Chapter 15
Discrete Dependent Variables and Duration Models

Abstract Probably the most common statistical technique in predictive modeling is the binary response, or logistic regression, model. This model is designed to predict either/or behavior such as "Will the customer buy?" or "Will the customer churn?" We discuss logistic regression and other discrete models such as discriminant analysis, multinomial logit, and count data methods. Duration models, the second part of this chapter, model the timing for an event to occur. One form of duration model, the hazard model, is particularly important because it can be used to predict how long the customer will remain as a current customer. It can also predict how long it will take before the customer decides to make another purchase, switch to an upgrade, etc. We discuss hazard models in depth.

Many database marketing phenomena we want to model are discrete. For example, consider predicting the brand of car a customer will choose in an upcoming car purchase. Or consider predicting which customers will respond to a direct mail offer. The brand choice or the response to the offer may be modeled to be a function of customer's demographic and purchase behavioral characteristics. However, the dependent variable is categorical (i.e., an identification of a brand or a response indicator). These are discrete dependent variables.

This chapter will discuss various statistical models that are designed to analyze discrete or what are also called qualitative dependent variables. We start with models for a binary response including the linear probability model, logit model (or logistic regression), probit model and discriminant analysis. In the next section, we introduce models for multinomial response that generalize the binary response models. Next, we briefly study models specially designed for count data, followed by the tobit model or censored regression. Finally, we discuss hazard models appropriate for analyzing duration data. The hazard model analyzes the time until an event occurs, so has both discrete and continuous aspects.

15.1 Binary Response Model

The dependent variable for the binary response model can take two different values. For example, a consumer responds to the promotional event ($Y = 1$) or will not ($Y = 0$). Or a customer purchases the firm's brand of car ($Y = 1$) or a competitor's brand ($Y = 0$). The specific values (0/1) assigned to each outcome of the dependent variable are arbitrary since all that matters is that we have a code for knowing which values of Y correspond to which outcomes. So we can assign $Y = 0$ for the response to the promotional event and $Y = 1$ for the non-response. Or it can be $Y =$ "Yes" for the response and $Y =$ "No" for the non-response.

In order to clarify the following discussion, consider a credit scoring model that has become a standard application for financial institutions deciding whether to grant credit to customers. The goal of the credit scoring model is to automate the credit-granting decision process by predicting the default probability for each credit applicant. The consumer's future response will be either default ($Y = 1$) or not default ($Y = 0$). Typically a customer's default behavior/response is modeled to be a function of her demographic and credit-related behavioral characteristics with a number of macro economic variables.

15.1.1 Linear Probability Model

Our goal is to model the default behavior of customer i. Let Y_i to be the default indicator variable for customer i that is assumed to be randomly drawn from a Bernoulli distribution with a mean of p_i. Hence, the probability that Y_i equals 1 is p_i while the probability that it equals 0 is $1 - p_i$. That is,

$$Y_i = \begin{cases} 1, P(Y_i = 1) = p_i \\ 0, P(Y_i = 0) = 1 - p_i \end{cases} \tag{15.1}$$

Our dependent variable Y_i will have a relationship with a set of independent variables by assuming that p_i is a function of the set of independent variables. That is, we assume that $p_i = F(\boldsymbol{\beta}'\mathbf{X}_i)$ where $\mathbf{X}_i$ is a vector of independent variables for customer i (e.g., customer's credit-related variables) and $\boldsymbol{\beta}$ is a corresponding parameter vector. Then $E(Y_i) = (1)(p_i) + (0)(1 - p_i) = p_i = F(\boldsymbol{\beta}'\mathbf{X}_i)$.

The key issue in a binary response model is the specification of the link function F. The simplest is to assume that F is linear, $p_i = F(\boldsymbol{\beta}'\mathbf{X}_i) = \boldsymbol{\beta}'\mathbf{X}_i$. Now since $E(Y_i|\mathbf{X}_i) = \boldsymbol{\beta}'\mathbf{X}_i$, we can derive the following linear probability model.

$$Y_i = \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i \tag{15.2}$$

where ε_i is the error term of customer i. A linear probability model is a traditional regression model with a binary dependent variable Y_i and a set

of independent variables $\mathbf{X}_i$. Consistent with the assumption for a classical linear regression, the expected value of the error term is 0, which can be seen in the following calculation:

$$\begin{aligned} E(\varepsilon_i) = E(Y_i - \boldsymbol{\beta}'\mathbf{X}_i) &= p_i(1 - \boldsymbol{\beta}'\mathbf{X}_i) + (1 - p_i)(0 - \boldsymbol{\beta}'\mathbf{X}_i) \\ &= p_i - p_i\boldsymbol{\beta}'\mathbf{X}_i - \boldsymbol{\beta}'\mathbf{X}_i + p_i\boldsymbol{\beta}'\mathbf{X}_i \\ &= p_i - \boldsymbol{\beta}'\mathbf{X}_i \\ &= 0 \end{aligned}$$

However, there are a number of shortcomings to the linear probability model. First, the error term in Equation 15.2 will violate the homoscedasticity assumption of classical linear regression model since Y_i is a binary discrete variable. Noting that ε_i can only take two values, $1 - \boldsymbol{\beta}'\mathbf{X}_i$ with probability of p_i and $-\boldsymbol{\beta}'\mathbf{X}_i$ with probability $1 - p_i$, we compute the variance of the error term as:

$$Var(\varepsilon_i) = E(\varepsilon_i^2) = p_i(1 - \boldsymbol{\beta}'\mathbf{X}_i)^2 + (1 - p_i)(-\boldsymbol{\beta}'\mathbf{X}_i)^2 = \boldsymbol{\beta}'\mathbf{X}_i(1 - \boldsymbol{\beta}'\mathbf{X}_i)^2$$

That is, the variance is not homoscedastic, but varies with the values of independent variables. The second problem associated with the linear probability model is more serious. We refer to it as the "unit interval" problem. Since p_i is the *probability* that $Y_i = 1$, its value should be bounded from 0 to 1. The predicted value of $\hat{p}_i = \hat{\boldsymbol{\beta}}'\mathbf{X}_i$ in the linear probability model is not guaranteed to be within the [0, 1] range. As a result, predictions can be impossible to interpret as probabilities. In addition, the heteroscedasticity, if not corrected for, can increase prediction error. Because of these shortcomings, the linear probability model is becoming less frequently used in database marketing even though it is computationally simple to use.

Several researchers have discussed ways of overcoming the shortcomings of linear probability models (Judge et al. 1985; Greene 1997). For example, Goldberger (1964) suggested correcting the heteroscedasticity problem by employing GLS (generalized least squares) estimation. Judge et al. (1985) proposed an inequality-restricted least squares approach to overcome the unit interval problem, however their remedies are sample-dependent.

15.1.2 Binary Logit (or Logistic Regression) and Probit Models

A direct way to remedy the unit interval problem is to find a link function that satisfies the [0, 1] constraint on p_i. One such function is a cumulative density function. The value of $p_i = F(\boldsymbol{\beta}'\mathbf{X}_i)$ or the probability of $Y_i = 1$ approaches to 1 as the value of $\boldsymbol{\beta}'\mathbf{X}_i$ goes to the plus infinity while it approaches to 0 as the value of $\boldsymbol{\beta}'\mathbf{X}_i$ goes to the minus infinity (see Fig. 15.1). Even though

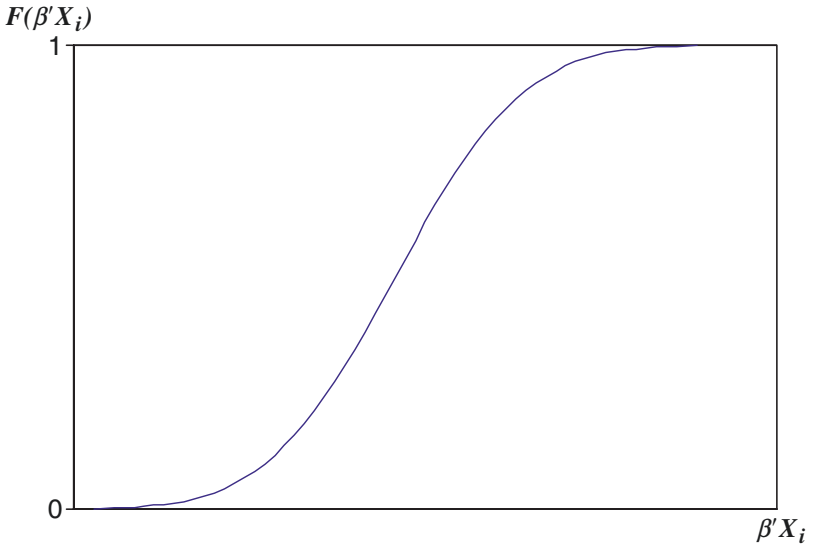

Fig. 15.1 Using a cumulative distribution function as the link function for a binary response model.

any cumulative distribution functions have this property, the following two cumulative density functions are most frequently used.

$$\text{Logistic cdf: } F(\boldsymbol{\beta}'\mathbf{X}_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{X}_i)} = \frac{1}{1+\exp(-\boldsymbol{\beta}'\mathbf{X}_i)} \tag{15.3a}$$

$$\text{Standard normal cdf: } F(\boldsymbol{\beta}'\mathbf{X}_i) = \int_{-\infty}^{\boldsymbol{\beta}'\mathbf{X}_i} \phi(t)dt = \Phi(\boldsymbol{\beta}'\mathbf{X}_i) \tag{15.3b}$$

where $\phi(\cdot)$ is the density of a standard normal distribution and $\Phi(\cdot)$ is the corresponding cumulative density. The model is called a binary logit or logistic regression model when the link function F is logistic, and a probit model when F is the standard normal. The shape of the logistic distribution is very similar to that of the normal distribution except in the tails, which are heavier (Greene 1997). Hence, its estimation results are also similar. It is difficult to show that the logistic is better (or worse) than the standard cumulative normal on theoretical grounds. However, the binary logit model may be more frequently used because of its mathematical convenience – once the logistic regression has been estimated, Equation 15.3a provides a convenient formula for calculating predicted probabilities. In contrast, the probit model requires a table look-up of the normal distribution to calculate predicted probabilities, as shown in Equation 15.3b.

The binary logit and probit models are estimated using the method of maximum likelihood. Each observation is treated as an independent random draw from the identical Bernoulli distribution. Hence, the joint probability

or the likelihood function of the binary response model with the sample size of n can be written as

$$L = \prod_{i=1}^{n} [F(\boldsymbol{\beta}'\mathbf{X}_i)]^{Y_i}[1 - F(\boldsymbol{\beta}'\mathbf{X}_i)]^{1-Y_i} \tag{15.4}$$

The estimation of the parameter vector $\boldsymbol{\beta}$ involves finding a set of parameters to maximize the likelihood function, Equation 15.4. It is difficult to derive an analytical solution for $\boldsymbol{\beta}$ because the likelihood function (except in the case of the linear probability model) is highly nonlinear. Therefore, the estimates $\boldsymbol{\beta}$ are found through an iterative search using Newton's BHHH method for the logit or probit model (Berndt et al. 1974).

Let us provide a simple example of the logistic regression model applied to credit scoring. Again our objective is to predict the prospect's default probability. For illustration, we assume that a customer's default likelihood is a function of her or his annual income and marital status, even though there are several other variables related to the default behavior. Annual income x_1 (measured in \$1,000) is a continuous variable while marital status x_2 is defined to be categorical ($x_2 = 1$ if the customer is single, divorced, or separated, and $x_2 = 0$ otherwise). The logistic regression model is applied to a sample of current customers whose default behaviors have been observed. We code $Y_i = 1$ if customer i defaults and $Y_i = 0$ if not. The following equation summarizes the estimation result.

$$P(Y_i = 1) = \frac{e^{0.5-0.01x_1+1.1x_2}}{1 + e^{0.5-0.01x_1+1.1x_2}} = \frac{1}{1 + e^{-(0.5-0.01x_1+1.1x_2)}} \tag{15.5}$$

Equation 15.5 indicates that there is a negative relationship between income (x_1) and default likelihood. A customer with higher income will have the lower chance of default. On the other hand, there is a positive relationship between marital status and the default probability. A customer who is single, divorced, or separated will have the higher probability of default. More specifically, the single customer with the income of \$40,000 ($x_1 = 40$ and $x_2 = 1$) is predicted to have a default probability of 0.08 while the married customer with the same income ($x_1 = 40$ and $x_2 = 0$) is predicted to have 0.03 probability of default. Hence, the marginal effect of the marital status x_2 (at $x_1 = 40$) is 0.05. This is the difference in default probabilities between a married customer and a single, divorced, or separated customer.

An intuitive way to interpret an individual logit parameter (β) is to consider the "odds ratio". First, the *odds* of a yes response ($Y_i = 1$) is defined to be $P(Y_i = 1)/P(Y_i = 0)$, i.e., the likelihood of the event happening relative to not happening. For example, an odds of "3", also known as "3 to 1 odds," means that the likelihood of defaulting is three times greater than the likelihood of not defaulting. Second, the *odds ratio* is the ratio of the odds when the independent variable equals $X_i + 1$ divided by the odds when the independent variable equals X_i. Hence the odds ratio shows by

what factor the odds change when the independent variable increases by one unit.

It can be shown using simple algebra that for logistic regression, the odds ratio equals $exp(\beta)$ – the exponentiation of a logistic regression parameter tells us the factor by which the odds change per unit change in the corresponding independent variable. For example, the coefficient for marital status in Equation 15.5 is $\beta = 1.1$. Since $exp(1.1) = 3$, this tells us that the odds of defaulting change by a factor of 3, which means an increase of 200%, if the customer is single, devorced, or separated ($x_2 = 1$) versus married ($x_2 = 0$). The coefficient for income is $\beta = -0.01$. Since $exp(-0.01) = 0.99$, that means that the odds of defaulting change by a factor of 0.99, which means a *decrease* by 1% ($(1 - 0.99) \times 100\% = -1\%$) per $1,000 increase in income.

In order to compare the impact of variables measured in different units, we can calculate the change in the odds per *standard deviation* change in the independent variable. Let σ = the standard deviations of the indendent variable of interest, then the odds ratio *per standard deviation change* can be shown to equal $exp(\beta\sigma)$. Hence if the standard deviation of income in our data is $15,000, we have $exp(-0.01 \times 15) = 0.86$, so a standard deviation increase in income on the odds of defaulting decreases the odds of defaulting by 14% ($(1 - 0.86) \times 100\% = -14\%$).

15.1.3 Logistic Regression with Rare Events Data

Researchers have addressed problems in the statistical analysis of rare events data using logistic regression or binary probit. In the social and epidemiological sciences, there are dozens to thousands of times fewer ones (events) than zeros (non-events), for example in the analysis of wars, coups, presidential vetoes and infections by uncommon diseases. In database marketing, response rates below 1% are not unusual. When applied to rare events data, logistic regression or binary probit can *under-estimate* customer response probability.

Statistically, the problem emerges from the fact that the statistical properties of linear regression models are invariant to the (unconditional) mean of the dependent variable. But the same is not true for logistic regression or binary probit (King and Zeng 2001). In fact, King and Zeng show that for the logistic regression model, when the mean of a binary dependent variable, or the frequency of events in the data, is very small, parameter estimates of logistic regression become more biased and predicted response probabilities become too pessimistic. There are two intuitive explanations for this. (1) King and Zeng argue that in rare events data, there are plenty of values available for the independent variables to understand the circumstances that cause a non-event, however, there are far fewer to understand the circumstances that cause an event. Those few values do not fully cover the tail of the logistic distribution, and so the model infers that there are fewer circumstances under

which the event will occur, resulting in an under-estimate of the probability the event occurs. King and Zeng show that the primary manifestation of this is downward bias in the constant term of the logistic regression.[1] (2) Parametric link functions such as those used for logit or probit lack flexibility. Logit and probit models assume specific shapes of the underlying link function (see Fig. 15.1), implying a given tail probability expression that remains invariant to observed data characteristics. As a result, these models cannot adjust for the case when there are not enough observations to fully span the range needed for estimating these link functions (see Kamakura et al. (2005) for further discussion).

The bias in logistic regression with rare events is potentially very important because it suggests that taking predicted logistic response probabilities literally under-estimates the actual likelihood of response. Too many customers will be deemed unprofitable and the firm will incur an opportunity loss by not contacting many customers who would have been profitable (a "Type II Error" as described in Chapter 10, Sect. 10.3.5.4).

Researchers have proposed three approaches to overcome the problem with using logistic regression (or probit) to analyze rare events data. These are all statistical approaches aimed at calculating unbiased individual-level predictions. When applying predictive models at the n-tile level, it is practical to use each n-tile's actual response rate as the prediction for customers in that n-tile (see Chapter 10, Sect. 10.3.5.1). Turning now to the statistical approaches to calculating unbiased individual-level predictions, the first is to adjust the coefficients and the predictions of the estimated logistic regression model. King and Zeng (2001, p. 147) describe how to adjust the maximum likelihood estimates of the logistic regression parameters to calculate "approximately unbiased" coefficients, $\tilde{\beta}$. When the $\tilde{\beta}$'s are inserted into the logistic equation for a given customer's set of independent variables, X_i, the resulting prediction is called $\tilde{\pi}_i$. King and Zeng then derive the following adjustment to predicting the probability of an event using logistic regression when events are rare:

$$P(Y_i = 1) = \tilde{\pi}_i + (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)X_i Var(\hat{\beta})X_i' \tag{15.6}$$

where $Var(\hat{\beta})$ is the estimated variance/covariance matrix of the estimated coefficients. First, since we are dealing with rare events data, $\tilde{\pi}_i$ will be small and so predictions using Equation 15.6 are adjusted upwards. Second, to the extent that we have a very large sample size, we have more information, $Var(\hat{\beta})$ will be relatively small, and there is less need for adjustment.[2]

[1] King and Zeng note that logistic regression coefficients estimated using maximum likelihood are biased but consistent. However, the bias tends toward zero if observations are randomly sampled and the percentage of events approaches 50%. This makes sense given our intuitive explanation for the bias.

[2] Software for implementing these adjustments, called "Zelig," is available at Professor King's website, http://gking.harvard.edu/stats.shtml. We thank Professor King for his insights on this issue and for making his software available.

A second approach to addressing the bias issue is "choice-based sampling." In choice-based sampling, the sample is constructed based on the value of the dependent variable. For example, if we were constructing a predictive model for customer churn, we would gather all the churners and all the non-churners, then randomly select 10,000 churners and 10,000 non-churners. The intuitive appeal of choice-based sampling is that we now have an equal (or at least more well-balanced) number of churners and non-churners, so being a churner is no longer a rare event. The problem is that choice-based sampling may induce a selection bias regarding the independent variables because there may be unobserved factors that systematically produce different distributions of independent variables for churners and non-churners (King and Zeng 2001; Donkers et al. 2003).

As a result, choice-based sampling produces biased results and corrections must be undertaken. One of the popular ones is "Weighted Exogenous Sampling Maximum-Likelihood" (WESML), developed by Manski and Lerman (1977) (see Singh (2005) for an application). King and Zeng (2001) propose a simpler technique they find is equivalent to other econometric solutions, and show that it performs similarly to WESML, although acknowledge that WESML can be more effective with large samples and with functional form misspecification. The King and Zeng adjustment is only to adjust the constant term in the maximum-likelihood-estimated logistic regression model:

$$\hat{\beta}_{0,adj} = \hat{\beta}_0 - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \tag{15.7}$$

where $\hat{\beta}_{0,adj}$ is the adjusted constant term, $\hat{\beta}_0$ is the MLE estimate of the constant term, τ is the percentage of "1's" (i.e., churn, respond, etc.) in the population, and $\bar{y}$ is the fraction of 1's in the choice-based sample. For example, τ might equal 2% but $\bar{y}$ could equal 50%. One can see that since $\tau < \bar{y}$, the adjusted constant term will be smaller than the MLE-estimated constant term.

Donkers et al. (2003) investigated the similar issue and derived the adjustment factor to the constant term of the logistic regression. Their adjustment formula is identical to Equation 15.7 except that they did not consider the population (or prior) percentage. That is, $\hat{\beta}_{0,adj} = \hat{\beta}_0 - \ln[\overline{y}/(1-\overline{y})]$.

Research is needed to investigate WESML as well as King and Zeng's adjustment in a database marketing context, and to find the conditions under which random sampling (with King and Zeng's adjustment Equation 15.6) is preferred to choice-based sampling (with either King and Zeng's adjustment Equation 15.7 or WESML). See Ben-Akiva et al. (1997) for further discussion and cautions regarding choice-based sampling.

A third approach to addressing the rare-events problem is to relax the logit or probit parametric link assumptions, which can be too restrictive for rare events data (Bult and Wansbeek 1995; Naik and Tsai 2004). Naik and Tsai (2004) proposed an isotonic single-index model and developed an efficient

algorithm for its estimation. Different from the logistic regression or probit model, its link function is flexible so that it encompasses all proper distribution functions and identifies the underlying distribution using information in the data rather than imposing a particular shape. More work is needed to investigate Naik and Tasi's method in a database marketing context.

15.1.4 Discriminant Analysis

Database marketers frequently use discriminant analysis as an alternative to logistic regression or probit in analyzing binary response data. Discriminant analysis is a multivariate technique identifying variables that explain the differences among several groups (e.g., respondents and non-respondents to mailing offers) and that classify new observations or customers into the previously defined groups.

Discriminant analysis involves deriving linear combinations of the independent variables ($\boldsymbol{\beta}'\mathbf{X}$) that will discriminate between *a priori* defined groups (e.g., responders and non-responders). The weights, called discriminant coefficients, are estimated in such a way that the misclassification error rates are minimized or the between-group variance relative to the within-group variance is maximized. Once the discriminant weights $\boldsymbol{\beta}$ are determined, the discriminant score ($\mathbf{y}_i = \boldsymbol{\beta}'\mathbf{x}_i$) for each customer i can be obtained by multiplying the discriminant weight associated with each independent variable by the customer's value on the independent variable and then summing over the set of independent variables. The resulting score for each customer can be transformed into a posterior probability that gives the likelihood of the customer belonging to each group.

We apply the discriminant analysis into the credit scoring data in the previous section where a logistic regression was applied. The dependent variable is binary (i.e., $Y_i = 1$ if customer i defaults and $Y_i = 0$ if not). And there are two independent variables, annual income x_1 and marital status x_2. The discriminant function (d) is estimated to be $d = -0.02x_1 + 1.87x_2$. That is, the discriminant coefficient is -0.02 for x_1 and 1.87 for x_2. Among defaulters whose Y_i is equal to 1, the mean value of x_1 is 30 and the mean of x_2 is 0.7. Among non-defaulters whose Y_i is equal to 0, the mean value of x_1 is 50 and the mean of x_2 is 0.3. Hence, the average discriminant score of defaulters is $d_{defaulters} = (-0.02)(30) + (1.87)(0.7) = 1.249$ while the average discriminant score of non-defaulters is $d_{non\text{-}defaulters} = (-0.02)(50) + (1.87)(0.3) = 0.469$. A single customer with annual income of \$40,000 is classified as a defaulter because her or his discriminant score is $(-0.02)(40) + (1.87)(1) = 1.79$, which is greater than the midpoint of $d_{defaulters}$ and $d_{non\text{-}defaulters} (= 0.859)$.

There are many studies on the relative performance of logistic regression and discriminant analysis in the analysis of binary dependent variables. In terms of computational burden, discriminant analysis is better. Ordinary

least squares can be used to estimate the coefficients of the linear discriminant function, while nonlinear optimization methods are required to estimate the coefficients of the logistic regression (Maddala 1983). However, computational simplicity is no longer an adequate criterion, considering the high-speed computers available now.

Amemiya and Powell (1983) found that if the independent variables are multivariate normal, the discriminant analysis estimator is the maximum-likelihood estimator and is asymptotically efficient. On the other hand, the discriminant analysis estimator is not consistent when the independent variables are not normal, but the logistic regression is and therefore more robust. Press and Wilson (1978) compared the performances of these two estimators in terms of the number of correct classification when the independent variables were dummy variables, and thus the assumption of normality was violated. They found that the logistic regression did slightly better than discriminant analysis.

15.2 Multinomial Response Model

Multinomial response models generalize binary response models to the situation of more than two possible outcomes or choice alternatives. Hence, the dependent variable for the multinomial response model takes more than two values. For example, consider a customer's choosing a brand of a car among J alternative brands. The consumer response will be the choice of the first brand ($Y = 1$), the second brand ($Y = 2$), or the Jth brand ($Y = J$).

It is much more complex to estimate multinomial response models than binary response models. However, the fundamental concepts, including the interpretation of results are identical. Marketers have frequently employed a multinomial logit model in analyzing multinomial response (or choice) data because it is mathematically more tractable. However, the multinomial logit fundamentally has a structural problem called the *IIA* (Independence of Irrelevant Alternatives) property (Maddala 1983; Hausman and McFadden 1984). The multinomial probit model avoids the *IIA* problem but it is computationally intense. More recently, McCulloch and Rossi (2000) proposed a simulation-based estimation technique called Gibbs sampling to overcome the computational problem of the multinomial probit model.

A multinomial logit is similar to a binomial logit, except that the number of choice (or response) alternatives is J. So we consider a consumer facing a choice problem among J alternatives. Then the probability of the consumer i's choosing alternative j can be written as:

$$P(Y_{ij} = j) = \exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\alpha}_j'\mathbf{z}_i) / \sum_{k=1}^{J} \exp(\boldsymbol{\beta}'\mathbf{x}_{ik} + \boldsymbol{\alpha}_k'\mathbf{z}_i) \tag{15.8}$$

Table 15.1 Estimates of logit coefficients for electric utility customers (From Gensch et al. 1990)

Independent variables	Estimates of logit coefficients	*t*-value
Invoice price	3.45	1.45
Energy losses	7.45	3.29[a]
Appearance	4.32	2.11[a]
Availability of spare parts	2.45	0.99
Clarity of bid document	1.62	0.36
Knowledgeable salesmen	2.78	1.12
Maintenance requirements	2.64	1.31
Warranty	8.22	4.05[a]

[a]Significant at 0.05 level

where $\mathbf{z}_i$ represents a set of independent variables describing characteristics of customer i (e.g., consumer's income), $\mathbf{x}_{ij}$ are a set of independent variables representing the attributes of alternatives (e.g., price of brand j faced by the customer i), and $\boldsymbol{\beta}$ and α are parameters to be estimated. The alternative specific parameters α_j indicate that the effect of an independent variable is different across different alternatives.

Multinomial response models have rarely been employed in database marketing. The reason may be that database marketers do not usually have competitors' data (e.g., which competitor's brand to choose). They only observe whether customers do or do not purchase their products (or react/no react to their promotional offers). However, there are some database marketing problems in which a multinomial response model can be useful.

Gensch et al. (1990) used customer research and the multinomial logit model to understand the preferences and the decision-making processes of ABB Electric's customers. ABB sold medium-sized power transformers, breakers, switchgear, relays, etc., to electric utilities in the North American market, and its major competitors included General Electric, Westinghouse, McGraw–Edison, and so on. Gensch et al. identified 8 attributes that customers used to select among 7 alternative suppliers including ABB. The multinomial logit (Equation 15.8) was applied to evaluate which attributes were the most salient or key in determining the choice among 7 suppliers. Table 15.1 illustrates the output of the multinomial logit. It indicates that warranty, energy losses and appearance are the key variables in determining which supplier to purchase from. Gensch et al. also found that the salient attributes identified by the logit model are quite different from the attributes customers say are most important in their choice.

As mentioned, multinomial logit has not been used frequently in database marketing. However, we can think of several situations where the multinomial logit can be useful. Suppose that we classify current customers into J clusters (or segments) using cluster analysis. We can apply the multinomial logit with segment membership of each customer serving as the dependent variable. The estimated multinomial logit model can be used to identify the

segment membership of potential customers. In addition, a multinomial logit model can be valuable when database marketers attempt to predict what products their customers will purchase. For example, insurance salesperson wants to know which products (e.g., term insurance, endowment insurance, and accident death benefits) the customer would be likely to buy when the salesperson needs to decide which products to cross-sell.

15.3 Models for Count Data

Some dependent variables are not categorical, but are discrete with an ordered metric. For example, the number of beers a consumer drinks in a week can be 0, 1, 2, and so on. Another example may be the number of mail orders a customer makes in a year. A multinomial logit model will not be appropriate because the dependent variable is ordered. One may apply a classical linear regression. But if there are many small numbers in the data, the discrete characteristic of the data will be prominent. As a result, the classical linear regression, which assumes a normal error term and hence a continuous dependent variable, may not work well either.

15.3.1 Poisson Regression

Let Y_i to be the value of the dependent variable for customer i. The Poisson regression model assumes that each $Y_i (i = 1, 2, \ldots, n)$ is a random variable independently drawn from a Poisson distribution with parameter λ_i. Its probability density function is

$$P(Y_i = y_i) = \lambda_i^{y_i} \exp(-\lambda_i)/y_i! \qquad y_i = 0, 1, 2, \ldots. \tag{15.9}$$

The dependent variable Y_i will have a relationship with a set of independent variables specified by a link function. The log-linear link function is frequently used in Poisson regression. That is,

$$\ln \lambda_i = \boldsymbol{\beta}' \mathbf{X}_i \tag{15.10}$$

where $\mathbf{X}_i$ is a vector of independent variables for customer i and $\boldsymbol{\beta}$ is a corresponding parameter vector. It can be shown that the mean equals the variance for the Poisson distribution. That is:

$$E(Y_i|\mathbf{X}_i) = Var(Y_i|\mathbf{X}_i) = \lambda_i = \exp(\boldsymbol{\beta}' \mathbf{X}_i)$$

The Poisson regression model is typically estimated by the method of maximum likelihood. Its log-likelihood function with the sample size of n can be written as

$$\ln L(\boldsymbol{\beta}|Y_i, X_i) = \sum_i (Y_i \ln \lambda_i - \lambda_i - \ln Y_i!) \propto \sum_i Y_i(\boldsymbol{\beta}'\mathbf{X}_i) - \sum_i \exp(\boldsymbol{\beta}'\mathbf{X}_i) \tag{15.11}$$

Parameters will be estimated by finding $\boldsymbol{\beta}$ maximizing the log-likelihood function in the Equation 15.11. Similar to the logit and probit, the log-likelihood function of the Poisson regression model is nonlinear and, hence, there is no analytical solution. An iterative search routine such as Gauss–Raphson can be applied to find the optimal solution.

15.3.2 Negative Binomial Regression

The Poisson regression model is often criticized because of its implicit assumption that the mean of the Poisson distribution is the same as its variance. There are a number of tests available (called tests for overdispersion) to determine whether this assumption is valid (Greene 1997). A number of researchers have found the assumption to be violated (Hausman et al. 1984; McCulloch and Nelder 1983). In that case, a more flexible model should be applied. A number of researchers have proposed several approaches to extend the Poisson regression model. We briefly discuss the most popular extension called a negative binomial regression.

Let the λ_i parameter of the Poisson distribution equal $\delta_i u_i$ where δ_i is the component observable to the researcher ($\ln \delta_i = \boldsymbol{\beta}'\mathbf{X}_i$) and u_i is the random error or the term for explaining unobserved cross-sectional heterogeneity, that is, differences between customers that are not explicitly measured by the researcher. Hence, $\ln \lambda_i = \ln \delta_i + \ln u_i = \boldsymbol{\beta}'\mathbf{X}_i + \ln u_i$. Then the conditional distribution of Y_i, given u_i, is Poisson with mean and variance of $\lambda_i = \delta_i u_i$. That is,

$$P(Y_i = y_i|u_i) = \lambda_i^{y_i} \exp(-\lambda_i)/y_i! = (\delta_i u_i)^{y_i} \exp(-\delta_i u_i)/y_i! \tag{15.12}$$

The unconditional distribution is simply the expected value of the conditional distribution integrated over the conditioning variable u_i. That is,

$$P(Y_i = y_i) = \int_0^\infty [(\delta_i u_i)^{y_i} \exp(-\delta_i u_i)/y_i!] g(u_i) du_i \tag{15.13a}$$

The choice of the density of u_i will determine the form of the unconditional distribution. For mathematical convenience, a gamma distribution is generally assumed for the density of u_i. Then the unconditional density of Y_i in Equation 15.13a becomes the density of the negative binomial distribution:

$$P(Y_i = y_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \left[\frac{\delta_i}{\delta_i + \theta}\right]^{y_i} \left[1 - \frac{\delta_i}{\delta_i + \theta}\right]^{\theta} \tag{15.13b}$$

The negative binomial distribution in Equation 15.13b has a mean of δ_i and a variance of $\delta_i(1 + \delta_i/\theta)$. In contrast to the Poisson regression, the mean is different from the variance.

Models for count data have not been used in database marketing. However, there are situations in which they could be useful. Econometricians used count models for the number of accidents on a natural-gas pipeline, the number of patents issued and so on. Similar database marketing examples include the number of customer complaints, the number of returns, and the number of responses to direct marketing offers.

We can easily fit Poisson regression and the negative binomial regression models using the SAS GENMOD procedure.

15.4 Censored Regression (Tobit) Models and Extensions

The dependent variables of our interest in database marketing often are censored. As defined by Wooldridge (2002, p. 517), "censored regression models generally apply when the variable to be explained is partly continuous but has positive probability mass at one or more points." A simple example is monthly expenditures from a catalog firm's customers. Expenditures are continuous, but there will be several customers who spend no money during a particular month. Hence expenditures is a continuous dependent variable but has a positive probability mass at zero, and has no observations less than zero.

This can be modeled using a (Type I) Tobit model as follows: Define y_i^* as a latent variable that reflects customer i's propensity for spending money on the firm's product in a given time period. Consider a sample of size n, $(y_1^*, y_2^*, \ldots, y_n^*)$. Those observations of $y^* \leq c$ will be recorded as the value c (c is usually zero as in the case of expenditures). The resulting sample of observations $y_1, y_2, \ldots, y_n$ is said to be a censored sample. Note that $y_i = y_i^*$ if $y_i^* > c$ and $y_i = c$ otherwise.

One might proceed by estimating y_i as a function of various independent variables using OLS regression. However, Wooldridge (2002, pp. 524–525) shows that the resulting estimates will be inconsistent, whether we use all n observations or just those for which $y_i > 0$. The problem is that OLS does not account for the underlying censoring process. The regression model specially designed to analyze the censored sample is the censored regression (or Tobit) model can be written as:

$$y_i^* = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i \tag{15.14a}$$

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases} \tag{15.14b}$$

Equation 15.14a shows that a *latent* variable follows a usual regression model, with error term ε_i having mean zero and variance σ^2, while Equation 15.14b shows that the dependent variable we observe can only be non-negative. Our problem is to estimate $\boldsymbol{\beta}$ and σ^2 using n observations of y_i and x_i. This

model is first studied by Tobin (1958). Because he related his study to the literature on the probit model, his model was nicknamed the Tobit model (Tobin's probit). If there are no censored observations, $E(y_i) = E(y_i^*) = \boldsymbol{\beta}'\mathbf{x}_i$ and a classical regression model can be applied. However, with censored observations, $E(y_i)$ is no longer equal to $\boldsymbol{\beta}'\mathbf{x}_i$. Restricting our attention to non-censored observations, we get

$$E(y_i|y_i > 0) = \boldsymbol{\beta}'\mathbf{x}_i + E(\varepsilon_i|\varepsilon_i > -\boldsymbol{\beta}'\mathbf{x}_i) = \boldsymbol{\beta}'\mathbf{x}_i + \sigma\frac{\phi_i}{\Phi_i} \tag{15.15}$$

where ϕ_i and Φ_i are the density function and distribution function of the standard normal evaluated at $\boldsymbol{\beta}'\mathbf{x}_i/\sigma$. We can see that $\mathbf{x}_i$ is correlated with ϕ_i/Φ_i because ϕ_i and Φ_i are both functions of $\mathbf{x}_i$ and hence if we run an OLS regression as a function just of $\mathbf{x}_i$, we obtain biased and inconsistent results due to omitted variables bias. If we use all observations, we get

$$\begin{aligned} E(y_i) &= P(y_i > 0)E(y_i|y_i > 0) + P(y_i = 0)E(y_i|y_i = 0) \\ &= \Phi_i(\boldsymbol{\beta}'\mathbf{x}_i + \sigma\frac{\phi_i}{\Phi_i}) + (1 - \Phi_i)(0) = \Phi_i\boldsymbol{\beta}'\mathbf{x}_i + \sigma\phi_i \end{aligned} \tag{15.16}$$

Still, an OLS regression will yield biased results because $\mathbf{x}_i$ is correlated with ϕ_i.

The Type I Tobit Model can be estimated using maximum likelihood (Wooldridge 2002, pp. 525–527). This can be done in SAS using Proc LIFEREG or QLIM.

An important extension of the Type I Tobit is to model the *process* by which a customer purchases at the level c (0) or greater. For example, we may want to model which types of customers are likely to buy in a given month, and if so, how much do they spend. This can be formulated as follows:

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \tag{15.17a}$$

$$y_i = \begin{cases} 0 & if\ y_i^* \leq 0 \\ y_i^* & if\ y_i^* > 0 \end{cases} \tag{15.17b}$$

$$z_i^* = \alpha' w_i + u_i \tag{15.17c}$$

$$z_i = \begin{cases} 1 & if\ z_i^* > 0 \\ 0 & if\ z_i^* \leq 0 \end{cases} \tag{15.17d}$$

The probit model (Equations 15.17c and 15.17d) determines whether the customer buys in a given period, and if so, Equations 15.17a and 15.17b determine how much the customer spends. Note that expenditures are observed only if the customer buys, but we acknowledge through the Type I Tobit component that when a customer buys, he or she must spend at least \$0. Crucial to the formulation is that the error term u_i of the probit is correlated with the error term of the Type I Tobit (ε_i). This introduces additional "selectivity bias" into the estimation of Equation 15.17a if not accounted for.

This model is estimable in LIMDEP using a two-stage maximum likelihood procedure (Greene 2002, p. E23.18).

A variation of Equation 15.17 is not to include the censoring restriction 15.17b (e.g., see Greene 2002, p. 710). This might be applicable if the dependent variable can be positive or negative, such as the case if we are looking at customer profitability. Many authors refer to this as a Type II Tobit model (e.g., Wooldridge 2002, p. 562). This model can also be estimated within LIMDEP (Greene 2002, pp. E23-1–E23-5).

Another related model is where there is selectivity, but data are observed for all customers, so the selection variable can be an independent variable in the regression model. The example would be the case where we wanted to determine if Internet usage affects customer profitability, but wanted to recognize that only certain customers "self-select" into using the Internet. The model would be:

$$Y_i = \beta' X_i + \delta z_i + \varepsilon_i \tag{15.18a}$$

$$z_i^* = \alpha' w_i + u_i \tag{15.18b}$$

$$z_i = \begin{cases} 1 & if\ z_i^* > 0 \\ 0 & if\ z_i^* \leq 0 \end{cases} \tag{15.18c}$$

This is a recursive model (w determines whether the customer uses the Internet via Equations 15.18b and 15.18c, and then using the Internet determines customer profitability via Equation 15.18a. The only difference between this model and a standard linear recursive model is that whereas traditional recursive models consist of two or more linear equations, each one feeding into the next, the case here is a "mixed" recursive model, where one equation is a nonlinear discrete variable model, a probit, and the second equation is a linear model. OLS estimation of Equation 15.18a will be biased if ε and u are correlated, because that will set up a correlation between z and ε. Greene (2002, p. E23-14) describes this model and how to estimate it within LIMDEP.

The above models (Equations 15.16–15.18) are very relevant for database marketing, but there are few applications to date. An exception is the study by Reinartz et al. (2005) which is discussed in Chapter 26. They present a variation of Equation 15.17, where the probit model determines the acquisition (or selection) process and two (censored) regressions characterize relationship duration and customer profitability. They use their model to balance resources between customer acquisition and retention efforts.

15.5 Time Duration (Hazard) Models

What is the probability that a customer in a telecommunication company will remain as a customer after a year? Or what is the attrition probability of

each customer in a month? Are attrition probabilities different depending on the customer's demographic characteristics? What is the expected duration of a customer's relationship with the firm? These questions can be addressed using time duration models, specifically the statistical technique called hazard modeling. We first discuss the characteristics of duration data. Then we discuss and criticize a traditional approach such as a logit model to analyze duration data. Finally, we introduce the hazard model specially designed to analyze duration data.

15.5.1 Characteristics of Duration Data

In order to understand the characteristics of duration data, let us consider an example of ABC newspaper. It has a database of its subscriber lists and keeps the records of subscribers who are or have been customers at least a month for the last 7 years. We randomly select 1,000 subscribers out of the database to study their purchase behavior. Figure 15.2 displays how long some of these customers have subscribed the ABC newspaper. We can exactly calculate the duration of subscription for some customers including Customers 1, 2 and 3 because they no longer subscribe the ABC newspaper. On the other hand, those customers such as Customers 4 and 1,000 still subscribe to the ABC newspaper. The duration information provided by these current customers is incomplete. We know when they began to subscribe, but we do not know when they stop. For example, we know that customer 4 has subscribed to the ABC newspaper for a year so far, but we do not exactly know how longer she will stay. The data are right-censored.

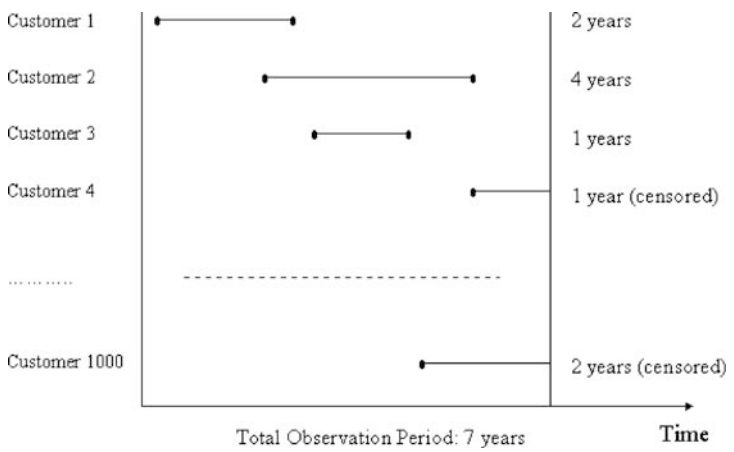

Fig. 15.2 Duration data for the ABC newspaper

The ABC newspaper recognizes that the concept of customer's lifetime value is important for its successful operations and customer retention is a critical component of customer's lifetime value. Hence, the ABC newspaper develops a model to understand factors determining a customer's subscription duration and forecasts how long each of the current and potential customers will stay. Based on this analysis, the ABC newspaper plans to develop various acquisition and retention strategies. The question is, what is the appropriate model?

15.5.2 Analysis of Duration Data Using a Classical Linear Regression

The classical regression model may be the simplest way to explain the relationship between the customer's subscription duration and her demographic characteristics (income, education, age, etc.). We might eliminate the incomplete (right-censored) observations – they are current customers – from the estimation because they will bias downward the mean duration since those observations actually have longer durations. Limiting our estimation sample only to non-current customers, we fit the regression model where the subscription duration of each customer is the dependent variable and the corresponding demographic characteristics are the independent or predictor variables.[3]

We apply the linear regression to the sample of previous customers with two customer characteristics, annual income (measured in thousand dollars) and sex (coded 1 if male and 0 if female customer). The estimated regression line is

$$\text{Subscription duration} = -1 + 0.2 \times \text{Income} + 0.5 \times \text{Sex} \tag{15.19}$$

This regression line allows us to predict the expected subscription duration of a customer given his or her income and sex. The expected subscription duration for a female customer with the annual income of \$20,000 is predicted to be 3 years ($= -1 + 0.2 \times 20 + 0.5 \times 0$). On the other hand, a male customer with the annual income of \$30,000 is predicted to maintain his subscription for 5.5 years ($= -1 + 0.2 \times 30 + 0.5 \times 1$) on average. In addition, the regression line indicates that there is a positive relationship between subscription duration and annual income. With the same income, a male would stay longer than a female by 1/2 year.

[3] The dataset of complete observations may result in sample selection bias when the characteristics of the complete observations are different from those of the right-censored observations. For example, current customers may be satisfied with the ABC newspaper, but previous customers may have switched to other newspapers. As a result, current customers may have longer subscription durations than previous customers.

The application of the regression line to current customers allows us to forecast how much longer they will remain with the ABC newspaper. Suppose that Customer 4 in Figure 15.2 is female with the annual income of $30,000. Her predicted duration is 5 years. Since she has subscribed for a year so far, we expect that she will remain four more years with the ABC newspaper. Moreover, we can use the results of the regression model to target potential customers. Computing the predicted duration for each of potential customers, we rank order them in terms of their subscription durations. Recognizing that a customer with the longer duration will generate bigger profits for the ABC newspaper, we only select customers with predicted duration greater than a specified cutoff.

However, a regression model has several limitations in analyzing duration data (Helsen and Schmittlein 1993). First, because of potential censoring bias, the regression model is not applied to all customers, but only those for whom we have observed their full lifespan. It will be problematic especially when the number of complete observations is small relative to the number of incomplete observations. Second, a regression model is very limited in helping marketers to manage the customer relationship. For example, marketers may want to know the attrition *probability* for each customer during the specified period of time. A bank prepares a special promotion to target customers who have high attrition probability during the upcoming month. A regression model cannot easily answer this question.[4] For more discussion on the limitation of a regression in the analysis of duration data, see Kalbfleisch and Prentice (1980) or Lawless (2003).

15.5.3 Hazard Models

Recently researchers from various disciplines have devoted considerable attention to the analysis of duration, survival time, lifetime, or failure time data. Engineers would like to know how long a light bulb lasts under various conditions. Medical researchers want to know how long an AIDS infected patient will live. Economists are interested in knowing the duration of unemployment. Subscription managers in newspaper want to know how long customers will subscribe to their newspapers. Customer managers at Verizon may have interest in knowing how long their customers will stay with Verizon before they switch to other carriers.

[4] We can partially overcome this limitation by dividing the observation period into even intervals. For each interval, a customer indicator (1 if a customer stays, and 0 if she does not) is used as a dependent variable. Because of its discrete nature, we now apply the binary logit or probit. However, the logit or probit has other shortcomings in modeling duration time such as arbitrarily determined time intervals. See Helsen and Schmittlein (1993) for more discussion.

Hazard models are specially designed to analyze duration data. Helsen and Schmittlein (1993) note several advantages over traditional tools such as a linear regression and discrete time probit or logistic regression in the analysis of duration data. Similar to the example given in the previous section, the variable of interest is the length of time that elapses from the beginning of an event either to the end of the event (for uncensored data) or the end of the observation period (for censored data). For the example in Fig. 15.2, we have observations, $t_1, t_2, \ldots, t_{1000}$ with $t_1 = 2, t_2 = 4, \ldots, t_{1000} = 2$. Note that the starting time of the event can be different across observations. Note also that we often have information on customer characteristics that will be related to the observed duration, $t_i (i = 1, \ldots, 1000)$. These characteristics are typically demographics such as family size, income and marital status, but hazard models can also include time-varying "covariates" such as purchase recency or the timing of previous marketing contacts.

Let us define T to be a random variable representing the duration time of interest (e.g., the time the customer stays with the company) and its (continuous) probability density function be $f(t)$, where t is a realization of T. Then its cumulative density function, or the probability the customer leaves the company before period t, is

$$F(t) = \int_0^t f(s)ds = P(T \leq t) \tag{15.20}$$

The survival function is defined as the probability that the length of the duration is at least t. That is, the survival function is $S(t) = 1 - F(t) = P(T > t)$. Researchers prefer directly to model the hazard function to the probability density function because of its mathematical convenience. The hazard rate is the probability that the event occurs at t, given that it has not occurred until t. That is, the hazard rate is a kind of a conditional probability. For an example in life insurance, the hazard rate measures the probability that a customer cancels the policy between periods t and $t + \Delta t$, where Δt is a short period of time, given that she maintains the policy for t years. Formally defined, the hazard rate is

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)/P(T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{[F(t + \Delta t) - F(t)]/S(t)}{\Delta t} = \frac{f(t)}{S(t)} \end{aligned} \tag{15.21}$$

Note that $-d \ln S(t)/dt = -[dS(t)/dt]/S(t) = -[-f(t)]/S(t) = h(t)$. The term, $-\ln S(t) = \Lambda(t)$, is called the integrated hazard function. It equals $\int_0^t h(s)ds$ since $-d\Lambda(t)/dt = h(t)$. Then the survival function $S(t) = \exp[-\Lambda(t)]$.[5]

[5] This equation is useful to calculate the survival probability (up to T^*) for current customers once we have estimated the parameters of the hazard function.

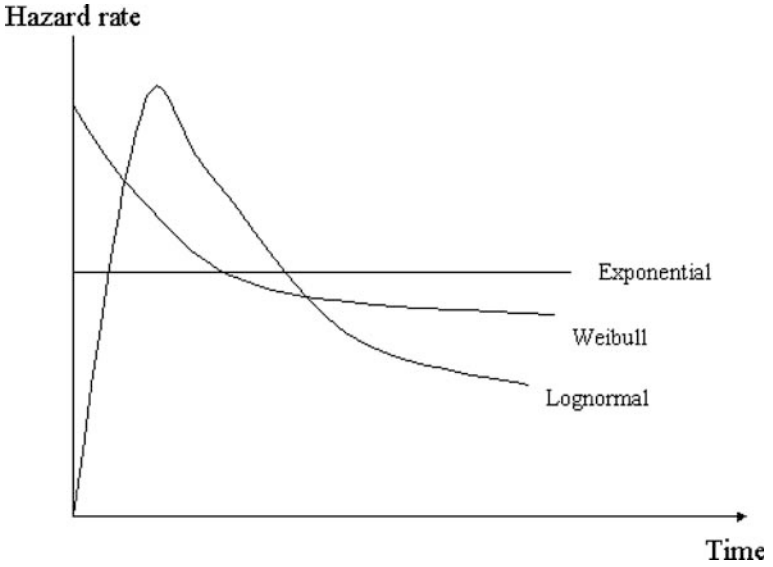

Fig. 15.3 Various hazard functions.

Hence, the density function $f(t)$ can be written as

$$f(t) = h(t)S(t) = h(t) \exp\left[-\int_0^t h(s)ds\right] \tag{15.22}$$

How do we model the hazard rate $h(t)$? The simplest model may be to assume $h(t)$ is a constant h_0. This model implies that the hazard rate is constant as t increases as shown on Fig. 15.3. With a constant hazard rate, the probability that a customer cancels the insurance policy given she maintains the policy for 1 month is the same as the probability that a customer cancels the policy given she maintains the policy for 10 years. Substituting $h(t) = h_0$ into the Equation 15.22, we can derive $f(t)$ that equals to $h_0 e^{-h_o t}$. That is, the probability density function corresponding to a constant hazard rate is an exponential distribution for duration time. The exponential distribution has a memoryless property that leads to the constant hazard rate.

For many applications, the constant hazard function is too restrictive. A natural extension allowing for monotonically increasing or decreasing hazard is to assume that $h(t) = \beta_0 + \beta_1 t$ where β_0 and β_1 are parameters to be estimated. If β_1 is zero, the model goes back to the constant hazard model. If β_1 is positive as shown in Fig. 15.3, the hazard rate is increasing with t. The case is said to have positive duration dependence. We occasionally find positive duration dependence in the analysis of shopping data. A shopper tends to have a low probability of repurchasing the product immediately after she buys it. The probability will increase as t increases because household inventory is depleting. On the other hand, if β_1 is negative, the hazard rate is monotonically decreasing with t. This negative duration dependence frequently occurs in the analysis of direct marketing data.

The exponential, Weibull, and log-logistic are the three most popular (parametric) hazard functions among researchers. Their probability density functions, hazard functions and survivor functions are:

	Probability density function: $f(t)$	**Hazard function: $h(t)$**	**Survival function: $S(t)$**
Exponential distribution	$\lambda \exp(-\lambda t)$	λ	$\exp(-\lambda t)$
Weibull distribution	$\lambda p(\lambda t)^{p-1} \exp[-(\lambda t)^p]$	$\lambda p(\lambda t)^{p-1}$	$\exp[(-\lambda t)^p]$
Log-logistic distribution	$\frac{\lambda p(\lambda t)^{p-1}}{[1+(\lambda t)^p]^2}$	$\frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}$	$\frac{1}{1+(\lambda t)^p}$

The shapes of these hazards are shown in Fig. 15.3. The exponential has a constant hazard while the Weibull can be monotonically increasing or decreasing function depending on the value of p. On the other hand, the log-logistic assumes that the hazard rate is monotonically increasing at the beginning and then decreasing later (Kalbfleisch and Prentice 1980).

The parameters of the hazard model can be estimated by maximum likelihood. Given the duration data of n samples, $t_1, t_2, \ldots, t_n$, the log-likelihood function can be written as

$$\ln L = \sum_{uncensored \;\; obs} \ln f(t|\theta) + \sum_{censored \;\; obs} \ln S(t|\theta) = \sum_{uncensored \;\; obs} h(t|\theta) + \sum_{all \;\; obs} \ln S(t|\theta) \quad (15.23)$$

Note that the only information available for the right-censored observations is the survivor rate. The above log-likelihood function is highly nonlinear so that an iterative search algorithm such as the BHHH method is used to find the optimal solution (Berndt et al. 1974). Hazard models can be estimated in SAS using either of two procedures. PROC PHREG may be more popular since it can handle time-varying covariates (e.g., marketing) and various forms of hazard functions. However, if the shapes of survival distribution and hazard function are known, PROC LIFEREG produces more efficient estimates with faster speed.

15.5.4 Incorporating Covariates into the Hazard Function

There are several approaches to incorporating independent variables in a hazard model. First, we can model the parameter of the parametric hazard rate $h(t)$ to be a function of independent variables. For example, the parameter λ of the exponential and the Weibull hazard is modeled as

$$\lambda_i = \exp(-\boldsymbol{\beta}' \mathbf{X}_i) \quad (15.24)$$

Table 15.2 Parameter estimates and hazard rates of a customer churn model (Modified from Van den Poel and Lariviere (2004)

Independent variables	Estimates ($\boldsymbol{\beta}$)	Relative hazard rate[a]
Interpurchase time	0.048	4.9
Product ownership	−6.856	99.9
Age	−0.022	2.2
Gender	0.879	140.8
Education level	−0.085	8.2
High social status	−0.593	44.7

[a]Relative hazard rate is calculated by $100 \times [\exp(\boldsymbol{\beta}) - 1]$.

where $\mathbf{X}_i$ is the vector of independent variables for observation i and $\boldsymbol{\beta}$ is the corresponding parameter vector. Cox (1972) has proposed a more flexible method called the proportional hazard model. He defines the hazard rate of observation i, $h_i(t\,|\underline{X})$ as

$$h_i(t\,|\underline{X}) = h_0(t)\psi_i(\underline{X}) = h_0(t)\exp(-\boldsymbol{\beta}'\mathbf{X}_i) \tag{15.25}$$

where $h_0(t)$ is the baseline hazard rate and $\psi_i(\underline{X})$ incorporates covariates (or independent variables) that may be time-varying. The baseline hazard is the hazard rate that describes the relationship between the hazard rate and time duration, and can specified as constant, exponential, Weibul, etc., as discussed above (see also Seetharaman and Chintagunta 2003).

Going back to the example of ABC newspaper, we employ a proportional hazard model to the duration data with two independent variables, income and sex. Allowing for monotonically increasing or decreasing hazards, we estimated the hazard function incorporating independent variables as

$$h_i(t\,|\underline{X}) = h_0(t)\psi_i(\underline{X}) = h_0(t)\ \exp(-\beta_2 \times \text{Income} - \beta_3 \times \text{sex}) \tag{15.26}$$

Upon estimation, we can evaluate from the β coefficients how each independent variable influences the hazard rate: $100 \times [\exp(\boldsymbol{\beta}) - 1]$ measures the percentage change of the hazard rate with respect to the unit change of the independent variable (Tuma and Hannan 1984).

We conclude this section with a real application of the hazard model provided by Poel and Larivière (2004). They applied the proportional hazard model to data from a European financial services company that offers banking and insurance services towards customers. The data set consists of a random sample of 47,157 customers, of whom 47% churned (uncensored sample) and the rest are current customers (censored sample). They considered various categories of independent variable to explain retention (or churn), but we report some of their estimates for an expositional purpose in Table 15.2.

Table 15.2 shows that customers whose interpurchase time increases experience shorter duration time. Every additional year in the average interpurchase time is associated with a $100 \times [\exp(0.048) - 1] = 4.9\%$ higher probability to churn. The more products owned by a customer the more likely she

is to stay with the company. An increase of one additional product lowers the switching probability with 99.9% (e.g., $100 \times [\exp(-6.856) - 1] = -99.9$). Older people are less inclined to leave the company. As a customer's age increases by one, her probability to leave decreases by 2.2%. Men (coded as 1) are 141% more likely to leave the company than females. More educated people have a somewhat (8.2%) lower attrition probability. Finally, customers with a high social status have a significantly lower attrition probability than customers who live in an area that is associated with a low social status.