9

Models for Nominal Data

9.1 INTRODUCTION

In Chapter 8 we extended the latent variable model for binary outcomes to incorporate ordered response variables. We managed this by adding a set of estimated "thresholds" that served to map the assumed probability density into discrete but ordered categories. Our ability to do this relies on the "parallel regressions" assumption: the relationship between covariates and the outcome variable is constant across different categories of the outcome.

Unordered, polychotomous dependent variables are simply variables in which the categories cannot be ordered in any mathematically meaningful way. These are also called *nominal* variables having more than the two categories found in a dichotomous variable. There are lots of good examples in the social sciences: vote choice (Christian Democrat, Social Democrat, Greens, etc.); occupation (doctor, lawyer, mechanic, astronaut, student, etc.); marital status (single, married, divorced, etc.); college major (art history, modern history, Greek history, etc.); language (French, German, Urdu, etc.); ethnicity (Serb, Croat, Bosniak, Avar, Lek, etc.); and many, many others. Sometimes these nominal groups can represent ascriptive categories. But often these groups are the objects of some choice process. This, in turn, has consequences for how we think about a statistical model and find relevant covariates. For example, there may be covariates relevant to the *choice*, such as price, color, or party platform. Or we might be interested in covariates relevant to the *chooser* or choice process, like income, age, or gender.

Constructing likelihood-based models for nominal variables builds on a generalization of the binomial distribution, the appropriately named *multinomial*. From there we develop a set of linear predictors that allow us to incorporate both covariates and link functions that map the linear predictor into the appropriate interval.

In case you were wondering ... 9.1 Categorical and multinomial distributions

Let S be some finite set with cardinality M. Define Z_i such that, for any $m \in S$, $\Pr(Z_i = m) = p_m$, $\sum_{m \in S} p_m = 1$, and $\Pr(Z_i = l) = 0 \quad \forall \quad l \notin S$. We say that Z_i follows a *categorical distribution*, with probability mass function denoted

$$Z_i \sim f_c(z_i; p_1, \ldots, p_M).$$

We can write the expectation of a categorical variable using the notion $\mathbb{1}_m$ to be the indicator variable taking on 1 if $Z_i = m$ and 0 otherwise. $\mathrm{E}[\mathbb{1}_m] = p_m$ and $\mathrm{var}(\mathbb{1}_m) = p_m(1 - p_m)$. The covariance between any two categories, $a, b \in S$, is given as $\mathrm{cov}(\mathbb{1}_a, \mathbb{1}_b) = -p_a p_b$. The Bernoulli distribution is a special case of the categorical distribution in which $M = 2$.

Now suppose we have n independent realizations of Z_i. Let $n_j \leq n, j \in S$ denote the number of realizations in category j such that $\sum_{j \in S} n_j = n$. Let $Y = (n_1, \ldots, n_M)$. We say that Y follows a multinomial distribution with parameter vector $\boldsymbol{\theta} = (n, p_1, \ldots, p_M)$ and probability mass function

$$Y \sim f_m(\mathbf{y}; \boldsymbol{\theta}) = \frac{n!}{\prod_{j \in S} n_j!} \prod_{j \in S} p_j^{n_j},$$

with $\mathrm{E}[Y] = (np_1, \ldots, np_M)$ and $\mathrm{var}(Y) = (np_1(1-p_1), \ldots, np_M(1-p_M))$. The covariance between any two categories, $a, b \in S$, is given as $\mathrm{cov}(\mathbb{1}_a, \mathbb{1}_b) = -np_a p_b$. The binomial distribution is a special case of the multinomial distribution in which $M = 2$.

9.2 MULTINOMIAL LOGIT

The multinomial logit model describes nominal outcomes such that the influence of each independent variable differs by outcome category. As a running example, consider data from the 2013 Australian National Election Survey describing how survey respondents voted for various parties in the lower house of parliament. Respondents could choose between the two major parties, Labor and Liberal, as well as the National Party, the Green Party, and "other," a residual category that included spoiling a ballot (Australia has compulsory voting). For our purposes we will simplify this to three categories, Labor, Coalition (Liberal and National), and Other. We are interested in how respondents' measured attributes relate to their vote choice. To derive a likelihood we build on the categorical distribution.

Suppose we have n survey respondents indexed by i. The outcome variable, Y_i, can take one and only one of M values, which we consider to be unordered categories. For example, we might have $Y_i \in \{\text{L,C,O}\}$. We *index* the elements of this set with j. For example $y_i = \text{C}$ would imply $j = 2$. Applying the categorical distribution, $\Pr(Y_i = m) = p_{im}$ and $\sum_{h=1}^{M} p_{ih} = 1$.

We want to allow p_{im} to vary as a function of $k - 1$ observation-level covariates along with a constant. That is, the response variable is an $n \times 1$ vector where element i records the category in which i falls. In our example, each cell is respondent i's vote intention. The covariate data is the typical $n \times k$ matrix in which each row represents an observation's values on $k - 1$ independent variables (plus a 1 for the intercept). In our example, covariates include income (in \$AU 10,000), union membership, religious affiliation (Protestant as reference), sex, and age.

Because covariates are observed at the level of the unit, i, and not the category, m, regression coefficients differ across outcome categories. This must be the case because a covariate that increases the chances that i chooses Labor must necessarily reduce the probability that she chooses at least one of the other options. In other words the multinomial model posits a different $\boldsymbol{\beta}_m$ specific to each outcome category.

We want to connect our linear predictor term, $\mathbf{x}_i^\top \boldsymbol{\beta}_m$, to the probabilities, p_{im}. Probabilities must be nonnegative, so we can use an exponential setup: $p_{im} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_m)$. To ensure that probabilities sum to 1 across outcome categories, we divide by the sum across all M categories:

$$\Pr(Y_i = m) \equiv p_{im} = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_m)}{\sum_{h=1}^{M} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_h)}. \tag{9.1}$$

The expression in Equation 9.1 is unidentified. To fix the model, we must choose a *baseline* or *reference* category against which other categories are compared. Computationally this is accomplished by constraining the $\boldsymbol{\beta}_m$ for a particular category, usually the first, to be zero. This results in the following restatement of the basic multinomial model:

$$\Pr(Y_i = m \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_m)}{1 + \sum_{j=2}^{M} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}. \tag{9.2}$$

Inspecting Equation 9.2 reveals an important insight: the multinomial logit model is essentially $M - 1$ different binary logits estimated simultaneously. It differs from actually estimating a series of binary logits in that we constrain the probabilities to sum to unity, gaining efficiency through joint estimation. But, unlike the ordered model considered in the previous chapter, the multinomial model does not constrain the $\boldsymbol{\beta}_m$ to be constant across categories; we estimate $M - 1$ different sets of regression parameters, each of which describes the log odds of being in category m versus the reference category. It then follows that any multinomial model can be reestimated with a different reference category. This will not change the overall fit of the model or its implications, but it will generally change the regression coefficients that you see on your computer screen and their immediate interpretations. But, as we will see later in this section, we can still recover pairwise comparisons across categories not used as a baseline.

Thinking of the multinomial model as a set of binary models also highlights another important fact: multinomial models are very demanding of the data. Multinomial models burn degrees of freedom rapidly as we estimate $M-1$ parameters for every new explanatory variable. In a multinomial framework all the problems of perfect separation and rare events that we encountered in binary data are amplified. With so many more predictor–outcome combinations, small numbers of observations in any of these cells can arise more easily.

The multinomial likelihood can be formed following our usual steps and using $\mathbb{1}_{ij}$ as an indicator variable taking on 1 when observation i is in the jth category and 0 otherwise. For notational simplicity, the equations include sums over all M categories. To identify the model, let $\boldsymbol{\beta}_1 = 0$ so $\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_1) = 1$.

$$\mathcal{L}_i = \prod_{h=1}^{M} \left(\frac{\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_h)}{\sum_{\ell=1}^{M} \exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_\ell)} \right)^{\mathbb{1}_{ih}},$$

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\prod_{h=1}^{M} \left(\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_h)\right)^{\mathbb{1}_{ih}}}{\sum_{\ell=1}^{M} \exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_\ell)},$$

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{h=1}^{M} \mathbb{1}_{ij}\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_h - \log \left(\sum_{\ell=1}^{M} \exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_\ell) \right).$$

This likelihood is nice in all the standard ways: globally concave and quickly converging, producing (in the limit) estimates that are consistent, normal, and efficient.

9.2.1 A Latent Variable Formulation

Like the binary logit model, the multinomial model can be derived in a latent variable framework. Suppose an individual, i, chooses among discrete alternatives in the set S, with a utility $U_i(m)$, associated with each choice, m. Like all of statistics, each utility has a stochastic part and a systematic part. That is, $U_i(m) = \mu_i(m) + \epsilon_{im}$. The systematic part is a function of variables associated with the individual and might consist of different weights for each

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

of these characteristics across alternatives: $\mu_i(m) = \mathbf{x}_i^\top \boldsymbol{\beta}_m$. Being clever, the individual chooses among the alternatives so as to maximize utility:

$$\begin{aligned}
\Pr(Y_i = m) &= \Pr(U_i(m) > U_i(d) \quad \forall \quad d \neq m \in S) \\
&= \Pr(\mu_i(m) + \epsilon_{im} > \mu_i(d) + \epsilon_{id} \quad \forall \quad d \neq m \in S) \\
&= \Pr(\mathbf{x}_i^\top \boldsymbol{\beta}_m + \epsilon_{im} > \mathbf{x}_i^\top \boldsymbol{\beta}_d + \epsilon_{id} \quad \forall \quad d \neq m \in S) \\
&= \Pr(\epsilon_{im} - \epsilon_{id} > \mathbf{x}_i^\top (\boldsymbol{\beta}_d - \boldsymbol{\beta}_m) \quad \forall \quad d \neq m \in S). \qquad (9.3)
\end{aligned}$$

If the difference between the stochastic component and that of any alternative is greater than the difference in the systematic parts, it has the highest utility, because either ϵ_{im} is large or $\mathbf{x}_i^\top \boldsymbol{\beta}_m$ is large, or both. The expression in Equation 9.3 relies on differences between coefficient vectors across categories, so once again we see that model identification requires that we fix one category as the baseline. To complete the model we need to choose an expression for the stochastic component, i.e., we specify the distribution of the error terms, ϵ_{im}. If we choose a multivariate normal distribution we arrive at the multinomial probit model. If we use a standard *type-I extreme value* distribution (EV-I), and the errors are i.i.d. across categories, we arrive at the multinomial logit model already introduced.

In case you were wondering ... 9.2 Extreme value distributions

We say that a random variable, V, follows a *generalized extreme value distribution* with parameter vector $\boldsymbol{\theta} = (\mu, \sigma, \eta)$:

$$V \sim f_{GEV}(v; \mu, \sigma, \eta)$$
$$\Pr(V \leq v) = \begin{cases} \exp\left(-\left(1 + \eta\frac{v-\mu}{\sigma}\right)^{1/\eta}\right) & \eta \neq 0, \\ \exp\left(-\exp\left(-\frac{v-\mu}{\sigma}\right)\right) & \eta = 0 \end{cases}.$$

The case where $\eta = 0$ is called the *type-I extreme value distribution*, also known as the *Gumbel*, *log-Weibull*, and the *double exponential* distribution. $\mathrm{E}[V] = \mu + \sigma\gamma$ and $\mathrm{var}(V) = \frac{\pi^2}{6}\sigma^2$, where γ is Euler's constant. Setting $\mu = 0$ and $\sigma = 1$, we have the standard type-I extreme value distribution, and we write $V \sim f_{EV_1}(v; 0, 1)$.

The choice of EV-I appears arbitrary. The following theorem justifies this choice by linking the EV-I distribution to the logistic distribution we are already familiar with from Chapters 3 and 8.

Theorem 9.1. *If* $A, B \overset{i.i.d.}{\sim} f_{EV_1}(0, 1)$ *then* $A - B \sim f_L(0, 1)$.

Proof The proof involves the convolution of two type-I extreme value distributions. Let $C = A - B$.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

$$
\begin{aligned}
F_C(c) &= \Pr(C \leq c) = \Pr(A - B \leq c) \\
&= \int_{b=-\infty}^{\infty} \Pr(A \leq c + b) f_{EV_1}(b; 0, 1) db \\
&= \int_{b=-\infty}^{\infty} F_{EV_1}(c + b; 0, 1) f_{EV_1}(b; 0, 1) db \\
&= \int_{b=-\infty}^{\infty} \exp(-\exp(-(c + b))) \exp(-b - \exp(-b)) db \\
&= \int_{b=-\infty}^{\infty} \exp(-b - \exp(-b)(1 + \exp(-c))) db.
\end{aligned}
$$

Let $u = \exp(b)$. This leads to:

$$
F_C(c) = \int_{u=0}^{\infty} \frac{1}{u^2} \exp\left(\frac{1}{u}(-1 - \exp(-c))\right) du.
$$

Let $v = \frac{1}{u}(-1 - \exp(-c))$. We now have

$$
F_C(c) = \frac{1}{1 + \exp(-c)} \int_{v=-\infty}^{0} \exp(v) dv = \frac{1}{1 + \exp(-c)}.
$$

The last expression is the CDF for the standard logistic distribution. □

From Theorem 9.1 we see that each pairwise difference between alternatives follows a logistic distribution, just as in the binary logit model. Once again we see that a multinomial logit model can be viewed as a collection of $M-1$ binary logits.

9.2.2 IIA

IIA stands for the *independence of irrelevant alternatives*. IIA is an assumption about the nature of the choice process: under IIA, an individual's choice does not depend on the availability or characteristics of inaccessible alternatives. IIA is closely related to the notion of transitive (or acyclic) preferences.

Returning to the Australian election example, suppose a voter is asked whether she prefers the Labor Party or the Liberal Party and she responds with "Liberal." The interviewer then reminds her that the Green Party is also fielding candidates, and she switches her choice. IIA says that the only admissible switch she could make is to the Green Party. She cannot say Labor because she could have chosen Labor before (when it was only Labor v. Liberal) but decided not to. More formally, IIA says that if you hold preferences {Liberal $\succ$ Labor} when those are the only two options, then you must also hold preferences {Green $\succ$ Liberal $\succ$ Labor} or {Liberal $\succ$ Labor$\succ$ Green} or {Liberal $\succ$ Green $\succ$ Labor} when the Green party is available. Orderings like {Labor $\succ$ Liberal $\succ$ Green}, in which Labor and Liberal switch positions once Green becomes available, are not admissible under the IIA assumption.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

The multinomial model implies that

$$\begin{aligned}
\frac{\Pr(Y_i = m)}{\Pr(Y_i = d)} &= \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_m)}{\sum_{\ell=1}^{M} \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_\ell)} \frac{\sum_{\ell=1}^{M} \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_\ell)}{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_d)}, \\
&= \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_m)}{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_d)}, \\
&= \exp[\mathbf{x}_i^\intercal (\boldsymbol{\beta}_m - \boldsymbol{\beta}_d)]. \qquad (9.4)
\end{aligned}$$

That is, the log ratio of the probabilities for any two alternatives m and d is just the values of the covariates times the difference between the two alternatives' coefficient vectors. Importantly, this means that *the ratio of the probabilities of choosing any two outcomes is invariant with respect to the other alternatives*. It only depends on the characteristics of the alternatives in question:

$$\frac{\Pr(Y_i = m|M_R)}{\Pr(Y_i = d|M_R)} = \frac{\Pr(Y_i = m|M_S)}{\Pr(Y_i = d|M_S)} \forall m, d \in R \subseteq S.$$

In other words, the IIA assumption is baked in to the standard multinomial model. The IIA assumption buys us enormous computational simplicity: rather than having to evaluate an M-dimensional integral, we can construct a series of binary logits. This assumption may or may not hold in applied settings. Violations of IIA imply that our model of the data-generating process is inaccurate.

Diagnosing IIA Violations

There are tests for violations of the IIA assumption. If the IIA assumption holds, then a model omitting any particular choice should return $\hat{\boldsymbol{\beta}}_m$'s for the remaining alternatives similar to those estimated under the full model. Conversely, if the $\hat{\boldsymbol{\beta}}_m$'s vary a lot when an alternative is omitted, the data likely violate IIA. A common statistical test built on this insight is known as the Hausman-McFadden test (Hausman and McFadden, 1984). The test statistic takes the form:

$$\begin{aligned}
H &= (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_u)^\intercal [\hat{\mathbf{V}}_r - \hat{\mathbf{V}}_u]^{-1} (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_u) \qquad (9.5) \\
H \,|_{H_0} &\sim \chi^2,
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_r$ are the estimates from the restricted model (i.e., the model with an omitted alternative), $\hat{\boldsymbol{\beta}}_u$ is the vector of estimates for the unrestricted model (i.e., the one with all the alternatives included), and $\hat{\mathbf{V}}_r$ and $\hat{\mathbf{V}}_u$ are the estimated variance-covariance matrices for the two sets of coefficients, respectively.[1] Under the null hypothesis that IIA holds, the test is distributed

[1] Let $\boldsymbol{\beta}_m$ be the coefficients for the choice category that is omitted from the restricted model but included in the unrestricted one. For these vectors and matrices to be conformable, we omit all the $\boldsymbol{\beta}_m$ elements from $\hat{\boldsymbol{\beta}}_u$ and $\hat{\mathbf{V}}_u$.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

χ^2 with degrees of freedom equal to the rank of $\mathbf{V}_r$, typically the number of estimated parameters in the restricted model. The logic of the test is as follows: if IIA holds, then both the restricted and unrestricted models will be consistent, but the unrestricted model will be more efficient (smaller variance). If IIA does not hold, however, then the unrestricted model is consistent (assuming nothing else has been left out), but the restricted model is not.

While IIA tests can be useful they should be viewed as heuristics rather than dispositive. The Hausman-McFadden test can actually return negative values. In the original paper Hausman and McFadden interpret this as evidence that IIA is satisfied, but such values clearly do not satisfy the asymptotic properties of the test statistic, thus p-values are meaningless in such cases. If there are many categories, and each one is tested sequentially for IIA violations, then we have the traditional multiple testing problem. Cheng and Long (2007) argue that IIA tests are underpowered and can give conflicting results in practice. In Section 9.5 we discuss modeling extensions that relax the IIA assumptions. Deciding whether these models are appropriate in any particular application usually requires more than a simple Hausman test. Comparisons across competing models can help.

9.3 AN EXAMPLE: AUSTRALIAN VOTERS IN 2013

We are now in a position to examine the Australian survey data from the 2013 election. We will fit the model twice, once with "Labor" as the reference category and once with "Coalition." Results in the form of a BUTON appear in Table 9.1.

Table 9.1 shows that multinomial models produce a lot of output since we are estimating $k \times (M - 1)$ parameters, one for each covariate category except the baseline. The size of the BUTON can therefore grow very quickly. Altering the reference category alters the coefficients you see but not their implications in terms of $\hat{p}_{im}$. To see this, examine the Coalition and Labor columns. Entries in the Coalition column describe how a covariate relates to the probability of choosing Coalition relative to Labor, while the coefficients in the Labor column describe the reverse relationship. As we would expect, the coefficients are exactly equal in absolute value but opposite in sign.

9.3.1 Evaluation

Before interpreting the output of a multinomial model, we have several tools available to describe model fit. With multinomial models we, of course, have access to the usual in-sample test statistics as well as the AIC and BIC. Most statistical software packages report the likelihood ratio model diagnostic comparing a null model ($M - 1$ intercepts) to the one specified. This tells us whether all coefficients estimated are equal to 0 for all M outcomes and k variables. This is typically not very informative, like most null models.

TABLE 9.1 *Multinomial logistic regression on vote choice in the 2013 Australian elections. The left half of the table uses Labor as the reference category, whereas the right uses Coalition.*

	Coalition	**Other**	**Labor**	**Other**
(Intercept)	−0.32	−0.33	0.32	−0.01
	(0.21)	(0.25)	(0.21)	(0.24)
income	0.05	0.01	−0.05	−0.03
	(0.01)	(0.01)	(0.01)	(0.01)
union member	−1.03	−0.21	1.03	0.82
	(0.10)	(0.11)	(0.10)	(0.12)
Catholic	−0.24	−0.35	0.24	−0.11
	(0.11)	(0.15)	(0.11)	(0.14)
not religious	−0.71	0.37	0.71	1.08
	(0.11)	(0.13)	(0.11)	(0.12)
other religion	−0.21	0.23	0.21	0.44
	(0.13)	(0.16)	(0.13)	(0.15)
female	−0.05	0.05	0.05	0.10
	(0.08)	(0.10)	(0.08)	(0.10)
age	0.02	−0.01	−0.02	−0.02
	(0.00)	(0.00)	(0.00)	(0.00)
n	3,342			
$\log \mathcal{L}$	−3,316			
AIC	6,664			
BIC	6,762			

In general, model evaluation is best conducted by comparing models to one another. Within the multinomial framework we can check individual sets of estimated coefficients by comparing an unrestricted model to a restricted version that excludes some specific variable $j \in \{1, 2, \ldots, k\}$. In the multinomial context, this restriction takes the form $\beta_{jm} = 0 \quad \forall \quad m$, a test of whether variable j is jointly significant across the $M - 1$ outcome categories. The corresponding likelihood ratio is distributed χ^2 with $M - 1$ degrees of freedom for each of the variables excluded.

The multinomial model allows us to test whether we have "too many" categories given our ability to distinguish between groups. In our Australian election example, if our covariates are unable to distinguish between Coalition and Other, then (for fixed covariates) we can combine the two categories, gaining efficiency and simplifying our model considerably. For concreteness, let's consider the model with Labor as the baseline. There are k covariates (including the intercept). An inability to distinguish Coalition and Other can be expressed as the null hypothesis

$$\hat{\beta}_{jC} = \hat{\beta}_{jO} \quad \forall \quad j \in \{2, 3, \ldots, k\}.$$

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 9.1 ***Wald test for combining categories in a multinomial logit***

```
# the myoz data from the online repository is needed here.
library(nnet)  #where the multinom() function lives
mnl.fit<-multinom(vc.simp ~   income2 + union + religion.simp +
  sex + age, Hess=T, model=T, data=myoz, maxit=200)  #model
Beta<-as.vector(t(coef(mnl.fit)))        #vectorizing coefficients
A<-rbind(                                #constraint matrix
  c(0,1,0,0,0,0,0,0,0,-1,0,0,0,0,0,0), #income_c - income_o
  c(0,0,1,0,0,0,0,0,0,0,-1,0,0,0,0,0), #union_c - union_o
  c(0,0,0,1,0,0,0,0,0,0,0,-1,0,0,0,0), #Catholic_c - Catholic_o
  c(0,0,0,0,1,0,0,0,0,0,0,0,-1,0,0,0), #no relig_c - no relig_o
  c(0,0,0,0,0,1,0,0,0,0,0,0,0,-1,0,0), #other relig_c - other relig_o
  c(0,0,0,0,0,0,1,0,0,0,0,0,0,0,-1,0), #female_c-female_o
  c(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,-1)  #age_c-age_o
  )
wt<-t(A%*%Beta)%*%solve(A%*%vcov(mnl.fit)%*%t(A))%*%(A%*%Beta) #Wald test
pchisq(wt, df=dim(A%*%Beta)[1], lower.tail=FALSE)
```

Stated in words, all coefficients (except the intercepts) for outcomes C and O are equal. This can be restated as a simple linear constraint: the differences in coefficients (except the intercepts) for each of the two outcomes are zero under the null, which leads naturally to a Wald test. Code Example 9.1 calculates the model in Table 9.1 and then performs the Wald test that the Coalition and Other can be combined. This produces a p-value ≈ 0, leading to the conclusion that we *can* distinguish Coalition from Other.[2]

Do Australian voters' choices between Labor and Coalition depend on the presence of Other alternatives? We can use Equation 9.6 to conduct a the Hausman-McFadden test for IIA. Leaving out the Other category gives us a test statistic of -2, implying that preferences between Labor and Coalition are not affected by the presence of Other, as IIA requires.

Prediction Heuristics

Alongside these hypothesis tests we can take advantage of a variety of tools for describing how well a multinomial model predicts outcomes, whether in- or out-of-sample. Many of the tools below are generalizations of those described for binary outcomes. All start by generating predicted probabilities for each observation across all M categories: $\hat{\mathbf{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iM})$. From these predictions it is common to define $\hat{y}_i = \arg_m \max\{\hat{p}_{im}\}$, i.e., observation i is classified into the category with the highest predicted probability.

One of the simplest diagnostic tools, a *confusion matrix*, derives from a comparison of $\hat{y}_i$ and y_i. Table 9.2 displays an in-sample version of this matrix for the model in Table 9.1. From the table we can see that the model

[2] This test can also be constructed as a likelihood ratio in which we constrain all the coefficients except the intercept for one of the *categories* to be 0.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

TABLE 9.2 *Confusion matrix for the classifications from the multinomial logistic regression in Table 9.1.*

		Predicted		
		Labor	**Coalition**	**Other**
Actual	Labor	396	665	20
	Coalition	246	1,305	21
	Other	243	419	27

TABLE 9.3 *The category-by-category "one-versus-all" confusion matrices.*

		Predicted	
		Labor	non-Labor
Actual	Labor	396	685
	non-Labor	489	1,772
		Coalition	non-Coalition
	Coalition	1,305	267
	non-Coalition	1,084	686
		Other	non-Other
	Other	27	662
	non-Other	41	2,612

is underpredicting Other and overpredicting Coalition and Labor. This is unsurprising since the distribution of observations is unbalanced across the three categories, with Other less frequently observed.

From the confusion matrix we can calculate a variety of interesting and useful quantities. For example, the overall accuracy of the model – also called the correct classification rate – is defined as the sum of the main diagonal of the confusion matrix divided by the total number of observations. In this example the model yields accuracy of 0.52. This implies error rate of 0.48. We can also construct M binary, "one-versus-all" confusion matrices in which we examine correct and incorrect prediction for each category separately. These are displayed in Table 9.3.

From the one-versus-all tables we can construct category-specific error rates and accuracy measures weighted by the prevalence of that category in the data. The *per-class error rate* and its mean across all three categories is displayed in Table 9.4. Note that the mean per-class error rate is substantially smaller than the overall error rate of 0.48 because averaging over the one-versus-all matrices has the effect of upweighting categories that are more prevalent in the sample.

Code Example 9.2 produces all the above calculations and tables.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

TABLE 9.4 *Per-class error.*

Labor	Coalition	Other	Mean Error
0.35	0.40	0.21	0.32

$\mathcal{R}$ **Code Example 9.2** ***Predictive diagnostics for multinomial logit***

```
pmnl<-predict(mnl.fit)
conmat<-table(mnl.fit$model[,1],pmnl,
  dnn=list("actual","predicted"))   #confusion matrix
sum(diag(conmat))/sum(conmat)       #overall accuracy
oneVall <- lapply(1:ncol(conmat),   #one v. all matrices
  function(i){
    v <- c(conmat[i,i],                  #true positives
      rowSums(conmat)[i] - conmat[i,i], #false negatives
      colSums(conmat)[i] - conmat[i,i], #false positives
      sum(conmat)-rowSums(conmat)[i]- colSums(conmat)[i] + conmat[i,i]);
    return(matrix(v, nrow = 2, byrow = T,
      dimnames=list(
        c(paste("actual",colnames(conmat)[i]),
          paste("actual non",colnames(conmat)[i])),
        c(paste("predicted",colnames(conmat)[i]),
          paste("predicted non",colnames(conmat)[i]))
        )
      )
    )
  }
)
pcerr<-lapply(oneVall, function(x) #per class error
  return(1-sum(diag(x))/sum(x))
)

smat <- matrix(0, nrow = 2, ncol = 2)
for(i in 1 : ncol(conmat)){smat<- smat + oneVall[[i]]}
1-sum(diag(smat))/sum(smat) #mean per class error
```

As with binary classifiers, we can generate ROC curves or separation plots, but in higher dimensions they become harder to interpret. One strategy is to simply generate separate ROC curves for each category. Figure 9.1 displays exactly such "one-versus-all" ROC plots. Multiple plots present a challenge similar to the per-class error: it is not obvious how to weigh each binary comparison when evaluating multiple models that may perform differently in their abilities to distinguish between particular categories. Higher dimensional ROC manifolds can be constructed and some have advocated for calculating single-number generalizations to the AUC (Li and Fine, 2008). Our take is that single-number summaries often mask important nuance, even in simple binary cases. That problem is accentuated in the more-complicated multinomial setting. Careful, problem-driven consideration of the costs of misclassification for particular categories, combined with a suite of diagnostic quantities, will enable a nuanced evaluation of competing models for nominal data.

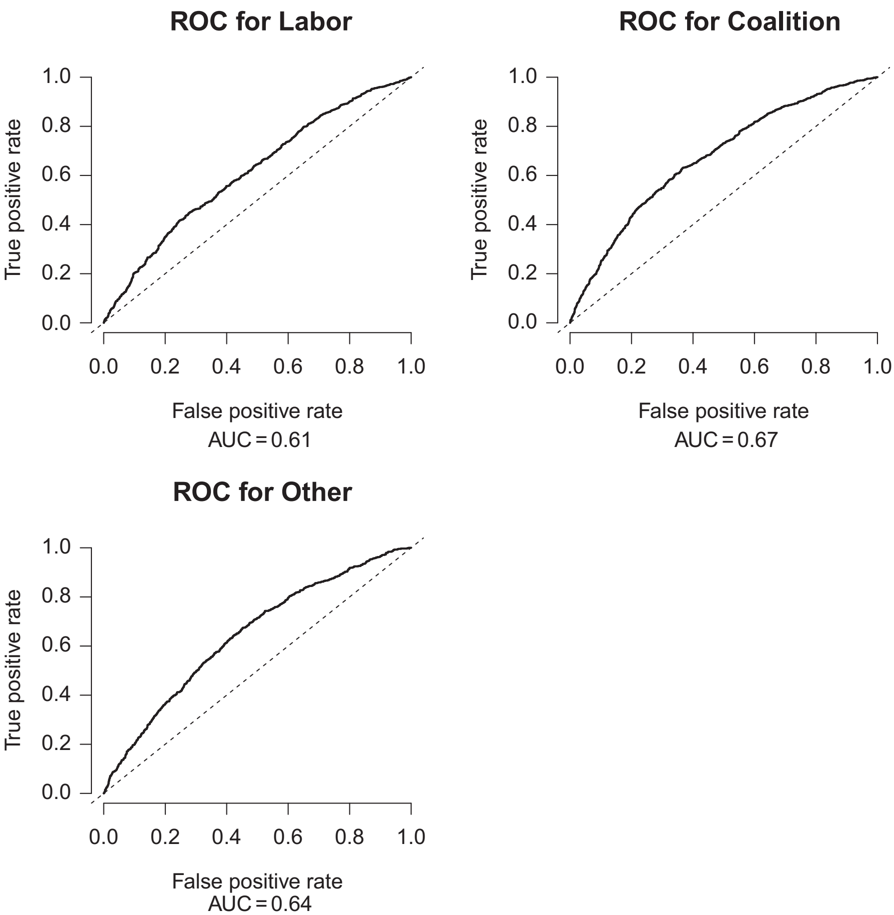


FIGURE 9.1 "One-v.-all" ROC curve diagnostics for the multinomial logit in Table 9.1.

9.3.2 Interpretation

How does one interpret the (really big) table of numbers generated from multinomial models? We proceed in the same way as in earlier chapters: simulate outcomes from the assumed data-generating process described by the model under meaningful scenarios. As usual, we include the systematic and stochastic components in all their glory. Here this means using the fundamental probability statement of the multinomial model from Equation 9.2. From there we can construct graphical displays, tables of first differences, or calculated marginal effects. With multiple outcome categories there are multiple comparisons that should be reflected in any interpretation.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

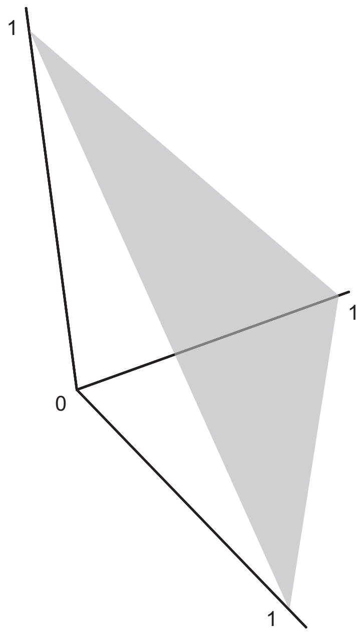


FIGURE 9.2 The three-dimensional unit simplex.

In the special – although common – case in which $M = 3$, a *ternary plot* is a useful way to display predicted probabilities. Ternary plots take advantage of the *probability simplex* to view a three dimensional object (a vector of predicted probabilities) in a convenient two-dimensional plane. This is possible because probabilities of choosing among the three options, e.g., Labor, Coalition, and Other, must sum to unity. Given predicted probabilities for any two categories, we can infer the third. The shaded triangular region in Figure 9.2 depicts exactly this three-dimensional probability simplex.

Ternary plots take the three-dimensional vector and plot it in the unit simplex triangle. Thus points on the vertices represent certainty that a respondent will choose that category. Points in the interior display the combination of probabilities simultaneously. In Figure 9.3 we construct a ternary plot in order to interpret how union membership relates to vote choice in the 2013 Australian election. In this plot we first construct a relevant scenario: a male, Protestant with income and age equal the median among union members. We then sample coefficient vectors from the multivariate normal distribution and generate predicted probabilities for each sample, one for a union member and one for a nonmember. We then plot these predicted probabilities along with the 95% confidence region. From here we can see that union members are substantially more likely to vote Labor than nonmembers. The difference between the groups is large relative to the estimation uncertainty.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

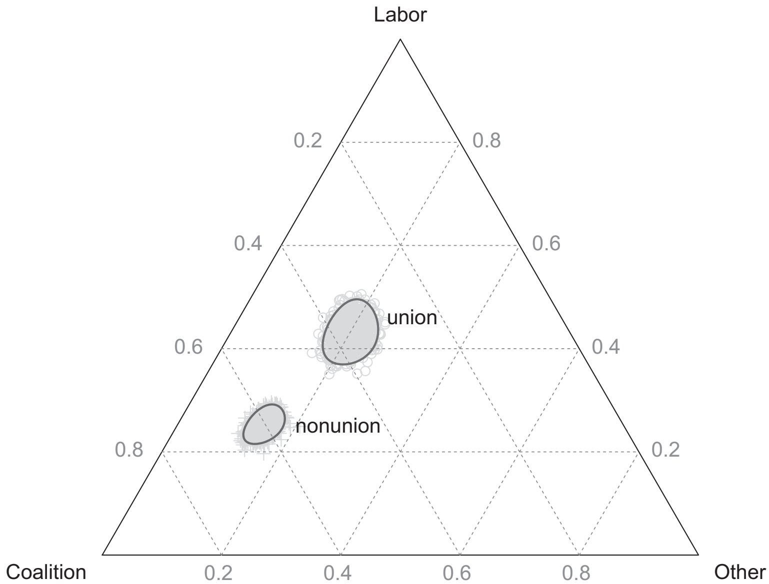


FIGURE 9.3 Using ternary plots to interpret multinomial logit models when $M = 3$. The curves define 95% confidence regions for predicted vote choice among Australians in 2013 as a function of union membership, other covariates fixed.

A second possibility for displaying results – one that generalizes beyond $M = 3$ – is shown in Figure 9.4. Here we plot cumulative probabilities, where the vertical height of each band reflects the predicted probability of falling into the category at particular levels of the covariate. In this case we are interested in displaying how vote choice in Australia is expected to vary across voters aged 20–70. The figure shows how support for the Coalition increases with age. The increase in support for the Coalition among older voters comes at the expense of Other and, especially, Labor. A strength of this plot is its clear display of trade-offs across categories at different levels of the covariates. A weakness is a failure to show estimation uncertainty.

We can also unstack the predicted probabilities to show estimation uncertainty, as in Figure 9.5. In this figure we are interested in looking at how vote choice relates to income. We construct a scenario in which we examine a nonunion, Protestant female of median age (among women). We vary income between its 20th and 80th percentiles. In this scenario we see that higher income respondents are much more likely to support the Coalition and less likely to support Labor. Chosing "Other" is unrelated to income.

Other Interpretation Strategies

While graphical displays accounting for uncertainty are useful, you will likely encounter other interpretation approaches in your reading. One such strategy

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

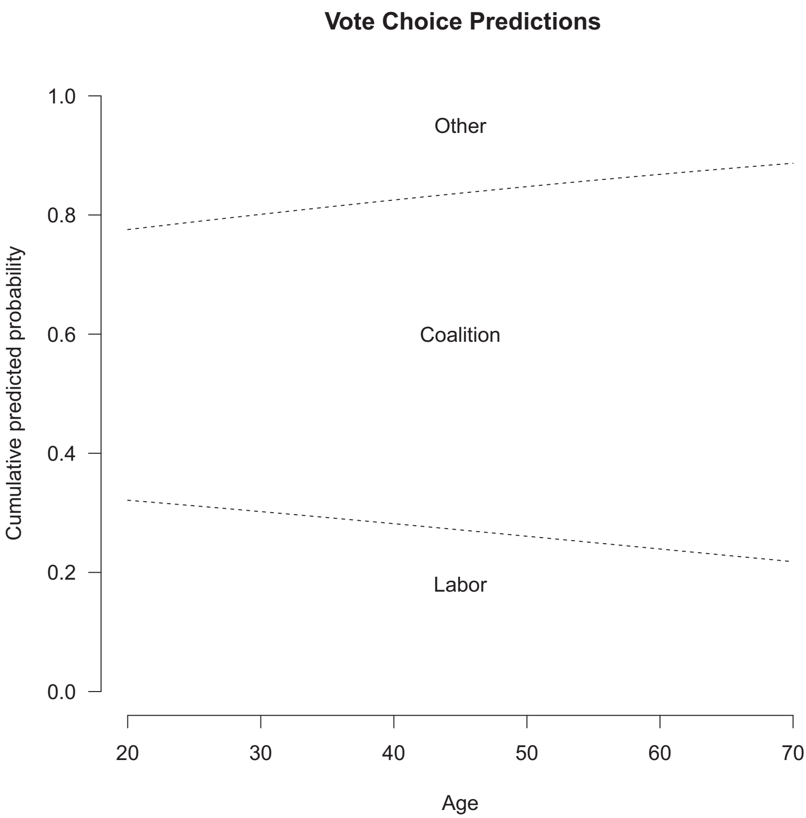


FIGURE 9.4 Predicted vote choice in the 2013 Australian federal elections across age cohorts. Older voters are more likely to support the Liberal-National Coalition.

is the dreaded *odds ratio*. Since the multinomial logit is a log-odds model, it may be useful to note that the log of the ratio of two probabilities is a linear function of the independent variables:

$$\log\left[\frac{\Pr(Y_i = m|\mathbf{x}_i)}{\Pr(Y_i = d|\mathbf{x}_i)}\right] = \mathbf{x}_i^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_d).$$

Since we set the coefficients of one category – the baseline – to zero for identification, we can calculate the log odds that i is in m relative to the baseline using:

$$\log\left[\frac{\Pr(Y_i = m|\mathbf{x}_i)}{\Pr(Y_i = 1|\mathbf{x}_i)}\right] = \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_m.$$

This approach is linear in the parameters. We can calculate hypothetical changes in the odds ratio for category m associated with a particular covariate x_j by exponentiation (i.e., $\exp(\hat{\beta}_{m,j})$). In this way we can inspect Table 9.1

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

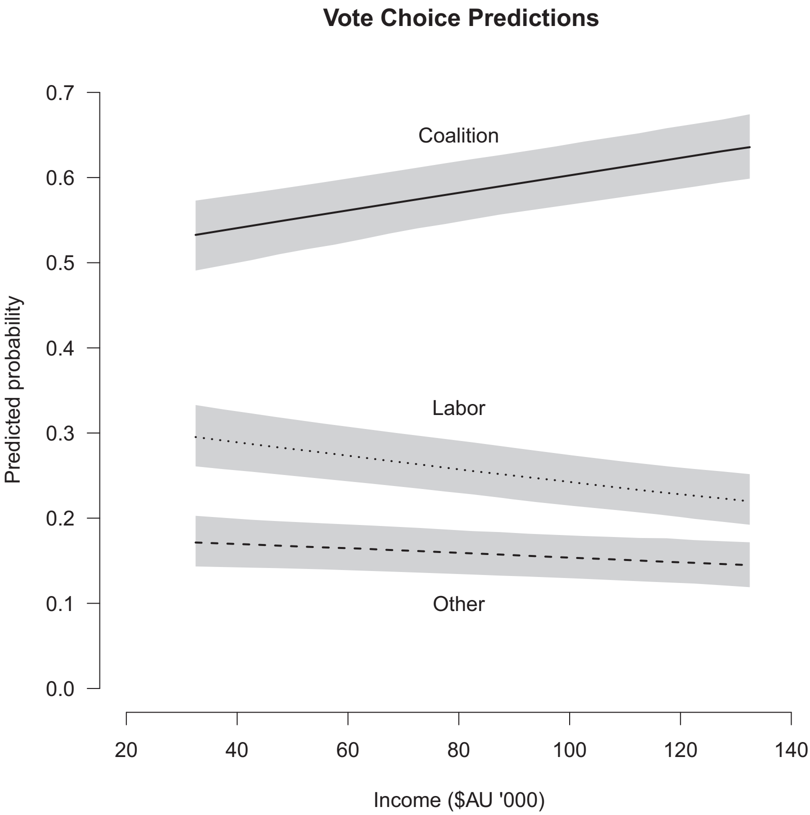


FIGURE 9.5 Predicted vote choice in the 2013 Australian federal elections for different income levels. Higher income voters were more likely to support the Liberal-National Coalition and less likely to vote for Labor.

and see that a Catholic has 21% lower odds of voting for the Coalition over Labor relative to a Protestant ($\exp(-0.24) = 0.79$); this relationship is large relative to the estimated standard error. However, such statements are cumbersome because the model is about comparing several categories. Simply exponentiating coefficients (or their differences) privileges some comparisons over others. Standard errors and statements of significance may vary depending on the comparison made. For example, looking back at Table 9.1, we see that Catholic is a "significant" predictor of voting for Labor relative to Other (left half), but it is not a "significant" predictor of voting for the Coalition relative to Other (right half). Relying on just the BUTON to view and interpret the implications of a multinomial model may not be the most effective means of communicating the model's substantive implications to audiences. It certainly fails to display many quantities that may be of interest to both readers and analysts.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 9.3 *Ternary plots and confidence regions*

```
B<-mvrnorm(1000,
  mu=c(coef(mnl.fit)[1,],coef(mnl.fit)[2,]),
  Sigma=vcov(mnl.fit))
X.u<-c(1,median(myoz$income2[myoz$union=="Yes"]),
  1,0,0,0,0,median(myoz$age[myoz$union=="Yes"]))
X.nu<-c(1,median(myoz$income2[myoz$union=="Yes"]),
  0,0,0,0,0,median(myoz$age[myoz$union=="Yes"]))

k<-dim(coef(mnl.fit))[2]
denom.u<- 1+exp(B[,1:k]%*%X.u)+exp(B[,(k+1):(2*k)]%*%X.u)  #denominator of
     multinomial
denom.nu<-1+exp(B[,1:k]%*%X.nu)+exp(B[,(k+1):(2*k)]%*%X.nu)#denominator of
     multinomial
pp.coal.u<-exp(B[,1:k]%*%X.u)/denom.u
pp.other.u<-exp(B[,(k+1):(2*k)]%*%X.u)/denom.u
pp.coal.nu<-exp(B[,1:k]%*%X.nu)/denom.nu
pp.other.nu<-exp(B[,(k+1):(2*k)]%*%X.nu)/denom.nu
union.pp<-cbind(pp.coal.u, pp.other.u, (1-pp.coal.u-pp.other.u))
nounion.pp<-cbind(pp.coal.nu, pp.other.nu, (1-pp.coal.nu-pp.other.nu))
colnames(union.pp)<-colnames(nounion.pp)<-c("Coalition", "Other", "Labor")

library(compositions)
#Getting ellipse radius for CR
#See van den Boogaart and Tolosana-Delgado 2013 p. 83
df1 = ncol(union.pp)-1
df2 = nrow(union.pp)-ncol(union.pp)+1
rconf = sqrt( qf(p=0.95, df1, df2)*df1/df2 )
rprob = sqrt( qchisq(p=0.95, df=df1) )

plot(acomp(union.pp), col = grey(0.8))
plot(acomp(nounion.pp), col = grey(0.8), pch=3, add=T)
isoPortionLines(, col=grey(0.5), lty=2, lwd=.7)
ellipses(mean(acomp(union.pp)),var(acomp(union.pp)),r=rprob,col="red", lwd=2)
ellipses(mean(acomp(nounion.pp)),var(acomp(nounion.pp)),r=rprob,col="blue",
     lwd=2)
text(.35,.2,"non\n union")
text(.52,.4,"union")
```

Another approach to interpretation – marginal effects calculations – does include comparisons across multiple outcome categories. To see this, we calculate the partial derivative for a particular covariate $j \in \{1, 2, \ldots, k\}$:

$$\frac{\partial \Pr(Y_i = m)}{\partial x_{ij}} = \Pr(Y_i = m|\mathbf{x}_i)\left[\beta_{jm} - \sum_{d \in S} \beta_{jd} \times \Pr(Y_i = d|\mathbf{x}_i)\right].$$

This expression depends not only on the values of the other covariates but also on all the other regression coefficients, $\beta_d, d \neq m$. As a result, the marginal effect of a particular covariate for a specific category may not have the same sign as an estimated coefficient for covariate j appearing in a BUTON. A full accounting of marginal effects would require yet another BUTON in which we display the

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

marginal effects of each variable for each category, each of which is conditional on a specific scenario embodied in $\mathbf{x}_i$. Whether a particular marginal effect is large relative to estimation uncertainty will depend on the scenario chosen, the coefficient for the variable of interest, and all the other coefficients in the model. As a result, this approach to interpretation is often best avoided.

9.4 CONDITIONAL MULTINOMIAL LOGIT

Thus far we have described categorical outcomes in terms of the characteristics of the "chooser," i. But the multinomial model can be flipped around. Suppose we are instead interested in how attributes of the categories relate to choices. In the voting example we might consider how party platforms or candidate attributes might make them more or less attractive to voters.

To formalize this version of the model, suppose we again have $k-1$ covariates. But now we let $\mathbf{w}_{im}$ be a k-vector with values describing how category m is experienced by individual i. The key distinction here is that regressors now vary across choice categories; they can also vary across individuals. For example, in the Australian survey data, outcome-level covariates might include the distance between a voter's left-right placement and that for each party; whether the respondent voted for party m in the last election; or whether party m contacted the voter in the course of this election campaign. All these clearly differ across parties for each respondent as well as across respondents. This leads to an altered model formulation:

$$\Pr(Y_i = m) = \frac{\exp(\mathbf{w}_{im}^\top \boldsymbol{\delta})}{\sum_{h=1}^{M} \exp(\mathbf{w}_{ih}^\top \boldsymbol{\delta})}.$$

The parameter vector $\boldsymbol{\delta}$ has a constant and a regression coefficient for each of $k-1$ regressors. This formulation goes by the name *conditional (multinomial) logit*.[3] Note that in the conditional logit model the *covariate values* differ across categories and the parameters are constant, whereas in the multinomial model the covariate values are fixed for each individual across choice categories, and the parameters differ. Another way to see this is to consider the ratio of the probabilities of choosing m over d under the conditional logit:

$$\frac{\Pr(Y_i = m)}{\Pr(Y_i = d)} = \exp[(\mathbf{w}_{im} - \mathbf{w}_{id})^\top \boldsymbol{\delta}]. \tag{9.6}$$

[3] Note a "conditional logit" model means different things in different disciplines. In economics, political science, and sociology "conditional logit" usually refers to the model described in this section, following McFadden (1974). In epidemiology, "conditional logit" refers to a matched case-control logit model, sometimes called "fixed effects/panel logit" in other disciplines. It is possible to show that the McFadden conditional logit is a special case of the fixed effects version, but we do not take that up here. Note that STATA has different model commands for each of these.

Equation 9.6 is just a restatement of the IIA condition. Comparing Equation 9.4 to Equation 9.6 highlights the difference between in the multinomial and conditional logit specifications.

Of course the multinomial and the conditional logits can each be viewed as special cases of the other. Their log-likelihood and methods of interpretation are nearly identical. It is also possible to build a model that combines both category- and chooser-specific covariates:

$$\Pr(Y_i = m) = \frac{\exp(\mathbf{w}_{im}^{\mathsf{T}}\boldsymbol{\delta} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_m)}{\sum_{h=1}^{M}\exp(\mathbf{w}_{ih}^{\mathsf{T}}\boldsymbol{\delta} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_h)}.$$

To illustrate these models, we reanalyze the 2013 Australian election data, including a party-level covariate that also differs by respondent: whether party m personally contacted the respondent. We fit two models. The first is a conditional multinomial logit including only the contact variable; we allow for differing intercepts across categories. The second is a combined model that includes the party-level contact variable as well as the same individual-level covariates as above. Coalition is the reference category for both models. The resulting BUTON for coefficients appears as Table 9.5.

The models in Table 9.5 present a useful revision to the model in Table 9.1. Using the BIC, we see that the mixed conditional/multinomial model outperforms both the conditional logit and the multinomial logit in Table 9.1. The new contact variable is precisely estimated away from zero. To a first approximation, a respondent is about 43% more likely to vote for a party that contacted her than one that did not ($\exp(0.36) \approx 1.43$). Adding in other covariates does not alter this basic conclusion. Other coefficient estimates are nearly identical to those in the multinomial model.

9.4.1 A Note on Data Structure

In terms of estimation, the trick involved in going from a multinomial model to a conditional or mixed multinomial involves the structure of the data set. In the pure multinomial model with just individual-specific covariates, the data are in a standard rectangular matrix in which the rows are individuals and the columns are variables. Table 9.6 displays a clip of the 2013 Australia election data in this form. Note how the contact variable is broken out by response category.

In order to estimate a conditional or mixed model, however, we need to think of the response category as the unit of analysis. But these response categories are now grouped by the individual or "case." Table 9.7 displays the same clip of the 2013 Australia data in this form. Each row is a person-category; each person accounts for M rows of the data set. The data set is now one $M \times k$ matrix for each individual, stacked on top of one another. Note that the category-level contact variable has been collapsed to a single column.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

TABLE 9.5 *Conditional and mixed conditional-multinomial logistic regression on vote choice for the 2013 Australian elections. Coalition is the reference category.*

	Conditional	**Mixed**
Contact	0.36	0.33
	(0.06)	(0.06)
Labor: intercept	−0.32	0.32
	(0.04)	(0.21)
Other: intercept	−0.72	0.04
	(0.05)	(0.24)
Labor: income		−0.05
		(0.01)
Other: income		−0.03
		(0.01)
Labor: union member		1.02
		(0.10)
Other: union member		0.81
		(0.12)
Labor: Catholic		0.24
		(0.11)
Other: Catholic		−0.11
		(0.14)
Labor: not religious		0.71
		(0.11)
Other: not religious		1.09
		(0.12)
Labor: other religion		0.22
		(0.13)
Other: other religion		0.43
		(0.15)
Labor: female		0.06
		(0.08)
Other: female		0.11
		(0.10)
Labor: age		−0.01
		(0.00)
Other: age		−0.02
		(0.00)
n	3,342	3,342
$\log \mathcal{L}$	−3,476	−3,302
AIC	6,957	6,639
BIC	6,976	6,743

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

TABLE 9.6 *A typical rectangular data structure for a multinomial model. Note the inclusion of the choice-specific variables.*

							Contact		
Respondent	**Vote Choice**	**Income**	**Union**	**Religion**	**Sex**	**Age**	**Labor**	**Coalition**	**Other**
1	Labor	20.0	No	Protestant	Female	43	1	1	0
2	Labor	20.0	No	Protestant	Male	61	1	1	1
3	Other	15.5	Yes	None	Male	52	1	1	1
4	Coalition	7.5	No	Protestant	Female	69	1	1	1
5	Other	1.2	No	Other	Female	61	1	1	1
6	Labor	9.5	Yes	None	Female	20	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3,342	Other	9.5	Yes	None	Female	40	1	1	0

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

TABLE 9.7 *A grouped-response data structure data structure enabling conditional and mixed logit estimation. Each row is a person-category.*

Respondent	Category	Vote Choice	Contact	Income	Union	Religion	Sex	Age
1	Coalition	FALSE	1	20	No	Protestant	Female	43
1	Labor	TRUE	1	20	No	Protestant	Female	43
1	Other	FALSE	0	20	No	Protestant	Female	43
2	Coalition	FALSE	1	20	No	Protestant	Male	61
2	Labor	TRUE	1	20	No	Protestant	Male	61
2	Other	FALSE	1	20	No	Protestant	Male	61
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3,342	Coalition	FALSE	1	9.5	Yes	None	Female	40
3,342	Labor	FALSE	1	9.5	Yes	None	Female	40
3,342	Other	TRUE	0	9.5	Yes	None	Female	40

The data in Tables 9.6 and 9.7 contain exactly the same information. Reformatting only serves computational convenience. Understanding how the data are organized is important for understanding how to effectively sample from the limiting distribution of the parameters and generate predicted probabilities.

Viewing the data in grouped-response format has the additional benefit of highlighting the connection between a conditional or mixed multinomial logit model and panel data. A panel data set in which each individual is observed repeatedly is also typically organized as stacked individual-level matrices. In fact a "conditional logit" is one way of analyzing binary panel data, something we take up further in Chapter 11.

9.5 EXTENSIONS

All of the models above rely on the IIA assumption. The IIA, in turn, results from the assumption that (1) the errors are independent across *categories*, (2) the errors are identically distributed, and (3) the errors all follow as EV-I distribution. A variety of alternatives have been developed to partially relax IIA or, equivalently, allow for unequal variance or correlation across the outcome categories. In formulating the likelihood, all these models continue to maintain the assumption of (conditional) independence across individuals, i.

9.5.1 Heteroskedastic Multinomial Logit

The *heteroskedastic multinomial logit* model, sometimes called the *heteroskedastic extreme value* model, allows the error variance to differ across outcome categories. This model might be useful in situations in which we are worried that a change in a covariate might produce different rates of substitution across choices (Bhat, 1996).[4]

Formally, we can restate the latent variable formulation as $U_i(m) = \mu_i(m) + \sigma_m \epsilon_{im}$, where $\epsilon_{im} \sim f_{EV_1}(1, 0)$.[5] To fix the model we require that $\sigma_1 = 1$ for the reference category. In this way we can write

$$\Pr(Y_i = m) = \frac{\exp\left[(\mathbf{w}_{im}^{\intercal}\boldsymbol{\delta} + \mathbf{x}_i^{\intercal}\boldsymbol{\beta}_m)/\sigma_m\right]}{\sum_{h=1}^{M} \exp\left[(\mathbf{w}_{ih}^{\intercal}\boldsymbol{\delta} + \mathbf{x}_i^{\intercal}\boldsymbol{\beta}_h)/\sigma_h\right]}. \tag{9.7}$$

Fitting the heteroskedasic model entails estimating an additional $M - 1$ scale parameters. Reestimating the mixed model in Table 9.5, allowing for heteroskedasticity, yields scale parameters that are indistinguishable from 1 and a BIC of 6,759.[6] In the 2013 Australia election data there is no benefit for this additional complication.

[4] In a conditional multinomial logit, the "marginal rate of substitution" between outcomes in terms of outcome-level covariates k and l is $\frac{\delta_k}{\delta_l}$.

[5] Or, equivalently, assume that $\epsilon_{im} \sim f_{EV_1}(1, \sigma_m)$.

[6] Recall that the standard multinomial model fixes $\sigma = 1$. So, when evaluating the $\hat{\sigma}_j$ from a heteroskedastic model we are interested in the null hypothesis that $\sigma_j = 1$.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

9.5.2 Nested Logit

The *nested logit* model partially relaxes the assumption that outcome categories are uncorrelated with each other. With this model we imagine that the choice set has subcomponents or nests. Formally, S can be partitioned into D disjoint subsets, $A_1, A_2, \ldots, A_D$. Under nested logit we retain the IIA assumption *within* nests but not across them. The *cross-nested logit* allows for choices to appear in more than one nest, but we do not take this up further in this volume.

One way for nested choice to arise in the real world is when people make choices sequentially: past choices affect the subsequent choice options. For example, Figure 9.6 displays part of the ballot confronting California voters in 2003. They were first asked to decide whether the incumbent governor, Gray Davis, should be removed from office. They were then asked who should replace him, presenting voters with a list of 135 candidates, including actors Arnold Schwarzenegger and Gary Coleman along with infamous publisher Larry Flynt. Voters were presented with the sets $A = \{$"Gray Davis", "not Gray Davis"$\}$ and then presented with $B = \{$"Gary Coleman", $\ldots\}$.

Conceptually the nested logit involves modeling $\Pr(Y_i = m)$ as the probability that the nest containing m is chosen times the probability that m is selected from among the choices in that nest:

$$\Pr(Y_i = m) = \Pr(A_d) \times \Pr(Y_i = m | m \in A_d) \quad \forall d \in \{1, \ldots, D\}. \tag{9.8}$$

The nested logit can be extended to several partitions and several layers of nesting, but we will only consider two layers here.

To fix the model, we can decompose each part of the product in Equation 9.8.[7] Let $V_{im} = \mathbf{w}_{im}^{\mathsf{T}}\boldsymbol{\delta} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_m$. This is simply the linear predictor term, but in the context of a choice situation, we can think of V_{im} as being the systematic part of i's utility function, evaluated for category m. The probability of choosing m out of all the elements of the nest A_d is simply a multinomial logit:

$$\Pr(Y_i = m | m \in A_d) = \frac{\exp(V_{im}/\lambda_d)}{\sum_{j \in A_d} \exp(V_{ij}/\lambda_d)}. \tag{9.9}$$

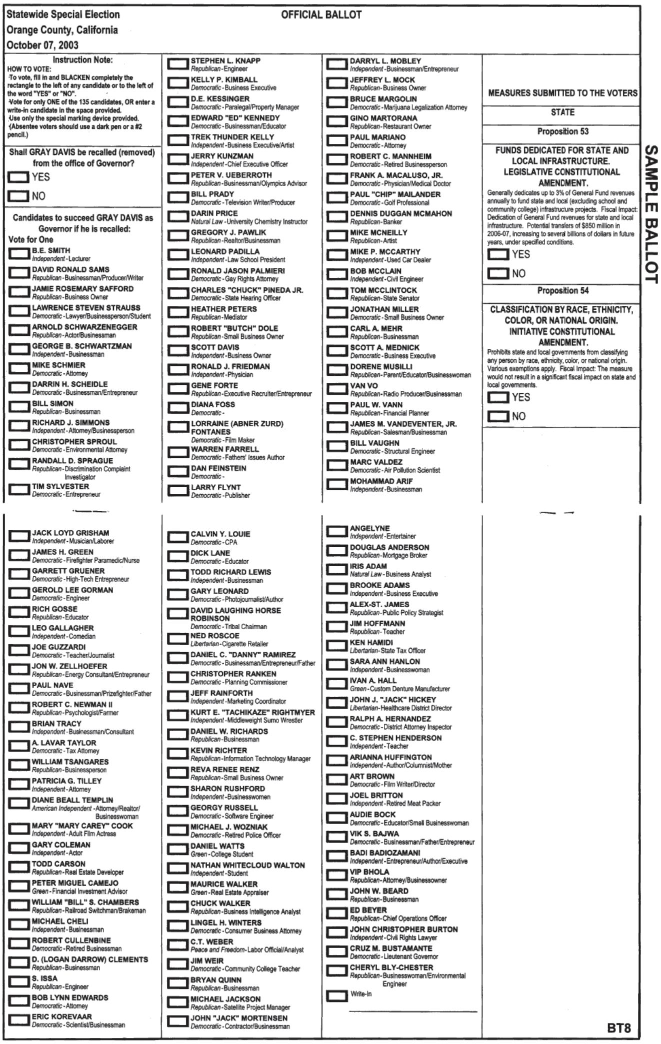
Equation 9.9 is similar to the expression for the heteroskedastic model in Equation 9.7, only here the error variances differ across nests rather than across categories. In the standard multinomial model the parameter $\lambda_d = 1$. To the extent that λ_d approaches 0 there is more homogeneity within the categories in nest d.[8]

Let $W_d = \mathbf{q}_d^{\mathsf{T}}\boldsymbol{\alpha}_d$ be any nest-level covariates (if such things exist) and associated parameters. We also define $Z_d = \log \sum_{m \in A_d} \exp(V_{im}/\lambda_d)$. Z_d is called the *inclusive value* or *log-sum*, and λ_d is the inclusive value or log-

[7] The nested logit model can also be derived from the latent variable constructing by assuming that the $M-$vector of errors, ϵ_i, follows a Generalized Extreme Value distribution.

[8] A λ_d outside the (0,1] interval is commonly viewed as evidence of model misspecification.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press



Statewide Special Election
Orange County, California
October 07, 2003

OFFICIAL BALLOT

Instruction Note:
HOW TO VOTE:
·To vote, fill in and BLACKEN completely the rectangle to the left of any candidate or to the left of the word "YES" or "NO".
·Vote for only ONE of the 135 candidates, OR enter a write-in candidate in the space provided.
·Use only the special marking device provided.
·(Absentee voters should use a dark pen or a #2 pencil.)

Shall GRAY DAVIS be recalled (removed) from the office of Governor?
YES
NO

Candidates to succeed GRAY DAVIS as Governor if he is recalled:
Vote for One

B.E. SMITH *Independent* - Lecturer
DAVID RONALD SAMS *Republican* - Businessman/Producer/Writer
JAMIE ROSEMARY SAFFORD *Republican* - Business Owner
LAWRENCE STEVEN STRAUSS *Democratic* - Lawyer/Businessperson/Student
ARNOLD SCHWARZENEGGER *Republican* - Actor/Businessman
GEORGE B. SCHWARTZMAN *Independent* - Businessman
MIKE SCHMIER *Democratic* - Attorney
DARRIN H. SCHEIDLE *Democratic* - Businessman/Entrepreneur
BILL SIMON *Republican* - Businessman
RICHARD J. SIMMONS *Independent* - Attorney/Businessperson
CHRISTOPHER SPROUL *Democratic* - Environmental Attorney
RANDALL D. SPRAGUE *Republican* - Discrimination Complaint Investigator
TIM SYLVESTER *Democratic* - Entrepreneur
STEPHEN L. KNAPP *Republican* - Engineer
KELLY P. KIMBALL *Democratic* - Business Executive
D.E. KESSINGER *Democratic* - Paralegal/Property Manager
EDWARD "ED" KENNEDY *Democratic* - Businessman/Educator
TREK THUNDER KELLY *Independent* - Business Executive/Artist
JERRY KUNZMAN *Independent* - Chief Executive Officer
PETER V. UEBERROTH *Republican* - Businessman/Olympics Advisor
BILL PRADY *Democratic* - Television Writer/Producer
DARIN PRICE *Natural Law* - University Chemistry Instructor
GREGORY J. PAWLIK *Republican* - Realtor/Businessman
LEONARD PADILLA *Independent* - Law School President
RONALD JASON PALMIERI *Democratic* - Gay Rights Attorney
CHARLES "CHUCK" PINEDA JR. *Democratic* - State Hearing Officer
HEATHER PETERS *Republican* - Mediator
ROBERT "BUTCH" DOLE *Republican* - Small Business Owner
SCOTT DAVIS *Independent* - Business Owner
RONALD J. FRIEDMAN *Independent* - Physician
GENE FORTE *Republican* - Executive Recruiter/Entrepreneur
DIANA FOSS *Democratic* -
LORRAINE (ABNER ZURD) FONTANES *Democratic* - Film Maker
WARREN FARRELL *Democratic* - Fathers' Issues Author
DAN FEINSTEIN *Democratic* -
LARRY FLYNT *Democratic* - Publisher
DARRYL L. MOBLEY *Independent* - Businessman/Entrepreneur
JEFFREY L. MOCK *Republican* - Business Owner
BRUCE MARGOLIN *Democratic* - Marijuana Legalization Attorney
GINO MARTORANA *Republican* - Restaurant Owner
PAUL MARIANO *Democratic* - Attorney
ROBERT C. MANNHEIM *Democratic* - Retired Businessperson
FRANK A. MACALUSO, JR. *Democratic* - Physician/Medical Doctor
PAUL "CHIP" MAILANDER *Democratic* - Golf Professional
DENNIS DUGGAN MCMAHON *Republican* - Banker
MIKE MCNEILLY *Republican* - Artist
MIKE P. MCCARTHY *Independent* - Used Car Dealer
BOB MCCLAIN *Independent* - Civil Engineer
TOM MCCLINTOCK *Republican* - State Senator
JONATHAN MILLER *Democratic* - Small Business Owner
CARL A. MEHR *Republican* - Businessman
SCOTT A. MEDNICK *Democratic* - Business Executive
DORENE MUSILLI *Republican* - Parent/Educator/Businesswoman
VAN VO *Republican* - Radio Producer/Businessman
PAUL W. VANN *Republican* - Financial Planner
JAMES M. VANDEVENTER, JR. *Republican* - Salesman/Businessman
BILL VAUGHN *Democratic* - Structural Engineer
MARC VALDEZ *Democratic* - Air Pollution Scientist
MOHAMMAD ARIF *Independent* - Businessman

MEASURES SUBMITTED TO THE VOTERS
STATE
Proposition 53
FUNDS DEDICATED FOR STATE AND LOCAL INFRASTRUCTURE. LEGISLATIVE CONSTITUTIONAL AMENDMENT.
Generally dedicates up to 3% of General Fund revenues annually to fund state and local (excluding school and community college) infrastructure projects. Fiscal Impact: Dedication of General Fund revenues for state and local infrastructure. Potential transfers of $850 million in 2006-07, increasing to several billions of dollars in future years, under specified conditions.
YES
NO
Proposition 54
CLASSIFICATION BY RACE, ETHNICITY, COLOR, OR NATIONAL ORIGIN. INITIATIVE CONSTITUTIONAL AMENDMENT.
Prohibits state and local governments from classifying any person by race, ethnicity, color, or national origin. Various exemptions apply. Fiscal Impact: The measure would not result in a significant fiscal impact on state and local governments.
YES
NO

SAMPLE BALLOT

JACK LOYD GRISHAM *Independent* - Musician/Laborer
JAMES H. GREEN *Democratic* - Firefighter Paramedic/Nurse
GARRETT GRUENER *Democratic* - High-Tech Entrepreneur
GEROLD LEE GORMAN *Democratic* - Engineer
RICH GOSSE *Republican* - Educator
LEO GALLAGHER *Independent* - Comedian
JOE GUZZARDI *Democratic* - Teacher/Journalist
JON W. ZELLHOEFER *Republican* - Energy Consultant/Entrepreneur
PAUL NAVE *Democratic* - Businessman/Prizefighter/Father
ROBERT C. NEWMAN II *Republican* - Psychologist/Farmer
BRIAN TRACY *Independent* - Businessman/Consultant
A. LAVAR TAYLOR *Democratic* - Tax Attorney
WILLIAM TSANGARES *Republican* - Businessperson
PATRICIA G. TILLEY *Independent* - Attorney
DIANE BEALL TEMPLIN *American Independent* - Attorney/Realtor/Businesswoman
MARY "MARY CAREY" COOK *Independent* - Adult Film Actress
GARY COLEMAN *Independent* - Actor
TODD CARSON *Republican* - Real Estate Developer
PETER MIGUEL CAMEJO *Green* - Financial Investment Advisor
WILLIAM "BILL" S. CHAMBERS *Republican* - Railroad Switchman/Brakeman
MICHAEL CHELI *Independent* - Businessman
ROBERT CULLENBINE *Democratic* - Retired Businessman
D. (LOGAN DARROW) CLEMENTS *Republican* - Businessman
S. ISSA *Republican* - Engineer
BOB LYNN EDWARDS *Democratic* - Attorney
ERIC KOREVAAR *Democratic* - Scientist/Businessman
CALVIN Y. LOUIE *Democratic* - CPA
DICK LANE *Democratic* - Educator
TODD RICHARD LEWIS *Independent* - Businessman
GARY LEONARD *Democratic* - Photojournalist/Author
DAVID LAUGHING HORSE ROBINSON *Democratic* - Tribal Chairman
NED ROSCOE *Libertarian* - Cigarette Retailer
DANIEL C. "DANNY" RAMIREZ *Democratic* - Businessman/Entrepreneur/Father
CHRISTOPHER RANKEN *Democratic* - Planning Commissioner
JEFF RAINFORTH *Independent* - Marketing Coordinator
KURT E. "TACHIKAZE" RIGHTMYER *Independent* - Middleweight Sumo Wrestler
DANIEL W. RICHARDS *Republican* - Businessman
KEVIN RICHTER *Republican* - Information Technology Manager
REVA RENEE RENZ *Republican* - Small Business Owner
SHARON RUSHFORD *Independent* - Businesswomen
GEORGY RUSSELL *Democratic* - Software Engineer
MICHAEL J. WOZNIAK *Democratic* - Retired Police Officer
DANIEL WATTS *Green* - College Student
NATHAN WHITECLOUD WALTON *Independent* - Student
MAURICE WALKER *Green* - Real Estate Appraiser
CHUCK WALKER *Republican* - Business Intelligence Analyst
LINGEL H. WINTERS *Democratic* - Consumer Business Attorney
C.T. WEBER *Peace and Freedom* - Labor Official/Analyst
JIM WEIR *Democratic* - Community College Teacher
BRYAN QUINN *Republican* - Businessman
MICHAEL JACKSON *Republican* - Satellite Project Manager
JOHN "JACK" MORTENSEN *Democratic* - Contractor/Businessman
ANGELYNE *Independent* - Entertainer
DOUGLAS ANDERSON *Republican* - Mortgage Broker
IRIS ADAM *Natural Law* - Business Analyst
BROOKE ADAMS *Independent* - Business Executive
ALEX-ST. JAMES *Republican* - Public Policy Strategist
JIM HOFFMANN *Republican* - Teacher
KEN HAMIDI *Libertarian* - State Tax Officer
SARA ANN HANLON *Independent* - Businesswoman
IVAN A. HALL *Green* - Custom Denture Manufacturer
JOHN J. "JACK" HICKEY *Libertarian* - Healthcare District Director
RALPH A. HERNANDEZ *Democratic* - District Attorney Inspector
C. STEPHEN HENDERSON *Independent* - Teacher
ARIANNA HUFFINGTON *Independent* - Author/Columnist/Mother
ART BROWN *Democratic* - Film Writer/Director
JOEL BRITTON *Independent* - Retired Meat Packer
AUDIE BOCK *Democratic* - Educator/Small Businesswoman
VIK S. BAJWA *Democratic* - Businessman/Father/Entrepreneur
BADI BADIOZAMANI *Independent* - Entrepreneur/Author/Executive
VIP BHOLA *Republican* - Attorney/Businessowner
JOHN W. BEARD *Republican* - Businessman
ED BEYER *Republican* - Chief Operations Officer
JOHN CHRISTOPHER BURTON *Independent* - Civil Rights Lawyer
CRUZ M. BUSTAMANTE *Democratic* - Lieutenant Governor
CHERYL BLY-CHESTER *Republican* - Businesswoman/Environmental Engineer
Write-In

BT8

FIGURE 9.6 An example of sequential choices confronting voters, from the 2003 California gubernatorial recall election.

sum coefficient commonly reported. With these expressions we can now define another logit,

$$\Pr(A_d) = \frac{\exp(W_d + \lambda_d^{\intercal} Z_d)}{\sum_{\ell=1}^{D} \exp(W_\ell + \lambda_\ell^{\intercal} Z_\ell)}.$$

The nested logit requires that we pre-specify the set of nests or meta categories. In most situations there are several possible nesting structures, and

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press

it may not be obvious which is "correct." In the California recall example, state law held that only those voting in favor of the recall could cast valid votes for the successor, conforming to the nested logit structure and providing an obvious way to construct a nesting structure. This provision was challenged in court and found unconstitutional during the 2003 recall campaign. As a result, in the actual election, voters could vote for *both* the retention of Gray Davis as well as his successor, a violation of the nesting assumption. In the 2013 Australian election data, we might imagine that voters first decide whether to vote for a mainstream party (Labor or Coalition) or to cast a protest vote (Other). Alternatively, we might imagine that voters first decide whether to vote left (Labor, Other) or right (Coalition). Evaluating these options involves another layer of model selection, requiring tools such as likelihood ratios, BIC, and out-of-sample evaluation. In the particular case of the Australian survey data, there is no benefit to including either of these nesting structures, based on the BIC.

9.5.3 Multinomial Probit

Recall the random utility specification: $U_i(m) = \mu_i(m) + \epsilon_{im}$. If we assume

$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iM}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

we arrive at the multinomial probit model. The covariance matrix $\boldsymbol{\Sigma}_{M\times M}$ allows for arbitrary correlation across choice categories. In other words, the multinomial probit does not require the IIA assumption. But the cost is a much heavier computational burden. To see this, note the choice probability for individual i:

$$\begin{aligned}\Pr(Y_i = m) &= \Pr(U_i(m) > U_i(d) \quad \forall \quad d \neq m \in S)\\ &= \Pr(\epsilon_{im} - \epsilon_{id} > \mu_i(d) - \mu_i(m) \quad \forall \quad d \neq m \in S).\end{aligned}$$

Since one category is fixed as a reference, this second expression involves numerically evaluating an $(M-1)$-dimensional integral. In general the researcher must place some constraints on the covariance matrix to identify the model.

9.5.4 Random Coefficients and Mixed Logit

The *mixed logit* model is a further generalization of the multinomial logit model that allows us to model arbitrary dependence across categories and relax the IIA assumption. The easiest way to conceive of a mixed logit is as a "random coefficient" model. That is, we imagine that the regression parameters are heterogeneous across the population but governed by some underlying distribution. We can then average across the possible values of the $\boldsymbol{\beta}$ to recover the choice probabilities across categories. In this model the $\boldsymbol{\beta}$ varies according to some distribution, with its own parameters given by $\boldsymbol{\theta}$. It is common to assume a normal distribution, implying that $\boldsymbol{\theta} = (\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, although others are feasible.

The mixed logit retains the logit probability, but treats it as a function of $\boldsymbol{\beta}$. That is,

$$\Pr(Y_i = m|\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{im}^{\mathsf{T}}\boldsymbol{\beta}_i)}{\sum_{\ell=1}^{M}\exp(\mathbf{x}_{i\ell}^{\mathsf{T}}\boldsymbol{\beta}_i)}.$$

But now we must specify the distribution for $\boldsymbol{\beta}$ and then integrate over it in the process of maximization:

$$\Pr(Y_i = m) = \int_{\boldsymbol{\beta}} \Pr(Y_i = m|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta}.$$

The necessity of integrating over $\boldsymbol{\beta}$ complicates estimation substantially, but simulation methods and modern computing power combine to make this problem surmountable. Mixed logit is widely viewed as the most flexible approach to relaxing IIA in the context of nominal data. It can accommodate a wider variety of distributional assumptions than the multinomial probit and, in many cases, is faster to estimate. Refitting the models from Table 9.5 as a mixed logits, treating the contact variable random, and following a normal distribution adds nothing. The original models are preferred on a BIC and likelihood ratio basis. In the 2013 Australia data there is no evidence of violations of IIA necessitating further complications.

9.6 CONCLUSION

In this chapter we generalized our treatment of binary and ordered categorical variables to include multi-category, unordered outcomes. The standard model is the multinomial logit, which can accomodate covariates describing attributes of both the outcome categories as well as the units. The multinomial logit model relies on the assumed IIA, which may be too restrictive for some applications. Several extensions provide ways to relax this assumption.

The generality of the multinomial model comes at a cost. Multinomial models ask a lot of the data, estimating a large number of parameters. These models can become quite rich – and complex – very rapidly with the inclusion of additional covariates or outcome categories. Decisions about what to include or leave out necessitate a hierarchy of model evaluation decisions. Once estimated, the presentation and interpretation of model results also entails additional care and effort, all the more so when the models involve nested outcome categories or hierarchical, random coefficients.

9.7 FURTHER READING

Applications

Eifert et al. (2010) use multinomial logit to analyze the choice of identity group in a collection of African countries. Glasgow (2001) profitably uses the mixed logit to analyze voter behavior in multi-party UK elections. Martin and Stevenson (2010) use the conditional logit to examine coalition bargaining and government formation.

Past Work

Alvarez and Nagler (1998) present an early systematic comparison of several of the models described in this chapter; also see Alvarez (1998).

Advanced Study

Train (2009) is the more-advanced text on the specification and computation of a variety of mulitnomial models. Bagozzi (2016) presents a "zero-inflated" multinomial model with application to international relations. Mebane and Sekhon (2004) discuss and extend the use of the multinomial model in the context of counts across categories. See van den Boogaart and Tolosana-Delgado (2013) on the calculation and plotting of confidence ellipses for compositional data.

Software Notes

In $\mathcal{R}$, the `multinom` function in the `nnet` package (Ripley and Venables, 2016; Venables and Ripley, 2002) only handles rectangular data and only fits a standard multinomial model with individual-level covariates. The `mlogit` (Croissant, 2013) and `mnlogit` (Hasan et al., 2016) libraries contain tools for restructuring data sets and estimating all the models discussed in this chapter. The `MNP` library (Imai and van Dyk, 2005) fits a Bayesian multinomial probit. The `compositions` library (van den Boogaart et al., 2014) has a variety of functions for working with compositional data such as the probability simplex.

https://doi.org/10.1017/9781316888544.012 Published online by Cambridge University Press