant. The hazard adjustment is not significant. All coefficients are within two standard errors of the parameters of the distribution from which the data were drawn, with the exception of the baseline hazard coefficient.

Table 9 contains fit statistics for the two models. The split adjusted hazard model clearly fits better. This difference in fit is also clearly visible in Figure 4, which shows the actual cumulative flow of returns along with the flows that are predicted by the regression and hazard models. The difference is even more dramatic, however, in Figure 5, which shows the actual return density along with the estimated return densities from the two models. The underlying exponential shape of the return density can be seen quite clearly in the actual returns. This shape is followed very closely by the split adjusted hazard function. The logit-regression, however, predicts a bell-

‡ (Another reason not to report these results is that GAUSS was unable to estimate standard errors for the coefficients, because the standard errors approach infinity.)

TABLE 6
Time-to-Return Regressions on Return Observations—Simulated Data From Exponential Distribution

Constant	3.407
(Standard error)	(0.541)
Price	−0.599
(Standard error)	(0.475)
Observations	207
R^2	0.008
Mean fitted return time	2.757
Standard error of residuals	2.381

TABLE 7
Logit Model of Return Rate—Simulated Data From Exponential Distribution

Constant—γ_1	2.964
(Standard error)	(0.241)
Price—γ_2	−0.776
(Standard error)	(0.214)

TABLE 8
Split Adjusted Hazard Model of Return Time and Rate—Simulated Data From Exponential Distribution

Baseline Hazard	
α_4	−1.625
(Standard error)	(0.306)
Hazard adjustment	
Price—β_1	0.412
(Standard error)	(0.254)
Logit split	
Constant—γ_1	2.834
(Standard error)	(0.265)
Price—γ_2	−0.708
(Standard error)	(0.233)

TABLE 9
Fit Statistics of Regression and Split Adjusted Hazard Models—Simulated Data From Exponential Distribution (Absolute Deviation from Actual Cumulative Return Proportion)

Cumulated over maximum observation time	
Cumulation time	10.000
Regression model with logit split	0.941
Split adjusted hazard model	0.229

shaped return pattern that is totally inappropriate. This bell-shaped return density is an artifact of the assumption of normally distributed errors. In Figure 5, the restrictive nature of this assumption is clear, as is the benefit of the flexible functional form we have chosen for the split adjusted hazard model.

5. CONCLUSION

As described above, the split adjusted hazard model outperforms the regression model. How would the difference in fit change with more data on customers? The hazard model could do a much better job of adjusting the predicted return rate through both the hazard adjustment and the return/nonreturn split. The hazard model is not unique in this respect, however, as the regression model could also add demographic or other terms to take household or item covariates into account. Still, the regression model does not approach the hazard model in usefulness. The possible differences in fit are too great. An even more serious problem with the regression model, however, is its ever present censorization bias. There is no attempt to correct for the bias that arises from not including in the model sales that will be returned but have not been returned yet. The return time regression also must be estimated with data only from items that are returned, and cannot take into account the information inherent in the much larger set of observations from items that have not been returned. From either a researcher's or a manager's point of view, it is disturbing to estimate a regression model that cannot use the overwhelming majority of the data.

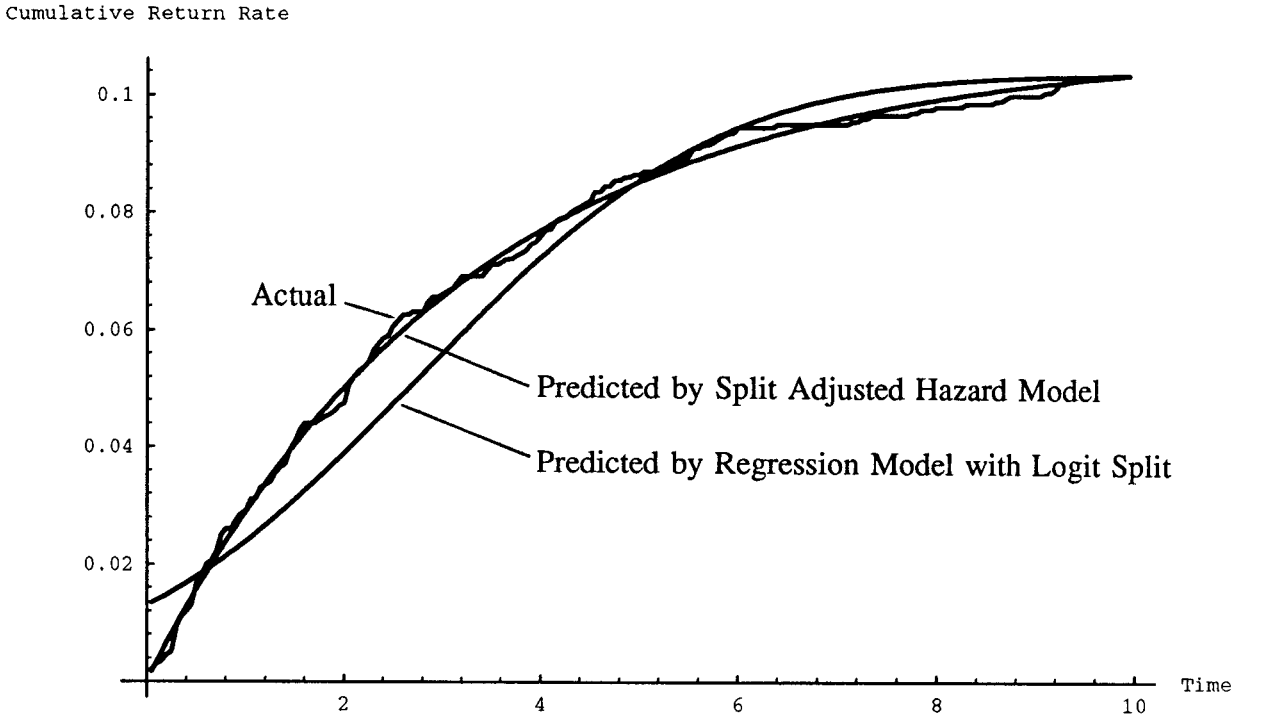


FIGURE 4
Cumulative return rate over time—simulated data from exponential distribution.

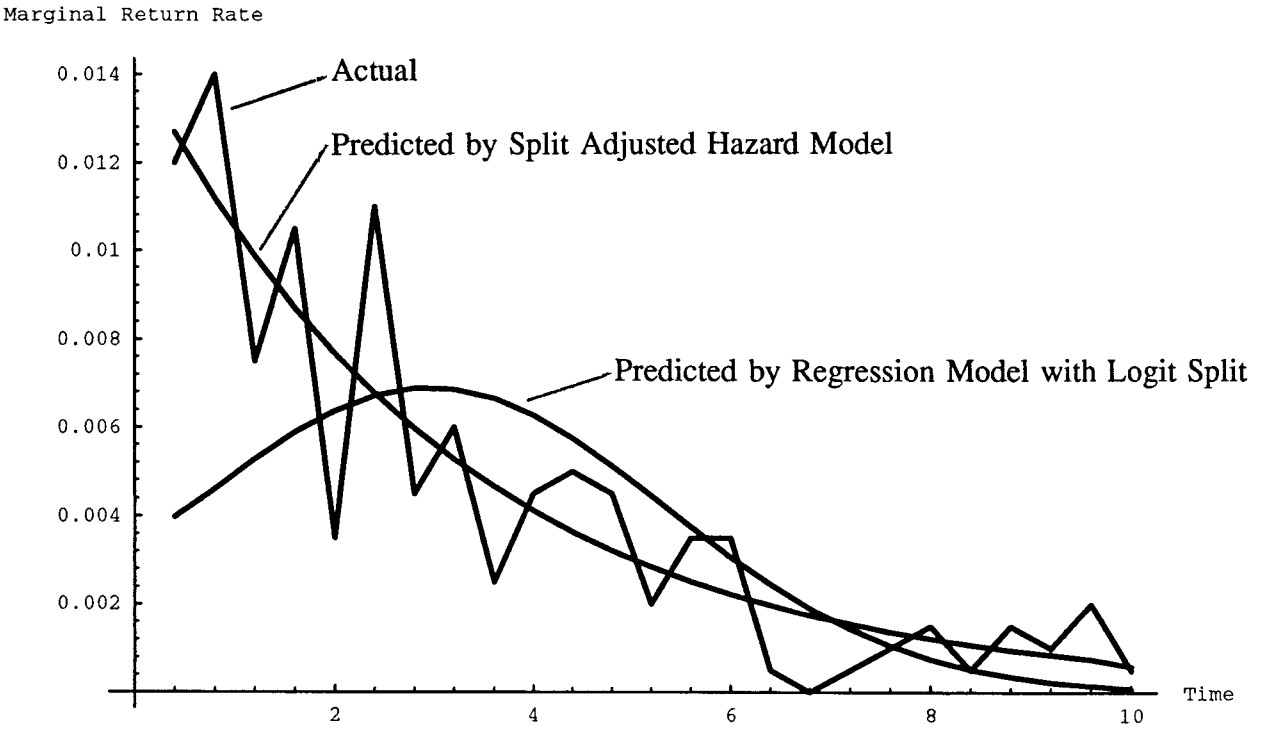


FIGURE 5
Density of return over time—simulated data from exponential distribution.

Another serious problem of the regression model is its inflexibility. The regression model was not an absolute failure with the actual return data from a direct marketer of apparel used here because the return pattern was close to bell-shaped. The regression model will predict only bell-shaped return patterns. The more unusual the return pattern becomes, the more dramatic will be the difference in fit between the flexible hazard model and the inflexible regression model. This was clearly seen above in the extremely poor fit of the regression model versus the hazard model using simulated return data from an exponential distribution. Another very plausible shape the regression model will have a very hard time with is a prolonged return pattern where returns have no sharp peak but do trail off after a few weeks. The regression model will have a very wide distribution, leading to a significant cumulative distribution in negative time periods. The regression model will also have a poorer fit the shorter the time to the peak return rate. This would become most extreme in the case of an exponential distribution of returns, where the return rate starts almost at time zero at its highest rate and then declines steadily. The hazard model is perfectly capable of taking on such a shape, but the bell-shape restriction on the regression model will lead to at least half of predicted returns occurring in negative time. The regression model's lack of flexibility, together with its use of a restricted data sample and an inability to handle sample truncation bias, suggest that it is an inferior model that should be used only with serious caveats.

One may ask, however, if the hazard model's added insight is worth the added trouble. What are the marginal costs of hazard function estimation versus regression? The hazard model is much less common, and its estimation and interpretation are not almost universally understood as are regression models. Regression estimation is easily performed by any number of statistical packages or even spreadsheet software. Such software, however, may not be capable of estimating the logit model of re-

turn rate. Software with a specific logit function or software capable of maximum likelihood estimation would be required to estimate the logit model shown here. Such software is readily available for a few hundred dollars. Estimating the logit model with such software should not be a problem for anyone familiar with standard statistical packages.

Is there software for estimating a split adjusted hazard model? There is no standard software package for performing the hazard function estimation shown here, although any package capable of optimization should be able to estimate the parameters by maximizing the likelihood function (all calculations in this study were done using the GAUSS software package). Writing the program that defines the likelihood function is not difficult for anyone familiar with maximum likelihood estimation. Therefore, the marginal cost of the split adjusted hazard function versus a sophisticated regression analysis is very small. The marginal cost versus a very simple regression analysis is not insignificant, but it should prove no problem for direct marketing companies that regularly use statistical methods.

The small marginal cost of estimating a split adjusted hazard model must be compared with the marginal benefit of the increased accuracy of returns estimation. Let us briefly review the marginal insight from this very small data set for the hazard model versus the regression model. First, let us again state that the hazard model fits and predicts better based on our fit measure. The absolute difference between predicted return rate and actual return rate is also smaller for the hazard model. We learn that the peak return time is at 3.012 weeks rather than the 3.828 that the regression model suggested. The cumulative return level at time zero for the hazard model is zero, while the regression model suggests immediate returns.

What impact would such knowledge have on a direct marketer's costs? Most companies refurbish and resell much of their returned merchandise. Such refurbishing operations must be staffed to handle the return flow in an optimal manner. If staffing is insufficient to handle the return flow, then merchandise sits, increasing inventory costs. If refurbishing is overstaffed, then money is unnecessarily spent on salaries. What if returned merchandise is simply thrown away? Then the timing of returns is of interest only insofar as it allows the return rate and its causes to be more accurately predicted. As discussed above, the hazard function allows the estimation of the "not yet returned" portion of a company's nonreturns. The unbiased estimation of the logit model of returns has its potentially greatest usefulness in the scoring of merchandise and households for return propensity. Small increases in scoring accuracy could potentially lead to significant savings.

Much work remains in the process of changing the state of the art in the modeling of direct marketing merchandise returns. We have demonstrated the usefulness of hazard modeling of returns using a simple model and a single small data set. The size of direct marketing data sets may pose new problems in the estimation of models such as this that require the maximization of relatively flat likelihood functions. Larger data sets, however, might solve other problems. One might expect the actual return distribution graphed in Figure 2 to be much more smooth as the number of observations is increased. This should allow a better fit for the models. Larger data sets also would allow the inclusion of more variables to describe both the customers and the merchandise, further improving the fit of the models and adding new insights into the phenomenon of returns.

Future research on larger data sets can also examine the usefulness of our split adjusted hazard approach to modeling returns in the context of scoring. If the model we propose is useful in developing customer or merchandise return scores, then its usefulness in direct marketing will be clearly established. Case studies would also be useful as they would allow the estimation of actual dollar savings that might result from the use of hazard models.

We look forward to the new knowledge that will be gained as both academic researchers and practitioners share their advances in the modeling of returns. The more we learn about returns, the more accurately companies will be able to score their customers and merchandise, and the more accurately they will be able to predict returns to streamline their operations for both cost savings and increased customer satisfaction.

TECHNICAL APPENDIX

A. Split Adjusted Hazard Model

Hazard rates and probability density and distribution functions are all jointly implied, as shown in equation (1).

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (1)$$

In (1), h is the hazard function, a function of time, t. F is the cumulative distribution function and f is the probability density function. One may estimate one of the functions f, F or h, and simply calculate the other functions. We estimate h because it does not have the integration or range restrictions that f and F have.

In choosing a baseline hazard function, flexibility is important. The fewer restrictions that must be placed on the functional form, the more closely it can fit the data and the less impaired by the researcher's biases it will be. We use the exponentiated quadratic functional form in equation (1), which is similar to the Box-Cox formulation in Flinn and Heckman (5) and Helsen and Schmittlein (6).

$$h(t|\underline{\alpha}, \delta = 1) = \frac{2\alpha_1\alpha_2}{\sqrt{\pi}} \exp[-(\alpha_2 t + \alpha_3)^2] + \exp(\alpha_4) \quad (2)$$

In equation (2), h, the hazard function, is a function of time and depends on a group of coefficients, $\underline{\alpha}$ (where underlining indicates a vector), and the fact that the item will be returned at some point in time (δ is a return indicator variable). The final term, $\exp(\alpha_4)$, is an integration constant that generally will be close to zero.

Adding an exponential adjustment term dependent on covariate $\underline{x}_t$, yields the adjusted hazard function shown in equation (3).

$$\begin{aligned} h(t|\underline{\alpha}, \underline{\beta}, \underline{x}_t, \delta = 1) &= h_0(t|\underline{\alpha}, \delta = 1)\theta(\underline{\beta}, \underline{x}_t) \\ &= \left(\frac{2\alpha_1\alpha_2}{\sqrt{\pi}} \exp[-(\alpha_2 t + \alpha_3)^2] + \exp(\alpha_4)\right) \exp(\underline{\beta}'\underline{x}_t) \end{aligned} \quad (3)$$

Finally, we can add the information contained in the non-return observations ($\delta = 0$). This leads to the split hazard function with the log-likelihood shown in equation (4).

$$\begin{aligned} \log L &= \sum_i \log [f(t_i|\underline{\alpha}, \underline{\beta}, \underline{x}_{it})] \\ &= \sum_i \log [f(t_i|\underline{\alpha}, \underline{\beta}, \underline{x}_{it}, \delta_i = 1)(\delta_i = 1) \\ &\quad + 1(\delta_i = 0)] \end{aligned} \quad (4)$$

In equation (4), the log of the likelihood of observing the data in question is the sum of the log-likelihood of observing the predicted returns at their precise times. Notice that the probability of observing a nonreturn is 1 for items that will not be returned. Also notice that we have moved from the hazard function to the probability density function that is implied in it. This density function is shown in equation (5).

$$f(t|\underline{\alpha}, \underline{\beta}, \underline{x}_t, \delta = 1) = h(t|\underline{\alpha}, \underline{\beta}, \underline{x}_t, \delta = 1) * S(t|\underline{\alpha}, \underline{\beta}, \underline{x}_t, \delta = 1) \quad (5)$$

In equation (5), S represents the survivor function, which is 1 minus the cumulative distribution function, $1 - F$. The functional form of S implicit in our definition of h is shown in equation (6).

$$\begin{aligned} S(t|\underline{\alpha}, \underline{\beta}, \underline{x}_t, \delta = 1) &= \exp[[-\alpha_1[-\text{erf}(\alpha_3) + \text{erf}(\alpha_2 t + \alpha_3)] \\ &\quad - \exp(\alpha_4)] \exp(\underline{\beta}'\underline{x}_t)] \end{aligned} \quad (6)$$

The problem with equation (4) is that we do not have δ. We must replace the deterministic return indicator with a probabilistic measure (π) dependent on covariate y. We use the logit function for π, as shown in equation (7).

$$\pi_i = \frac{\exp(\underline{y}_i\underline{\gamma})}{1 + \exp(\underline{y}_i\underline{\gamma})} \quad (7)$$

The log-likelihood using the logit probability of return function is shown in equation (8).

$$\begin{aligned} \log L = \sum_i \log [&f(t_i|\underline{\alpha}, \underline{\beta}, \underline{x}_{it}, \delta_i = 1)\pi_i(R_i = 1) \\ &+ [(1 - \pi_i) + S_{it}\pi_i](R_i = 0)] \end{aligned} \quad (8)$$

In equation (8), R is an indicator of whether the item was an observed return.

B. Joint Estimation of the Logit and Regression Models

The bias in the regression parameters that may result from the inability to include not-yet-returned observations can be reduced somewhat by the joint estimation of the logit return/nonreturn model and the time-to-return regression model. This is two-stage process. The first stage is the maximum likelihood estimation of the logit model in equation (7) above. A function of the resulting probability of return for each actual return is then entered as a new variable in the regression. The estimated coefficient of this new term, P in equation (9), is an estimate of the product of two theoretical terms. The first is the correlation coefficient (ρ) between the time of return and the estimated logit term. The second element of the product is the standard deviation (σ) of the logit error for that observation. Also in equation (9), Φ^{-1} represents the inverse normal cumulative distribution function and ϕ represents the corresponding normal probability density function.

$$t_i = \underline{\beta}' \underline{x}_i - \mathrm{P}\,\frac{\phi[\Phi^{-1}(\hat{\pi}_i)]}{\hat{\pi}_i}, \text{ where } \mathrm{P} = \rho\sigma \qquad (9)$$

REFERENCES

1. Amemiya, Takeshi (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature,* 19 (December), 1483–1536.
2. Berman, Phyllis and Feldman, Amy (1992), "Trouble in Bean Land," *Forbes*, 150 (1), 42–44.
3. Bliske, W. R., and Murthy, D. N. P. (1992), "Product Warranty Management—I: A Taxonomy for Warranty Policies," *European Journal of Operational Research,* 62, 127–148.
4. Fenvessy, Stanley J. (1992), "Fulfillment Planning: An Overview," in Edward Nash (ed.), *The Direct Marketing Handbook,* New York: McGraw-Hill.
5. Flinn, C. J., and Heckman, J. (1982), "Models for the Analysis of Labor Force Dynamics," in R. L. Bassman and G. F. Rhodes Jr. (eds.), *Advances in Econometrics,* Greenwich, CT: JAI Press Inc., 3–34.
6. Helsen, Kristiaan, and Schmittlein, David C. (1993), "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models," *Marketing Science,* 11 (Fall), 395–414.
7. Hess, James, Chu, Wujin, and Gerstner, Eitan (1996). "Controlling Product Returns in Direct Marketing," *Marketing Letters*, 7 (October), 307–317.
8. Jain, Dipak, and Vilcassim, Naufel (1991), "Investigating Household Purchase Timing Decisions: A Conditional Hazard Function Approach," *Marketing Science,* 10 (Winter), 1–23.
9. Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.