Chapter 11. Categorical Dependent Variables and Survival Models

Abstract. Extending Chapter 9, this chapter considers dependent variables having more than two possible categorical alternatives. As in Chapter 9, we often interpret a categorical variable to be an attribute possessed or choice made by an individual, household or firm. By allowing more than two alternatives, we substantially broaden the scope of applications to complex social science problems of interest.

The pedagogic approach of this chapter follows the pattern established in earlier chapters; we begin with homogeneous models in Section 11.1 followed by the Section 11.2 model that incorporates random effects, thus providing a heterogeneity component. Section 11.3 introduces an alternative method for accommodating time patterns through transition, or Markov, models. Although transition models are applicable in the Chapter 10 generalized linear models, they are particularly useful in the context of categorical dependent variables. Many repeated applications of the idea of transitioning gives rise to survival models, another important class of longitudinal models. Section 11.4 develops this link.

11.1 Homogeneous models

We now consider a response that is an unordered categorical variable. We assume that the dependent variable \( y \) may take on values \( 1, 2, \ldots, c \), corresponding to \( c \) categories. We first introduce the homogenous version so that this section does not use any subject-specific parameters nor does it introduce terms to account for serial correlation.

In many social science applications, the response categories correspond to an attribute possessed or choices made by individuals, households or firms. Some applications of categorical dependent variable models include:

- employment choice, such as Valletta (1999E),
- mode of transportation choice, such as the classic work by McFadden (1978E),
- choice of political party affiliation, such as Brader and Tucker (2001O) and
- marketing brand choice, such as Jain et al. (1994O).

Example 11.1 Political party affiliation

Brader and Tucker (2001O) studied the choice made by Russian voters of political parties in 1995-1996 elections. Their interest was in assessing effects of meaningful political party attachments during Russia’s transition to democracy. They examined a \( T = 3 \) wave survey of \( n = 2,841 \) respondents, taken: (1) three to four weeks before the 1995 parliamentary elections, (2) immediately after the parliamentary elections and (3) after the 1996 presidential elections. This survey design was modeled on the American National Election Studies (see Appendix F).

The dependent variable was the political party voted for, consisting of \( c = 10 \) parties including the Liberal Democratic Party of Russia, Communist Party of the Russian Federation, Our Home is Russia, and others. The independent variables included social characteristics such as education, gender, religion, nationality, age, and location (urban versus rural), economic
characteristics such as income and employment status and political attitudinal characteristics such
attitudes toward market transitions and privatization.

11.1.1 Statistical inference

For an observation from subject $i$ at time $t$, denote the probability of choosing the $j$th
category as $\pi_{it,j} = \text{Prob}(y_{it} = j)$, so that $\pi_{it,1} + \ldots + \pi_{it,c} = 1$. In the homogeneous framework, we
assume that observations are independent from one another.

In general, we will model these probabilities as a (known) function of parameters and use
maximum likelihood estimation for statistical inference. We use the notation $y_{itj}$ to be an indicator
of the event $y_{it} = j$. The likelihood for the $i$th subject at the $t$th time point is:

$$
\prod_{j=1}^{c} (\pi_{it,j})^{y_{itj}} = \begin{cases}
\pi_{it,1} & \text{if } y_{it} = 1 \\
\pi_{it,2} & \text{if } y_{it} = 2 \\
\vdots & \vdots \\
\pi_{it,c} & \text{if } y_{it} = c
\end{cases}
$$

Thus, assuming independence among observations, the total log-likelihood is

$$
L = \sum_{i} \sum_{j=1}^{c} y_{itj} \ln \pi_{it,j} .
$$

With this framework, standard maximum likelihood estimation is available. Thus, our main task
is to specify an appropriate form for $\pi$.

11.1.2 Generalized logit

Like standard linear regression, generalized logit models employ linear combinations of
explanatory variables of the form:

$$
V_{it,j} = x_{it} \beta_{j} .
$$

(11.1)

Because the dependent variables are not numerical, we cannot model the response $y$ as a linear
combination of explanatory variables plus an error. Instead we use the probabilities

$$
\text{Prob}(y_{it} = j) = \pi_{it,j} = \frac{\exp(V_{it,j})}{\sum_{k=1}^{c} \exp(V_{it,k})} , \quad j = 1, 2, \ldots, c
$$

(11.2)

Note here that $\beta_{j}$ is the corresponding vector of parameters that may depend on the alternative, or
choice, whereas the explanatory variables $x_{it}$ do not. So that probabilities sum to one, a
convenient normalization for this model is $\beta_{c} = 0$. With this normalization and the special case of
c $= 2$, the generalized logit reduces to the logit model introduced in Section 9.1.

Parameter interpretations

We now describe an interpretation of coefficients in generalized logit models, similar to
Section 9.1.1 for the logistic model. From equations (11.1) and (11.2), we have

$$
\ln \left( \frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = c)} \right) = V_{it,j} - V_{it,c} = x_{it}' \beta_{j} .
$$

The left-hand side of this equation is interpreted to be the logarithmic odds of choosing choice $j$
compared to choice $c$. Thus, as in Section 9.1.1, we may interpret $\beta_{j}$ as the proportional change in
the odds ratio.
Generalized logits have an interesting nested structure that we will explore briefly in Section 9.1.5. That is, it is easy to check that conditional on not choosing the first category, the form of Prob($y_{it} = j \mid y_{it} \neq 1$) has a generalized logit form in equation (11.2). Further, if $j$ and $h$ are different alternatives, we note that

$$\frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = j) + \text{Prob}(y_{it} = h)} = \frac{\exp(V_{it,j})}{\exp(V_{it,j}) + \exp(V_{it,h})} = \frac{1}{1 + \exp(x_t'(\beta_h - \beta_j))}.$$ 

This has a logit form that was introduced in Section 9.1.1.

**Special case – Intercept only model**

To develop intuition, we now consider the model with only intercepts. Thus, let $x_{it} = 1$ and $\beta_j = \beta_{0,j} = \alpha_j$. With the convention $\alpha_c = 0$, we have

$$\text{Prob}(y_{it} = j) = \pi_{it,j} = \frac{e^{\alpha_j}}{e^{\alpha_1} + e^{\alpha_2} + ... + e^{\alpha_c} + 1}$$

and

$$\ln \left( \frac{\text{Prob}(y_{it} = j)}{\text{Prob}(y_{it} = c)} \right) = \alpha_j.$$ 

From the second relation, we may interpret the $j$th intercept $\alpha_j$ to be the logarithmic odds of choosing alternative $j$ compared to alternative $c$.

**Example 11.2 Job security**

This is a continuation of the Example 9.1 on the determinants of job turnover, based on the work of Valetta (1999E). The Chapter 9 analysis of this data considered only the binary dependent variable dismissal, the motivation being that this is the main source of job insecurity. Valetta (1999E) also presented results from a generalized logit model, his primary motivation being that the economic theory describing turnover implies that other reasons for leaving a job may affect dismissal probabilities.

For the generalized logit model, the response variable has $c = 5$ categories: dismissal, left job because of plant closures, “quit,” changed jobs for other reasons and no change in employment. The latter category is the omitted one in Table 11.1. The explanatory variables of the generalized logit are same as the probit regression described in Example 9.1; the estimates summarized in Example 9.1 are reproduced here for convenience.

Table 11.1 shows that turnover declines as tenure increases. To illustrate, consider a typical man in the 1992 sample where we have $t = 16$ and focus on dismissal probabilities. For this value of time, the coefficient associated with tenure for dismissal is $-0.221 + 16 \times (0.008) = -0.093$ (due to the interaction term). From this, we interpret an additional year of tenure to imply that the dismissal probability is $\exp(-0.093) = 91\%$ of what it would be otherwise, representing a decline of 9%.

Table 11.1 also shows that the generalized coefficients associated with dismissal are similar to the probit fits. The standard errors are also qualitatively similar, although higher for the generalized logits when compared to the probit model. In particular, we again see that the
coefficient associated with the interaction between tenure and time trend reveals an increasing dismissal rate for experienced workers. The same is true for the rate of quitting.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probit Regression Model</th>
<th>Generalized Logit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dismissed</td>
<td>Plant Closed</td>
</tr>
<tr>
<td>Tenure</td>
<td>-0.084</td>
<td>-0.221</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Time Trend</td>
<td>-0.002</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Tenure × (Time Trend)</td>
<td>0.003</td>
<td>0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Logarithmic Sector Employment</th>
<th>Probit Regression Model</th>
<th>Generalized Logit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dismissed</td>
<td>Plant Closed</td>
</tr>
<tr>
<td>Change in Logarithmic Sector Employment</td>
<td>0.094</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.123)</td>
</tr>
<tr>
<td>Tenure × (Change in Logarithmic Sector Employment)</td>
<td>Probit Regression Model</td>
<td>Generalized Logit Model</td>
</tr>
<tr>
<td></td>
<td>Dismissed</td>
<td>Plant Closed</td>
</tr>
<tr>
<td>Tenure × (Change in Logarithmic Sector Employment)</td>
<td>0.020</td>
<td>-0.061</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.023)</td>
</tr>
</tbody>
</table>

Notes: Standard errors in parentheses. Omitted category is no change in employment. Other variables controlled for consisted of education, marital status, number of children, race, years of full-time work experience and its square, union membership, government employment, logarithmic wage, the U.S. employment rate and location.

### 11.1.3 Multinomial (conditional) logit

Similar to equation (11.1), an alternative linear combination of explanatory variables is

\[ V_{it,j} = x_{it,j}' \beta. \]  (11.3)

Here, \( x_{it,j} \) is a vector of explanatory variables that depends on the \( j \)th alternative, whereas the parameters \( \beta \) do not. Using the expression in equation (11.2) is the basis for the multinomial logit model, also known as the conditional logit model. With this specification, the total log-likelihood is

\[ L(\beta) = \sum_{i=1}^{c} \sum_{j=1}^{c} y_{it,j} \ln \pi_{it,j} = \sum_{i=1}^{c} \left( \sum_{j=1}^{c} y_{it,j} x_{it,j}' \beta - \ln \left( \sum_{k=1}^{c} \exp(x_{it,k}' \beta) \right) \right). \]  (11.4)

This straightforward expression for the likelihood enables maximum likelihood inference to be easily performed.

The generalized logit model is a special case of the multinomial logit model. To see this, consider explanatory variables \( x_{it} \) and parameters \( \beta_{t} \) each of dimension \( K \times 1 \). Define

\[
\begin{pmatrix}
0 \\
\vdots \\
0 \\
\vdots \\
0 \\
\end{pmatrix}
= x_{it} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_{1} \\ \beta_{2} \\ \vdots \\ \beta_{c} \end{pmatrix}.
\]
Specifically, $x_{it,j}$ is defined as $j-1$ zero vectors (each of dimension $K \times 1$), followed by $x_t$ and then followed by $c-j$ zero vectors. With this specification, we have $x'_{it,j} \beta = x'_t \beta_j$. Thus, a statistical package that performs multinomial logit estimation can also perform generalized logit estimation through the appropriate coding of explanatory variables and parameters. Another consequence of this connection is that some authors use the descriptor multinomial logit when referring to the Section 11.1.2 generalized logit model.

Moreover, through similar coding schemes, multinomial logit models can also handle linear combinations of the form:

$$V_{it,j} = x'_{it,1,j} \beta + x'_{it,2} \beta_j. \quad (11.5)$$

Here, $x_{it,1,j}$ are explanatory variables that depend on the alternative whereas $x_{it,2}$ do not. Similarly, $\beta_j$ are parameters that depend on the alternative whereas $\beta$ do not. This type of linear combination is the basis of a mixed logit model. As with conditional logits, it is customary to choose one set of parameters as the baseline and specify $\beta_c = 0$ to avoid redundancies.

To interpret parameters for the multinomial logit model, we may compare alternatives $h$ and $k$ using equations (11.2) and (11.3), to get

$$\ln \left( \frac{\text{Prob}(y_{it} = h)}{\text{Prob}(y_{it} = k)} \right) = (x_{it,h} - x_{it,k})' \beta. \quad (11.6)$$

Thus, we may interpret $\beta_j$ as the proportional change in the odds ratio, where the change is the value of the $j$th explanatory variable, moving from the $k$th to the $h$th alternative.

With equation (11.2), note that $\pi_{it,1} / \pi_{it,2} = \exp(V_{it,1}) / \exp(V_{it,2})$. This ratio does not depend on the underlying values of the other alternatives, $V_{it,j}$, for $j = 3, \ldots, c$. This feature, called the independence of irrelevant alternatives, can be a drawback of the multinomial logit model for some applications.

**Example 11.3 Choice of yogurt brands**

We now consider a marketing data set introduced by Jain et al. (1994) that was further analyzed by Chen and Kuo (2001). These data, obtained from A. C. Nielsen, are known as scanner data because they are obtained from optical scanning of grocery purchases at check-out. The subjects consist of $n=100$ households in Springfield, Missouri. The response of interest is the type of yogurt purchased, consisting of four brands: Yoplait, Dannon, Weight Watchers and Hiland. The households were monitored over a two-year period with the number of purchases ranging from 4 to 185; the total number of purchases is $N=2,412$.

The two marketing variables of interest are PRICE and FEATURES. For the brand purchased, PRICE is recorded as price paid, that is, the shelf price net of the value of coupons redeemed. For other brands, PRICE is the shelf price. FEATURES is a binary variable, defined to be one if there was a newspaper feature advertising the brand at time of purchase, and zero otherwise. Note that the explanatory variables vary by alternative, suggesting the use of a multinomial (conditional) logit model.

Tables 11.2 and 11.3 summarize some important aspects of the data. Table 11.2 shows that Yoplait was the most frequently (33.9%) selected type of yogurt in our sample whereas Hiland was the least frequently selected (2.9%). Yoplait was also the most heavily advertised, appearing in newspaper advertisements 5.6% of the time that the brand was chosen. Table 11.3 shows that Yoplait was also the most expensive, costing 10.6 cents per ounce, on average. Table 11.3 also shows that there are several prices that were far below the average, suggesting some potential influential observations.
TABLE 11.2 Summary Statistics by Choice of Yogurt

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Choices</td>
<td>818</td>
<td>970</td>
<td>553</td>
<td>71</td>
<td>2,412</td>
</tr>
<tr>
<td>Number of Choices, in Percent</td>
<td>33.9</td>
<td>40.2</td>
<td>22.9</td>
<td>2.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Feature Averages, in Percent</td>
<td>5.6</td>
<td>3.8</td>
<td>3.0</td>
<td>3.7</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 11.3 Summary Statistics for Prices

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>0.107</td>
<td>0.108</td>
<td>0.003</td>
<td>0.193</td>
<td>0.019</td>
</tr>
<tr>
<td>Dannon</td>
<td>0.082</td>
<td>0.086</td>
<td>0.019</td>
<td>0.111</td>
<td>0.011</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>0.079</td>
<td>0.079</td>
<td>0.004</td>
<td>0.104</td>
<td>0.008</td>
</tr>
<tr>
<td>Hiland</td>
<td>0.054</td>
<td>0.054</td>
<td>0.025</td>
<td>0.086</td>
<td>0.008</td>
</tr>
</tbody>
</table>

A multinomial logit model was fit to the data, using the following specification for the systematic component

\[ V_{it,j} = \alpha_j + \beta_1 \text{PRICE}_{it,j} + \beta_2 \text{FEATURE}_{it,j}, \]

using Hiland as the omitted alternative. The results are summarized in Table 11.4. Here, we see that each parameter is statistically significantly different from zero. Thus, the parameter estimates may be useful when predicting the probability of choosing a brand of yogurt. Moreover, in a marketing context, the coefficients have important substantive interpretations. Specifically, we interpret the coefficient associated with FEATURES to suggest that a consumer is \( \exp(0.4914) = 1.634 \) times more likely to purchase a product that is featured in a newspaper ad compared to one that is not. For the PRICE coefficient, a one cent decrease in price suggests that a consumer is \( \exp(0.3666) = 1.443 \) times more likely to purchase a brand of yogurt.

Table 11.4 Yogurt Multinomial Logit Model Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimate</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>4.450</td>
<td>23.78</td>
</tr>
<tr>
<td>Dannon</td>
<td>3.716</td>
<td>25.55</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>3.074</td>
<td>21.15</td>
</tr>
<tr>
<td>FEATURES</td>
<td>0.491</td>
<td>4.09</td>
</tr>
<tr>
<td>PRICE</td>
<td>-36.658</td>
<td>-15.04</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>10,148</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>10,138</td>
<td></td>
</tr>
</tbody>
</table>

11.1.4 Random utility interpretation

In economic applications, we think of an individual choosing among \( c \) categories where preferences among categories are determined by an unobserved utility function. For the \( i \)th individual at the \( r \)th time period, use \( U_{irj} \) for the utility of the \( j \)th choice. To illustrate, assume that the individual chooses the first category if \( U_{ir1} > U_{irj} \) for \( j = 2, \ldots, c \) and denote this choice as \( y_{ir} = 1 \). We model utility as an underlying value plus random noise, that is, \( U_{irj} = V_{irj} + \epsilon_{irj} \), where \( V_{irj} \) is specified in equation (11.4). The noise variable is assumed to have an extreme-value distribution of the form

\[ F(a) = \text{Prob}(\epsilon_{irj} \leq a) = \exp(-e^{-a}) . \]
This form is computationally convenient. Omitting the observation-level subscripts \(\{it\}\) for the moment, we have

\[
\text{Prob}(y = 1) = \text{Prob}(U_i > U_j \text{ for } j = 2, \ldots, c) = \text{Prob}(\varepsilon_j < \varepsilon_i + V_i - V_j \text{ for } j = 2, \ldots, c)
\]

\[
= E\{\text{Prob}(\varepsilon_j < \varepsilon_i + V_i - V_j \text{ for } j = 2, \ldots, c | \varepsilon_i)\} = E[F(\varepsilon_i + V_i - V_j) \cdots F(\varepsilon_i + V_i - V_c)]
\]

\[
= E \exp[-(\varepsilon_i + V_i - V_j)] \cdots \exp[-(\varepsilon_i + V_i - V_c)]
\]

\[
= E \exp[k_v \exp(-\varepsilon_i)],
\]

where \(k_v = \sum_{j=2}^{c} \exp(V_j - V_1)\). Now, it is a pleasant exercise in calculus to show, with the distribution function given above, that \(E \exp[k_v \exp(-\varepsilon_i)] = \frac{1}{k_v + 1}\). Thus, we have

\[
\text{Prob}(y = 1) = \frac{1}{1 + \sum_{j=2}^{c} \exp(V_j - V_1)} \frac{\exp(V_1)}{\sum_{j=1}^{c} \exp(V_j)}.
\]

Because this argument is valid for all alternatives \(j = 1, 2, \ldots, c\), the random utility representation yields the multinomial logit model.

### 11.1.5 Nested logit

To mitigate the problem of independence of irrelevant alternatives, we now introduce a type of hierarchical model known as a nested logit model. To interpret the nested logit model, in the first stage one chooses an alternative (say the first type) with probability

\[
\pi_{it,1} = \text{Prob}(y_{it} = 1) = \frac{\exp(V_{it,1})}{\exp(V_{it,1}) + \left(\sum_{k=2}^{c} \exp(V_{it,k}/\rho)\right)^\rho}.
\]

(11.7)

Then, conditional on not choosing the first alternative, the probability of choosing the any one of the other alternatives follows a multinomial logit model with probabilities

\[
\frac{\pi_{it,j}}{1 - \pi_{it,1}} = \text{Prob}(y_{it} = j | y_{it} \neq 1) = \frac{\exp(V_{it,j}/\rho)}{\sum_{k=2}^{c} \exp(V_{it,k}/\rho)}, \quad j = 2, \ldots, c.
\]

(11.8)

In equations (11.7) and (11.8), the parameter \(\rho\) measures the association among the choices \(j = 2, \ldots, c\). The value of \(\rho = 1\) reduces to the multinomial logit model that we interpret to mean independence of irrelevant alternatives. We also interpret \(\text{Prob}(y_{it} = 1)\) to be a weighted average of values from the first choice and the others. Conditional on not choosing the first category, the form of \(\text{Prob}(y_{it} = j | y_{it} \neq 1)\) has the same form as the multinomial logit.

The advantage of the nested logit is that it generalizes the multinomial logit model in a way such that we no longer have the problem of independence of irrelevant alternatives. A disadvantage, pointed out by McFadden (1981E), is that only one choice is observed; thus, we do not know which category belongs in the first stage of the nesting without additional theory regarding choice behavior. Nonetheless, the nested logit generalizes the multinomial logit by allowing alternative “dependence” structures. That is, one may view the nested logit as a robust alternative to the multinomial logit and examine each one of the categories in the first stage of the nesting.
11.1.6 Generalized extreme value distribution

The nested logit model can also be given a random utility interpretation. To this end, return to the random utility model but assume that the choices are related through a dependence within the error structure. McFadden (1978E) introduced the generalized extreme-value distribution of the form:

\[ F(a_1, ..., a_c) = \exp\left(-G(e^{-a_1}, ..., e^{-a_c})\right). \]

Under regularity conditions on \( G \), McFadden (1978E) showed that this yields

\[ \text{Prob}(y = j) = \text{Prob}(U_j > U_k \text{ for } k = 1, ..., c, k \neq j) = \frac{e^{y_j} G_j(e^{-v_j}, ..., e^{-v_c})}{G(e^{-v_1}, ..., e^{-v_c})}, \]

where \( G_j(x_1, ..., x_c) = \frac{\partial}{\partial x_j} G(x_1, ..., x_c) \) is the \( j \)th partial derivative of \( G \).

**Special cases**

1. Let \( G(x_1, ..., x_c) = x_1 + ... + x_c \). In this case, \( G_j = 1 \) and \( \text{Prob}(y = j) = \frac{\exp(V_j)}{\sum_{k=1}^c \exp(V_k)} \). This is the multinomial logit case.

2. Let \( G(x_1, ..., x_c) = x_1 + \left( \sum_{k=2}^c x_k^{1/\rho} \right)^\rho \). In this case, \( G_1 = 1 \) and

\[ \text{Prob}(y = 1) = \frac{\exp(V_1)}{\exp(V_1) + \left( \sum_{k=2}^c \exp(V_k / \rho) \right)^\rho}. \]

Additional calculations show that

\[ \text{Prob}(y = j \mid y \neq 1) = \frac{\exp(V_j / \rho)}{\sum_{k=2}^c \exp(V_k / \rho)}. \]

This is the nested logit case.

Thus, the generalized extreme-value distribution provides a framework that encompasses the multinomial and conditional logit models. Amemiya (1985E) provides background on more complex nested models that utilize the generalized extreme-value distribution.

11.2 Multinomial logit models with random effects

Repeated observations from an individual tend to be similar; in the case of categorical choices, this means that individuals tend to make the same choices from one observation to the next. This section models that similarity through a common heterogeneity term. To this end, we augment our systematic component with a heterogeneity term and, similar to equation (11.4), consider linear combinations of explanatory variables of the form

\[ V_{it,j} = z_{it,j} a_i + x'_{it,j} \beta. \]  

(11.9)
As before, \( a_i \) represents the heterogeneity term that is subject-specific. The form of equation (11.9) is quite general and includes many applications of interest. However, to develop intuition, we focus on the special case

\[
V_{it,j} = a_{ij} + x'_{it,j} \beta .
\]

(11.10)

Here, intercepts vary by individual and alternative but are common over time.

With this specification for the systematic component, the conditional (on the heterogeneity) probability that the \( i \)th subject at time \( t \) chooses the \( j \)th alternative is

\[
\pi_{ij}(a_i) = \frac{\exp(V_{it,j})}{\sum_{k=1}^{c} \exp(V_{it,k})} = \frac{\exp(a_{ij} + x'_{it,j} \beta)}{\sum_{k=1}^{c} \exp(a_{ik} + x'_{it,k} \beta)}, \quad j = 1, 2, \ldots, c ,
\]

(11.11)

where we now denote the set of heterogeneity terms as \( \alpha_i = (\alpha_{i1}, \ldots, \alpha_{ic})' \). From the form of this equation, we see that a heterogeneity term that is constant over alternatives \( j \) does not affect the conditional probability. To avoid parameter redundancies, a convenient normalization is to take \( \alpha_{ic} = 0 \).

For statistical inference, we begin with likelihood equations. Similar to the development in Section 11.1.1, the conditional likelihood for the \( i \)th subject is

\[
L(y_i | a_i) = \prod_{j=1}^{c} (\pi_{ij}(a_i))^{y_{ij}} = \prod_{t=1}^{T_i} \left[ \frac{\exp(\sum_{j=1}^{c} y_{it,j} (a_{ij} + x'_{it,j} \beta))}{\sum_{k=1}^{c} \exp(\alpha_{ik} + x'_{it,k} \beta)} \right].
\]

(11.12)

We assume that \( \{a_i\} \) is i.i.d. with distribution function \( G_a \), that is typically taken to be multivariate normal. With this convention, the (unconditional) likelihood for the \( i \)th subject is

\[ L(y_i) = \int L(y_i | a)dG_a(a). \]

Assuming independence among subjects, the total log-likelihood is \( L = \sum_i L(y_i) \).

Relation with nonlinear random effects Poisson model

With this framework, standard maximum likelihood estimation is available. However, for applied work, there are relatively few statistical packages available for estimating multinomial logit models with random effects. As an alternative, one can look to properties of the multinomial distribution and link it to other distributions. Chen and Kuo (2001S) recently surveyed these linkages in the context of random effects models and we now present one link to a nonlinear Poisson model. Statistical packages for nonlinear Poisson models are readily available; with this link, they can be used to estimate parameters of the multinomial logit model with random effects.

To this end, an analyst would instruct a statistical package to “assume” that the binary random variables \( y_{it,j} \) are Poisson distributed with conditional means \( \pi_{it,j} \) and, conditional on the heterogeneity terms, are independent. This is a nonlinear Poisson model because, from Section 10.5.3, a linear Poisson model takes the logarithmic (conditional) mean to be a linear function of explanatory variables. In contrast, from equation (11.11), log \( \pi_{it,j} \) is a nonlinear function. Of course, this “assumption” is not valid. Binary random variables have only two outcomes and thus cannot have a Poisson distribution. Moreover, the binary variables must sum to one (that is, \( \sum_{j} y_{it,j} = 1 \)) and thus are not even conditionally independent. Nonetheless, with this “assumption” and the Poisson distribution (reviewed in Section 10.5.3), the conditional likelihood interpreted by the statistical package is:
Up to the constant, this is the same conditional likelihood as in equation (11.12) (see Exercise 10.1). Thus, a statistical package that performs nonlinear Poisson models with random effects can be used to get maximum likelihood estimates for the multinomial logit model with random effects. See Chen and Kuo (2001S) for a related algorithm based on a linear Poisson model with random effects.

**Example 11.3 Choice of yogurt brands - Continued**

To illustrate, we used a multinomial logit model with random effects on the yogurt data introduced in Example 11.1. Following Chen and Kuo (2001S), random intercepts for Yoplait, Dannon and Weight Watchers were assumed to follow a multivariate normal distribution with an unstructured covariance matrix. Table 11.5 shows results from fitting this model, based on the nonlinear Poisson model link and using SAS PROC NLMIXED. Here, we see that the coefficients for FEATURES and PRICE are qualitatively similar to the model without random effects, reproduced for convenience from Table 11.4. They are qualitatively similar in the sense that they have the same sign and same degree of statistical significance. Overall, the AIC statistic suggests that the model with random effects is the preferred model.

<table>
<thead>
<tr>
<th>Table 11.5 Yogurt Multinomial Logit Model Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Yoplait</td>
</tr>
<tr>
<td>Dannon</td>
</tr>
<tr>
<td>Weight Watchers</td>
</tr>
<tr>
<td>FEATURES</td>
</tr>
<tr>
<td>PRICE</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
</tr>
<tr>
<td>AIC</td>
</tr>
</tbody>
</table>

As with binary dependent variables, conditional maximum likelihood estimators have been proposed; see, for example, Conaway (1989S). Appendix 11A provides a brief introduction to these alternative estimators.

### 11.3 Transition (Markov) Models

Another way of accounting for heterogeneity is to trace the development of a dependent variable over time and represent the distribution of its current value as a function of its history. To this end, define \( H_{it} \) to be the history of the \( i \)th subject up to time \( t \). For example, if the explanatory variables are assumed to be non-stochastic, then we might use \( H_{it} = \{y_{i1}, \ldots, y_{i(t−1)}\} \). With this information set, we may partition the likelihood for the \( i \)th subject is

\[
L(y_i) = f(y_{i1}) \prod_{t=2}^{T} f(y_{it} | H_{it}),
\]
(11.13)
where $f(y_{it} | H_{it})$ is the conditional distribution of $y_{it}$ given its history and $f(y_{i1})$ is the marginal distribution of $y_{i1}$. To illustrate, one type of application is through a conditional generalized linear model (GLM) of the form

$$f(y_{it} | H_{it}) = \exp \left( \frac{y_{it}^T \theta_{it} - b(\theta_{it})}{\phi} + S(y_{it}, \phi) \right)$$

where $E(y_{it} | H_{it}) = b'(\theta_{it})$ and $\text{Var}(y_{it} | H_{it}) = \phi b''(\theta_{it})$. Assuming a canonical link, for the systematic component, one could use

$$\theta_{it} = g(E(y_{it} | H_{it})) = x_{it}^T \beta + \sum_j \phi_j y_{it,j-1}.$$ 

See Diggle, Heagerty, Liang and Zeger (2002S, Chapter 10) for further applications of and references to general transition GLMs. We focus on categorical responses.

**Unordered categorical response**

To simplify our discussion of unordered categorical responses, we also assume discrete unit time intervals. To begin, we consider Markov models of order 1. Thus, the history $H_{it}$ need only contain $y_{i,t-1}$. More formally, we assume that

$$\pi_{it,jk} = \text{Prob}(y_{it} = k | y_{i,t-1} = j) = \text{Prob}(y_{it} = k | \{y_{i,t-1} = j, y_{i,t-2}, \ldots, y_{i,1}\}).$$

That is, given the information in $y_{i,t-1}$, there is no additional information content in $\{y_{i,t-2}, \ldots, y_{i,1}\}$ about the distribution of $y_{it}$.

Without covariate information, it is customary to organize the set of transition probabilities $\pi_{it,jk}$ as a matrix of the form

$$\Pi_{it} = \begin{pmatrix}
\pi_{it,11} & \pi_{it,12} & \cdots & \pi_{it,1c} \\
\pi_{it,21} & \pi_{it,22} & \cdots & \pi_{it,2c} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{it,c1} & \pi_{it,c2} & \cdots & \pi_{it,cc}
\end{pmatrix}.$$ 

Here, each row sums to one. With covariate information and an initial state distribution $\text{Prob}(y_{i1})$, one can trace the history of the process knowing only the transition matrix $\Pi_{it}$. We call the row identifier, $j$, the state of origin and the column identifier, $k$, the destination state.

For complex transition models, it can be useful to graphically summarize the set of feasible transitions under consideration. To illustrate, Figure 11.1 summarizes an employee retirement system with $c = 4$ categories. Here,

- 1 denotes active continuation in the pension plan,
- 2 denotes retirement from the pension plan,
- 3 denotes withdrawal from the pension plan and
- 4 denotes death.

For this system, the circles represent the nodes of the graph and correspond to the response categories. The arrows, or arcs, indicate the modes of possible transitions. This graph indicates that movement from state 1 to states 1, 2, 3 or 4 is possible, so that we would assume $\pi_{ij} \geq 0$, for $j = 1, 2, 3, 4$. However, once an individual is in states 2, 3, or 4, it is not possible to move from those states (known as absorbing states). Thus, we use $\pi_{ij} = 1$ for $j = 2, 3, 4$ and $\pi_{ik} = 0$, for $j = 2, 3, 4$ and $k \neq j$. Note that although death is certainly possible (and even eventually certain) for those in retirement, we assume $\pi_{24} = 0$ with the understanding that the plan has paid pension benefits at retirements and need no longer be concerned with additional transitions after exiting.
the plan, regardless of the reason. This assumption is often convenient because it is difficult to track individuals having left active membership in a benefit plan.

![Figure 11.1 Graphical Summary of a Transition Model for a Hypothetical Employment Retirement System.](image)

For another example, consider the modification summarized in Figure 11.2. Here, we see that retirees are now permitted to re-enter the workforce so that \( \pi_{21} \) may be positive. Moreover, now the transition from retirement to death is also explicitly accounted for so that \( \pi_{24} \geq 0 \). This may be of interest in a system that pays retirement benefits as long as a retiree lives. We refer to Haberman and Pitacco (1999O) for many additional examples of Markov transition models that are of interest in employee benefit and other types of actuarial systems.

![Figure 11.2 A Modified Transition Model for a Employment Retirement System.](image)

We can parameterize the problem by choosing a multinomial logit, one for each state of origin. Thus, we use

\[
\pi_{it,jk} = \frac{\exp(V_{it,jk})}{\sum_{h=1}^{c} \exp(V_{it,jh})}, \quad j, k = 1, 2, \ldots, c, \quad (11.14)
\]

where the systematic component \( V_{it,jk} \) is given by

\[
V_{it,jk} = x_{it,jk}' \beta_j. \quad (11.15)
\]
As discussed in the context of employment retirement systems, in a given problem one assumes that a certain subset of transition probabilities are zero, thus constraining the estimation of $\beta_j$.

For estimation, we may proceed as in Section 11.1. Define

$$y_{it,jk} = \begin{cases} 1 & \text{if } y_{it} = k \text{ and } y_{it-1} = j \\ 0 & \text{otherwise} \end{cases}$$

With this notation, the conditional likelihood is

$$f(y_{it} \mid y_{it-1}) = \prod_{j=1}^c \prod_{k=1}^c (\pi_{it,jk})^{y_{it,jk}}. \quad (11.16)$$

Here, in the case that $\pi_{it,jk} = 0$ (by assumption), we have that $y_{it,jk} = 0$ and use the convention that $0^0 = 1$.

To simplify matters, we assume that the initial state distribution, $\text{Prob}(y_{i1})$, is described by a different set of parameters than the transition distribution, $f(y_{it} \mid y_{i,t-1})$. Thus, to estimate this latter set of parameters, one only needs to maximize the partial log-likelihood

$$L_p = \sum_{i} \sum_{t=2}^T \ln f(y_{it} \mid y_{i,t-1}) \quad (11.17)$$

where $f(y_{it} \mid y_{i,t-1})$ is specified in equation (11.16). In some cases, the interesting aspect of the problem is the transition. In this case, one loses little by focusing on the partial likelihood. In other cases, the interesting aspect is the state, such as the proportion of retirements at a certain age. Here, a representation for the initial state distribution takes on greater importance.

In equation (11.15), we specified separate components for each alternative. Assuming no implicit relationship among the components, this specification yields a particularly simple analysis. That is, we may write the partial log-likelihood as

$$L_p = \sum_{j=1}^c L_{p,j}(\beta_j)$$

where, from equations (11.14)-(11.16), we have

$$L_{p,j}(\beta_j) = \sum_{i} \sum_{t=2}^T \ln f(y_{it} \mid y_{i,t-1} = j) = \sum_{i} \sum_{t=2}^T \left\{ \sum_{k=1}^c y_{it,jk} \exp(x'_{it,jk} \beta_j) - \ln \left( \sum_{k=1}^c \exp(x'_{it,jk} \beta_j) \right) \right\},$$

as in equation (11.4). Thus, we can split up the data according to each (lagged) choice and determine maximum likelihood estimators for each alternative, in isolation of the others.

**Example 11.3 Choice of yogurt brands - Continued**

To illustrate, we return to the Yogurt data set. We now explicitly model the transitions between brand choices, as denoted in Figure 11.3. Here, purchases of yogurt occur intermittently over a two-year period; the data are not observed at discrete time intervals. By ignoring the length of time between purchases, we are using what is sometimes referred to as “operational time.” In effect, we are assuming that one’s most recent choice of a brand of yogurt has the same effect on today’s choice, regardless as to whether the prior choice was one day or one month ago. This assumption suggests future refinements to the transition approach in modeling yogurt choice.
Tables 11.6a and 11.6b show the relation between the current and most recent choice of yogurt brands. Here, we call the most recent choice the “origin state” and the most current choice the “destination state.” Table 11.6a shows that there are only 2,312 observations under consideration; this is because initial values from each of 100 subjects are not available for the transition analysis. For most observation pairs, the current choice of the brand of yogurt is the same as chosen most recently, exhibiting “brand loyalty.” Other observation pairs can be described as “switchers.” Brand loyalty and switching behavior is more apparent in Table 11.6b, where we rescale counts by row totals to give (rough) empirical transition probabilities. Here, we see that customers of Yoplait, Dannon and Weight Watchers exhibit more brand loyalty compared to those of Hiland who are more prone to switching.

Table 11.6a Yogurt Transition Counts

<table>
<thead>
<tr>
<th>Destination State</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>654</td>
<td>65</td>
<td>41</td>
<td>17</td>
<td>777</td>
</tr>
<tr>
<td>Dannon</td>
<td>71</td>
<td>822</td>
<td>19</td>
<td>16</td>
<td>928</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>44</td>
<td>18</td>
<td>473</td>
<td>5</td>
<td>540</td>
</tr>
<tr>
<td>Hiland</td>
<td>14</td>
<td>17</td>
<td>6</td>
<td>30</td>
<td>67</td>
</tr>
<tr>
<td>Total</td>
<td>783</td>
<td>922</td>
<td>539</td>
<td>68</td>
<td>2,312</td>
</tr>
</tbody>
</table>

Table 11.6b Yogurt Transition Empirical Probabilities, in Percent

<table>
<thead>
<tr>
<th>Destination State</th>
<th>Yoplait</th>
<th>Dannon</th>
<th>Weight Watchers</th>
<th>Hiland</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoplait</td>
<td>84.2</td>
<td>8.4</td>
<td>5.3</td>
<td>2.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Dannon</td>
<td>7.7</td>
<td>88.6</td>
<td>2.0</td>
<td>1.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Weight Watchers</td>
<td>8.1</td>
<td>3.3</td>
<td>87.6</td>
<td>0.9</td>
<td>100.0</td>
</tr>
<tr>
<td>Hiland</td>
<td>20.9</td>
<td>25.4</td>
<td>9.0</td>
<td>44.8</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Of course, Tables 11.6a and 11.6b do not account for changing aspects of price and features. In contrast, these explanatory variables are captured in the multinomial logit fit, displayed in Table 11.7. Table 11.7 shows that purchase probabilities for customers of Dannon, Weight Watchers and Hiland are more responsive to a newspaper ad than Yoplait customers.
Moreover, compared to the other three brands, Hiland customers are not price sensitive in that changes in PRICE have relatively impact on the purchase probability (it is not even statistically significant).

Table 11.6b suggests that prior purchase information is important when estimating purchase probabilities. To test this, it is straightforward use a likelihood ratio test of the null hypothesis $H_0: \beta_j = \beta$, that is, the components do not vary by origin state. Table 11.7 shows that the total (minus two times the partial) log-likelihood is $2,379.8 + \ldots + 281.5 = 6,850.3$. Estimation of this model under the null hypothesis yields a corresponding value of 9,741.2. Thus, the likelihood ratio test statistic is $LRT = 2,890.9$. There are 15 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis, corroborating the intuition that the most recent type of purchase has a strong influence in the current brand choice.

### Table 11.7 Yogurt Transition Model Estimates

<table>
<thead>
<tr>
<th>State of Origin</th>
<th>Variable</th>
<th>Yoplait</th>
<th>Estimate</th>
<th>t-stat</th>
<th>Dannon</th>
<th>Estimate</th>
<th>t-stat</th>
<th>Weight Watchers</th>
<th>Estimate</th>
<th>t-stat</th>
<th>Hiland</th>
<th>Estimate</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yoplait</td>
<td>5.952</td>
<td>12.75</td>
<td>4.125</td>
<td>9.43</td>
<td>4.266</td>
<td>6.83</td>
<td>0.215</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dannon</td>
<td>2.529</td>
<td>7.56</td>
<td>5.458</td>
<td>16.45</td>
<td>2.401</td>
<td>4.35</td>
<td>0.210</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weight Watchers</td>
<td>1.986</td>
<td>5.81</td>
<td>1.522</td>
<td>3.91</td>
<td>5.699</td>
<td>11.19</td>
<td>-1.105</td>
<td>-1.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FEATURES</td>
<td>0.593</td>
<td>2.07</td>
<td>0.907</td>
<td>2.89</td>
<td>0.913</td>
<td>2.39</td>
<td>1.820</td>
<td>3.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRICE</td>
<td>-41.257</td>
<td>-6.28</td>
<td>-48.989</td>
<td>-8.01</td>
<td>-37.412</td>
<td>-5.09</td>
<td>-13.840</td>
<td>-1.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 11.8 Tax Preparers Transition Empirical Probabilities, in Percent

<table>
<thead>
<tr>
<th>Origin State</th>
<th>Destination State</th>
<th>PREP = 0</th>
<th>PREP = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREP = 0</td>
<td>546</td>
<td>89.4</td>
<td>10.6</td>
</tr>
<tr>
<td>PREP = 1</td>
<td>486</td>
<td>8.4</td>
<td>91.6</td>
</tr>
</tbody>
</table>
Table 11.9 provides a more formal assessment with a fit of a logit transition model. To assess whether or not the transition aspect is an important piece of the model, we can use a likelihood ratio test of the null hypothesis $H_0: \beta_j = \beta$, that is, the coefficients do not vary by origin state. Table 11.9 shows that the total (minus two times the partial) log-likelihood is $361.5 + 264.6 = 626.1$. Estimation of this model under the null hypothesis yields a corresponding value of 1,380.3. Thus, the likelihood ratio test statistic is $LRT = 754.2$. There are 4 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis, corroborating the intuition that most recent choice is an important predictor of the current choice.

To interpret the regression coefficients in Table 11.9, we use the summary statistics in Section 9.1.3 to describe a “typical” tax filer and assume that $\text{LNTP} = 10$, $\text{MR} = 23$ and $\text{EMP} = 0$. If this tax filer had not previously chosen to use a preparer, the estimated systematic component is $V = -10.704 + 1.024(10) - 0.072(23) + 0.352(0) = -2.12$. Thus, the estimated probability of choosing to use a preparer is $\exp(-2.12)/(1+\exp(-2.12)) = 0.107$. Similar calculations show that, if this tax filer had chosen to use a preparer, then the estimated probability is 0.911. These calculations are in accord with the estimates in Table 11.8 that do not account for the explanatory variables. This illustration points out the importance of the intercept in determining these estimated probabilities.

### Table 11.9 Tax Preparers Transition Model Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>PREP = 0</th>
<th>PREP = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.704</td>
<td>0.208</td>
</tr>
<tr>
<td>LNTP</td>
<td>1.024</td>
<td>0.104</td>
</tr>
<tr>
<td>MR</td>
<td>-0.072</td>
<td>0.047</td>
</tr>
<tr>
<td>EMP</td>
<td>0.352</td>
<td>0.750</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>361.5</td>
<td>264.6</td>
</tr>
</tbody>
</table>

### Higher order Markov models

There are strong serial relationships in the Taxpayer data and these may not be completely captured by simply looking at the most recent choice. For example, it may be that a tax filer who uses a preparer for two consecutive periods has a substantially different choice probability than a comparable filer who does not use a preparer in one period but elects to use a preparer in the subsequent period.

It is customary in Markov modeling to simply expand the state space to handle higher order time relationships. To this end, we may define a new categorical response, $y_{it}^* = \{y_{it}, y_{i,t-1}\}$. With this new response, the order 1 transition probability, $f(y_{it}^* | y_{i,t-1}^*)$, is equivalent to an order 2 transition probability of the original response, $f(y_{it} | y_{i,t-1}, y_{i,t-2})$. This is because the conditioning events are the same, $y_{i,t-1}^* = \{y_{i,t-1}, y_{i,t-2}\}$, and because $y_{i,t-1}$ is completely determined by the conditioning event $y_{i,t-1}^*$. Expansions to higher orders can be readily accomplished in a similar fashion.

To simplify the exposition, we consider only binary outcomes so that $c = 2$. Examining the transition probability, we are now conditioning on four states, $y_{i,t-1}^* = \{y_{i,t-1}, y_{i,t-2}\} = \{(0,0), (0,1), (1,0), (1,1)\}$. As above, one can split up the (partial) likelihood into four components, one for each state. Alternatively, one can write the logit model as

$$\text{Prob}(y_{it} | y_{i,t-1}, y_{i,t-2}) = \logit(V_{it})$$

with
\[ V_{it} = x_{it}^\prime \beta_1 I(y_{i,t-1} = 0, y_{i,t-2} = 0) + x_{it}^\prime \beta_2 I(y_{i,t-1} = 1, y_{i,t-2} = 0) \\
+ x_{it}^\prime \beta_3 I(y_{i,t-1} = 0, y_{i,t-2} = 1) + x_{it}^\prime \beta_4 I(y_{i,t-1} = 1, y_{i,t-2} = 1), \]

where \( I(.) \) is the indicator function of a set. The advantage of running the model in this fashion, as compared to splitting it up into four distinct components, is that one can test directly the equality of parameters and consider a reduced parameter set by combining them. The advantage of the alternative approach is computational convenience; one performs a maximization procedure over a smaller data set and a smaller set of parameters, albeit several times.

**Example 11.4 Income tax payments and tax preparers - Continued**

To investigate the usefulness of a second order component, \( y_{i,t-2} \), in the transition model, we begin in Table 11.10 with empirical transition probabilities. Table 11.10 suggests that there are important differences in the transition probabilities for each lag 1 origin state (\( y_{i,t-1} = \text{Lag PREP} \)) between levels of the lag two origin state (\( y_{i,t-2} = \text{Lag 2 PREP} \)).

Table 11.11 provides a more formal analysis by incorporating potential explanatory variables. The total (minus two times the partial) log-likelihood is 469.4. Estimation of this model under the null hypothesis yields a corresponding value of 1,067.7. Thus, the likelihood ratio test statistic is \( LRT = 567.3 \). There are 12 degrees of freedom for this test statistic. Thus, this provides strong evidence for rejecting the null hypothesis. With this data set, estimation of the model incorporating lag one differences yields total (partial) (minus two times) log-likelihood of 490.2. Thus, the likelihood ratio test statistic is \( LRT = 20.8 \). With 8 degrees of freedom, comparing this test statistic to a chi-square distribution yields a \( p \)-value of 0.0077. Thus, the lag two component is statistically significant contribution to the model.

<p>| Table 11.10 Tax Preparers Order 2 Transition Empirical Probabilities, in Percent |
|-------------------------------|-----------------------------------|
| <strong>Origin State</strong>               | <strong>Destination State</strong>             |</p>
<table>
<thead>
<tr>
<th>Lag PREP</th>
<th>Lag 2 PREP</th>
<th>Count</th>
<th>PREP = 0</th>
<th>PREP = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>376</td>
<td>89.1</td>
<td>10.9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>28</td>
<td>67.9</td>
<td>32.1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>43</td>
<td>29.4</td>
<td>70.6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>327</td>
<td>6.1</td>
<td>93.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 11.11 Tax Preparers Order 2 Transition Model Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State of Origin</strong></td>
</tr>
<tr>
<td>Lag PREP = 0, Lag 2 PREP = 0</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>LNTPI</td>
</tr>
<tr>
<td>MR</td>
</tr>
<tr>
<td>EMP</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
</tr>
</tbody>
</table>


Chapter 11. Categorical Dependent Variables and Survival Models

Just as one can incorporate higher order lags into a Markov structure, it is also possible to bring in the time spent in a state. This may be of interest in a model of health states, where we might wish to accommodate the time spent in a “healthy” state or an “at-risk” state. This phenomenon is known as “lagged duration dependence.” Similarly, the transition probabilities may depend on the number of prior occurrences of an event, known as “occurrence dependence.” For example, when modeling employment, we may wish to allow transition probabilities to depend on the number of previous employment spells. For further considerations of these and other specialized transition models, see Lancaster (1990E) and Haberman and Pitacco (1999O).

11.4 Survival Models

Categorical data transition models, where one models the probability of movement from one state to another, are closely related to survival models. In survival models, the dependent variable is the time until an event of interest. The classic example is time until death (the complement of death being survival). Survival models are now widely applied in many scientific disciplines; other examples of events of interest include the onset of Alzheimer’s disease (biomedical), time until bankruptcy (economics) and time until divorce (sociology).

Like the data studied elsewhere in this text, survival data are longitudinal. The cross-sectional aspect typically consists of multiple subjects, such as persons or firms, under study. There may be only one measurement on each subject but the measurement is taken with respect to time. This combination of cross-sectional and temporal aspects gives survival data their longitudinal flavor. Because of the importance of survival models, it is not uncommon for researchers to equate the phrase “longitudinal data” with survival data.

Some events of interest, such as bankruptcy or divorce, may not happen for a specific subject. It is common that an event of interest may not have yet occurred within the study period so that the data are (right) censored. Thus, the complete observation times may not be available due to the design of the study. Moreover, firms may merge or be acquired by other firms and subjects may move from a geographical area, leaving the study. Thus, the data may be incomplete due to events that are extraneous to the research question under consideration, known as random censoring. Censoring is a regular feature of survival data; large values of a dependent variable are more difficult to observe than small values, other things being equal. In Section 7.4 we introduced mechanisms and models for handling incomplete data. For the repeated cross-sectional data described in this text, models for incompleteness have become available only relatively recently (although researchers have long been aware of these issues, focusing on attrition). In contrast, models of incompleteness have been historically important and one of the distinguishing features of survival data.

Some survival models can be written in terms of the Section 11.3 transition models. To illustrate, suppose that \( Y_i \) is the time until an event of interest and, for simplicity, assume that it is discrete positive integer. From knowledge of \( Y_i \), we may define \( y_{it} \) to be one if \( Y_i = t \) and zero otherwise. With this notation, we may write the likelihood

\[
\text{Prob}(Y_i = n) = \text{Prob}(y_{i1} = 0, \ldots, y_{in} = 1, y_{in} = 1)
\]

\[
= \text{Prob}(y_{i1} = 0) \prod_{t=2}^{n-1} \text{Prob}(y_{it} = 0 | y_{i,t-1} = 0) \text{Prob}(y_{in} = 1 | y_{i,n-1} = 0),
\]

in terms of transition probabilities \( \text{Prob}(y_{it} | y_{i,t-1}) \) and the initial state distribution \( \text{Prob}(y_{i1}) \). Note that in Section 11.3 we considered \( n \) to be the non-random number of time units under consideration whereas here it is a realized value of a random variable.
Example 11.5 – Time until bankruptcy

Shumway (2001O) examined the time to bankruptcy for 3,182 firms listed on Compustat Industrial File and the CRSP Daily Stock Return File for the New York Stock Exchange over the period 1962-1992. Several explanatory financial variables were examined, including working capital to total assets, retained to total assets, earnings before interest and taxes to total assets, market equity to total liabilities, sales to total assets, net income to total assets, total liabilities to total assets and current assets to current liabilities. The data set included 300 bankruptcies from 39,745 firm years.

See also Kim et al. (1995O) for a similar study on insurance insolvencies.

Survival models are frequently expressed in terms of continuous dependent random variables. To summarize the distribution of \( Y \), define the hazard function

\[
h(t) = \frac{-\partial}{\partial t} \ln \text{Prob}(Y > t),
\]

the “instantaneous” probability of failure, conditional on survivorship up to time \( t \). This is also known as the force of mortality in actuarial science, as well as the failure rate in engineering. A related quantity of interest is the cumulative hazard function,

\[
H(t) = \int_0^t h(s) \, ds.
\]

This quantity can also be expressed as the minus the log survival function, and conversely, \( \text{Prob}(Y > t) = \exp(-H(t)) \).

Survival models regularly allow for “non-informative” censoring. Thus, define \( \delta \) to be an indicator function for right-censoring, that is,

\[
\delta = \begin{cases} 
1 & \text{if } Y \text{ is censored} \\
0 & \text{otherwise}
\end{cases}
\]

Then, the likelihood of a realization of \((Y, \delta)\), say \((y, d)\), can be expressed in terms of the hazard function and cumulative hazard as

\[
\delta \left\{ \begin{array}{ll}
\text{Prob}(Y > y) & \text{if } Y \text{ is censored} \\
-\frac{\partial}{\partial y} \text{Prob}(Y > y) & \text{otherwise}
\end{array} \right\}
\]

\[
= (\text{Prob}(Y > y))^d \cdot \left( h(y) \text{Prob}(Y > y) \right)^{1-d}
\]

\[
= h(y)^{1-d} \exp(-H(y)).
\]

There are two common methods for introducing regression explanatory variables, one is the accelerated failure time model and the other is the Cox proportional hazard model. Under the former, one essentially assumes a linear model in the logarithmic time to failure. We refer to any standard treatment of survival models for more discussion of this mechanism. Under the latter, one assumes that the hazard function can be written as the product of some “baseline” hazard and a function of a linear combination of explanatory variables. To illustrate, we use

\[
h(t) = h_0(t) \exp (x' \beta), \quad (11.18)
\]
where \( h_0(t) \) is the baseline hazard. This is known as a proportional hazards model because if one takes the ratio of hazard functions for any two sets of covariates, say \( x_1 \) and \( x_2 \), one gets
\[
\frac{h(t \mid x_1)}{h(t \mid x_2)} = \frac{h_0(t) \exp(x_1 \beta)}{h_0(t) \exp(x_2 \beta)} = \exp((x_1 - x_2) \beta),
\]
that the ratio is independent of time \( t \).

To express the likelihood function for the Cox model, let \( H_0 \) be the cumulative hazard function associated with the baseline hazard function \( h_0 \). Let \( (Y_1, \delta_1), \ldots, (Y_n, \delta_n) \) be independent and assume that \( Y_i \) follows a Cox proportional hazard model with regressors \( x_i \). Then the likelihood is
\[
L(\beta, h_0) = \prod_{i=1}^{n} \left( h(Y_i) \exp(-H(Y_i)) \right)^{\delta_i} \exp(\left( h_0(Y_i) \exp(x_i' \beta) \right)^{1-\delta_i} \exp(-h_0(Y_i) \exp(x_i' \beta)) \right).
\]

Maximizing this in terms of \( h_0 \) yields the partial likelihood
\[
L_p(\beta) = \prod_{i=1}^{n} \left( \frac{\exp(x_i' \beta)}{\sum_{j \in R(Y_i)} \exp(x_j' \beta)} \right)^{1-\delta_i},
\]
where \( R(t) \) is the set of all \( \{Y_1, \ldots, Y_n\} \) such that \( Y_i \geq t \), that is, the set of all subjects still under study at time \( t \).

From equation (11.19), we see that inference for the regression coefficients depends only on the ranks of the dependent variables \( \{Y_1, \ldots, Y_n\} \), not their actual values. Moreover, equation (11.19) suggests (and it is true) that large sample distribution theory has properties similar to the usual desirable (fully) parametric theory. This is mildly surprising because the proportional hazards model is semi-parametric; in equation (11.18) the hazard function has a fully parametric component, \( \exp(x_i' \beta) \), but also contains a nonparametric baseline hazard, \( h_0(t) \). In general, nonparametric models are more flexible than parametric counterparts for model fitting but result in less desirable large sample properties (specifically, slower rates of convergence to an asymptotic distribution).

An important feature of the proportional hazards model is that it can readily handle time dependent covariates of the form \( x_i(t) \). In this case, one can write the partial likelihood as
\[
L_p(\beta) = \prod_{i=1}^{n} \left( \frac{\exp(x_i'(Y_i))}{\sum_{j \in R(Y_i)} \exp(x_j'(Y_j))} \right)^{1-\delta_i}.
\]

Maximization of this likelihood is somewhat complex but can be readily accomplished with modern statistical software.

To summarize, there is a large overlap between survival models and the longitudinal and panel data models considered in this text. Survival models are concerned with dependent variables that are time until an event of interest whereas the focus of longitudinal and panel data models is broader. Because they concern time, survival models using conditioning arguments extensively in model specification and estimation. Also because of the time element, survival models heavily involve censoring and truncation of variables (it is often more difficult to observe large values of a time variable, other things being equal). Like longitudinal/panel data models, survival models address repeated observations on a subject. Unlike longitudinal/panel data models, survival models also address repeated occurrences of an event (such as marriage). To track events over time, survival models may be expressed in terms of stochastic processes. This
formulation allows one to model many complex data patterns of interest. There are many excellent applied introductions to survival modeling, see for example, Klein and Moeschberger (1997B) and Singer and Willet (2003EP). For a more technical treatment, see Hougaard (2000B).

**Appendix 11A. Conditional likelihood estimation for multinomial logit models with heterogeneity terms**

To estimate the parameters $\beta$ in the presence of the heterogeneity terms $\alpha_{ij}$, we may again look to conditional likelihood estimation. The idea is to condition the likelihood using sufficient statistics, introduced in Appendix 10A.2.

From equations (11.4) and (11.11), the log-likelihood for the $i$th subject is

$$
\ln L(y_i | \alpha_i) = \sum_{t}^{c} \sum_{j=1}^{y_{ij}} \ln \pi_{ij} = \sum_{t}^{c} \left( \sum_{j=1}^{y_{ij}} \ln \frac{\pi_{ij}}{\pi_{itc}} + \ln \pi_{itc} \right) 
$$

$$
= \sum_{t}^{c} \left( \sum_{j=1}^{y_{ij}} \left( \alpha_{ij} + (x_{itj} - x_{itc})'\beta \right) + \ln \pi_{itc} \right),
$$

because $\ln \frac{\pi_{ij}}{\pi_{itc}} = V_{ij} - V_{itc} = \alpha_{ij} + (x_{itj} - x_{itc})'\beta$. Thus, using the factorization theorem in Appendix 10A.2, $\Sigma_t y_{ij}$ is sufficient for $\alpha_{ij}$. We interpret $\Sigma_t y_{ij}$ to be the number of choices of alternative $j$ in $T_i$ time periods.

To calculate the conditional likelihood, we let $S_{ij}$ be the random variable representing $\Sigma_t y_{ij}$ and let $sum_{ij}$ be the realization of $\Sigma_t y_{ij}$. With this, the distribution of the sufficient statistic is

$$
\text{Prob}(S_{ij} = sum_{ij}) = \sum_{B_i} \prod_{j=1}^{c} \left( \pi_{ij} \right)^{y_{ij}},
$$

where $B_i$ is the sum over all sets of the form $\{y_{ij} : \Sigma_t y_{ij} = sum_{ij}\}$. By sufficiency, we may take to $\alpha_{ij} = 0$ without loss of generality. Thus, the conditional likelihood of the $i$th subject is

$$
\frac{L(y_i | \alpha_i)}{\text{Prob}(S_{ij} = sum_{ij})} = \frac{\exp \left( \sum_{t}^{c} \sum_{j=1}^{y_{ij}} (x_{itj} - x_{itc})'\beta + \ln \pi_{itc} \right)}{\sum_{B_i} \prod_{j=1}^{c} (\pi_{ij})^{y_{ij}}}
$$

$$
= \frac{\exp \left( \sum_{t}^{c} \sum_{j=1}^{y_{ij}} (x_{itj} - x_{itc})'\beta + \ln \pi_{itc} \right)}{\sum_{B_i} \exp \left( \sum_{t}^{c} \sum_{j=1}^{y_{ij}} (x_{itj} - x_{itc})'\beta + \ln \pi_{itc} \right)}.
$$

As in Appendix 9A.2, this can be maximized in $\beta$. However, it is computationally intensive.