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 \( p_i = F(\beta'X_i) \) where \( X_i \) is a vector of independent variables for customer \( i \) (e.g., customer’s credit-related variables) and \( \beta \) is a corresponding parameter vector. Then \( E(Y_i) = (1)(p_i) + (0)(1 - p_i) = p_i = F(\beta'X_i) \).

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, \( p_i = F(\beta'X_i) = \beta'X_i \).

Now since \( E(Y_i|X_i) = \beta'X_i \), we can derive the following linear probability model.

\[
Y_i = \beta'X_i + \varepsilon_i \tag{15.2}
\]

where \( \varepsilon_i \) 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