Discrete Choice Models I

1 Introduction

A discrete choice model is one in which decision makers choose among a set of alternatives.[1] To fit within a discrete choice framework, the set of alternatives – the choice set – needs to exhibit three characteristics: (i) alternatives need to be mutually exclusive, (ii) alternatives must be exhaustive, and (iii) the number of alternatives must be finite. We have already looked at various discrete choice models. For example, binary response models and ordered response models are both types of discrete choice models. Having already examined these specific types of discrete choice models, we are going to take a step back and think about discrete choice models in terms of a more general framework.

1.1 Derivation of Choice Probabilities

In general, discrete choice models are usually derived in a random utility model (RUM) framework in which decision makers are assumed to be utility maximizers.[2] The basic setup is the following. A decision maker, labeled n, faces a choice among J alternatives. The decision maker obtains a certain level of utility from each alternatives. The utility that decision maker n obtains from any alternative j is U_{nj}, $j = 1 \ldots J$. This utility is known to the decision maker but not the analyst. The decision maker chooses the alternative with the highest utility: choose alternative i if and only if $U_{ni} > U_{nj} \forall j \neq i$. The analyst can't observe the decision maker's utility. However, the analyst can observe some attributes of the alternatives, labeled $x_{nj} \forall j$, and some attributes of the decision maker, labeled s_n. The analyst can also specify a function that relates these observed factors to the decision maker's utility. This function is denoted $V_{nj} = V(x_{nj}, s_n) \forall j$ and is called representative utility.

Because there are aspects of utility that the researcher does not or cannot observe, $V_{nj} \neq U_{nj}$. As a result, utility is decomposed as $U_{nj} = V_{nj} + \epsilon_{nj}$, where ϵ_{nj} captures the factors that influence utility but that are not in V_{nj}. In effect, ϵ_{nj} is simply the difference between U_{nj} and V_{nj}. You can think of V_{nj} as the systematic component of a decision maker's utility and ϵ_{nj} as the stochastic component. The researcher does not know $\epsilon_{nj} \forall j$, and therefore treats these terms as random. The joint density of the random vector $\epsilon_n = \{\epsilon_{n1}, \ldots, \epsilon_{nJ}\}$ is denoted $f(\epsilon_n)$. With this density, the analyst can make probability statements about the choice of the decision maker. In other words, the probability that decision maker n choose alternative i is simply:

$$\begin{aligned} P_{ni} &= \text{Prob}(U_{ni} > U_{nj} \forall j \neq i) \\ &= \text{Prob}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj} \forall j \neq i) \\ &= \text{Prob}(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \forall j \neq i) \end{aligned} \tag{1}$$

[1] Much of these notes is heavily based on Ken Train's (2007) excellent book.

[2] Note that this derivation does not rule out other types of behavior. In other words, discrete choice models are consistent with utility maximization, but also other types of behavior.

This probability is a cumulative distribution i.e. the probability that each random term $\epsilon_{nj} - \epsilon_{ni}$ is below the observed quantity $V_{ni} - V_{nj}$. Using the density $f(\epsilon_n)$, this cumulative probability can be written as:

$$\begin{aligned} P_{ni} &= \text{Prob}(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}) \\ &= \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}) f(\epsilon_n) d\epsilon_n \end{aligned} \tag{2}$$

where $I(\cdot)$ is the indicator function equal to 1 when the expression in parentheses is true and 0 otherwise. As you can see, this is a multidimensional integral over the density of the unobserved portion of utility, $f(\epsilon_n)$.

Note that you get different discrete choice models depending on how you specify this density i.e. depending on what assumptions you make about the distribution about the unobserved portion of utility.[3] The integral only takes a closed form solution for certain specifications of $f(\cdot)$. For example, logit and nested logit have closed form solutions; they are derived under the assumption that the unobserved portion of utility is distributed iid extreme value (logit) and a type of generalized extreme value (nested logit). Probit is derived under the assumption that $f(\cdot)$ is multivariate normal and mixed logit is derived under the assumption that the unobserved portion of utility comprises a part that follows any distribution desired by the analyst and a part that is iid extreme value. With probit and mixed logit, the integral has no closed form solution and we have to evaluate it numerically through simulation.

1.2 Identification of Choice Models

1.2.1 Only Differences in Utility Matter

Note that in all discrete choice models, the absolute level of utility is irrelevant to both the decision maker and the analyst. If a constant is added to the utility of all alternatives, then the alternative with the highest utility does not change. Similarly, the level of utility does not matter for the analyst. The choice probability is $P_{ni} = \text{Prob}(U_{ni} > U_{nj}) = \text{Prob}(U_{ni} - U_{nj} > 0)$, which depends only on the difference in utility and not its absolute level. The fact that only differences in utility matter has implications for the identification of discrete choice models. In particular, it means that the only parameters that can be estimated are those that captures differences across alternatives. This can be thought of in a number of ways.

Alternative-Specific Constants

Oftentimes, we will want to specify the observed component of utility to include a constant: $V_{nj} = x_{nj}\beta + k_j$ where k_j is a constant specific to alternative j. This constant captures the average effect on utility of all factors that are not included in the model. When alternative-specific constants are included in the model, ϵ_{nj} has zero mean by construction. However, since only differences in utility matter, only differences in alternative-specific constants matter. Any model with the same difference in constants is equivalent. Put differently, it is impossible to estimate, say, two constants in a two-alternative scenario, because an infinite number of values of the two constants produce the same difference and hence the same choice probabilities. As a result, it is necessary for the analyst to set the overall level of these constants. This is typically done by

[3] You can think of $f(\epsilon_n)$ as the distribution of the unobserved portion of utility within a population of people who face the same observed portion of utility.

normalizing the value of one of the alternative-specific constants. Consider the following example with two alternatives:

$$U_1 = \alpha X_1 + \beta X_1 + k_1^0 + \epsilon_1$$
$$U_2 = \alpha X_2 + \beta X_2 + k_2^0 + \epsilon_2$$

where $K_1^0 - k_2^0 = d$, We can't estimate K^0 and k_2^0 separately. However, we can estimate the following setup:

$$U_1 = \alpha X_1 + \beta X_1 + \epsilon_1$$
$$U_2 = \alpha X_2 + \beta X_2 + k_2 + \epsilon_2$$

where $k_2 = d$, which is the difference in the original unnormalized constants. With J alternatives, at most $J - 1$ alternative-specific constants can be entered, with one constant normalized to 0. It does not matter which constant is normalized to 0, but it does mean that all other constants have to be interpreted relative to whichever one is set to 0.

Sociodemographic Variables

The exact same issue arises with sociodemographic variables i.e. variables relating to attributes of the decision maker. By definition, attributes of the decision maker do not vary over the alternatives. As a result, they can only enter the model in ways that create differences in utility over the alternatives. Consider the following example with two alternatives:

$$U_1 = \alpha X_1 + \beta X_1 + \theta_1^0 Y + \epsilon_1$$
$$U_2 = \alpha X_2 + \beta X_2 + \theta_2^0 Y + \epsilon_2$$

where Y is some sociodemographic variable and θ_1^0 and θ_2^0 capture the effect of this variable on the utility of alternatives 1 and 2 respectively. We might expect that $\theta_1^0 \neq \theta_2^0$ i.e. that the effect of the variable differs across alternatives. The problem is that only differences in utility matter and so we cannot estimate the absolute levels of θ_1^0 and θ_2^0. Again, we have to normalize one of these parameters to 0. Now the model becomes:

$$U_1 = \alpha X_1 + \beta X_1 + \epsilon_1$$
$$U_2 = \alpha X_2 + \beta X_2 + \theta_2 Y + \epsilon_2$$

where $\theta_2 = \theta_2^0 - \theta_1^0$. We now interpret θ_2 as the differential impact of the variable on the utility of alternative 2 compared to alternative 1.

Number of Independent Error Terms

Recall that the choice probabilities in a discrete choice model are:

$$P_{ni} = \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}) f(\epsilon_n) d\epsilon_n \tag{3}$$

This probability is a J-dimensional integral over the density of the J error terms in $\epsilon_n = \{\epsilon_{n1}, \ldots, \epsilon_{nJ}\}$.

However, we can reduce the dimension of the integral by again remembering that only differences in utility matter. With J errors, there are $J-1$ error differences. Thus, we can express the choice probabilities as:

$$\begin{aligned}
P_{ni} &= \text{Prob}(U_{ni} > U_{nj} \forall j \neq i) \\
&= \text{Prob}(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \forall j \neq i) \\
&= \text{Prob}(\tilde{\epsilon}_{nji} < V_{ni} - V_{nj}) \\
&= \int I(\tilde{\epsilon}_{nji} < V_{ni} - V_{nj}) g(\tilde{\epsilon}_{ni}) d\tilde{\epsilon}_{ni}
\end{aligned} \tag{4}$$

where $\tilde{\epsilon}_{nji} = \epsilon_{nj} - \epsilon_{ni}$ is the difference in errors for alternatives i and j. $\tilde{\epsilon}_{ni} = \{\tilde{\epsilon}_{n1i}, \ldots, \tilde{\epsilon}_{nJi}\}$ is the (J-1)-dimensional vector of error differences and $g(\cdot)$ is the density of these error differences. In effect, any model specified with an error for each alternative: $\epsilon_n = \{\epsilon_{n1}, \ldots, \epsilon_{nJ}\}$ with density $f(\epsilon_n)$ is equivalent to a model with J-1 errors defined as $\tilde{\epsilon}_{njk} = \epsilon_{nj} - \epsilon_{nk}$ for any k and density $g(\tilde{\epsilon}_{nk})$ derived from $f(\epsilon_n)$. Because choice probabilities can always be expressed as depending only on $g(\tilde{\epsilon}_{nk})$, one dimension of the density of $f(\epsilon_n)$ is not identified and must be normalized.

The normalization of $f(\epsilon_n)$ is dealt with in different ways. For some models, like the logit model, the distribution of the error terms is sufficiently restrictive that the normalization occurs automatically with the assumptions that come with using the distribution. For other models, such as probit, identification is obtained only by specifying the model in terms of error differences i.e. in terms of $g(\cdot)$ without reference to $f(\cdot)$. The bottom line is that you need to check that your model is properly identified. More on this later.

1.2.2 The Overall Scale of Utility is Irrelevant

We have already seen that adding a constant to the utility of all of the choices does not change the decision maker's choice. Multiplying the utility of all of the choices does not change his choice either. In other words, the alternative with the highest utility is the same irrespective of how utility is scaled. The two models shown below are equivalent for any $\lambda > 0$.

$$\begin{aligned}
U^0_{nj} &= V_{nj} + \epsilon_{nj} \\
U^1_{nj} &= \lambda V_{nj} + \lambda \epsilon_{nj}
\end{aligned} \tag{5}$$

As a result, the analyst has to normalize the scale of utility. Typically, we normalize the scale of utility by normalizing the variance of the error terms. This is what we did when we looked at logit and probit earlier. The scale of utility and the variance of the error terms are directly related. When utility is multiplied by λ, the variance of each ϵ_{nj} changes by λ^2: $\text{var}(\lambda\epsilon_{nj}) = \lambda^2 \text{var}(\epsilon_{nj})$. In other words, normalizing the variance of the error terms is equivalent to normalizing the scale of utility.

Normalization with iid Errors

When the errors are assumed to be iid, then normalizing the scale is straightforward. In effect, you just normalize the error variance to some number that is convenient. Because all the errors have the same variance by assumption, normalizing the variance of any of them sets the variance for all of them.

Whatever we use to normalize the scale affects how we interpret the coefficients. Suppose we have the following model: $U^0_{nj} = x_{nj}\beta + \epsilon^0_{nj}$ where the variance of the error terms is $\text{var}(\epsilon^0_{nj}) = \sigma^2$. Let's normalize the scale by setting the error variance to 1 as in a probit model. The model now becomes the following: $U^1_{nj} = x_{nj}\left(\frac{\beta}{\sigma}\right) + \epsilon^1_{nj}$ with $\text{var}(\epsilon^1_{nj}) = 1$. In other words, the original coefficients are divided by the standard deviation of the unobserved portion of utility. The new coefficients $\frac{\beta}{\sigma}$, therefore, reflect the effect of the observed variables *relative* to the standard deviation of the unobserved factors.

What happens if we set the error variance to $\frac{\pi^2}{6} = 1.64$ as in a standard logit model? Well, now the normalized model would be: $U^1_{nj} = x_{nj}\left(\frac{\beta}{\sigma}\right)\sqrt{1.64} + \epsilon^1_{nj}$ with $\text{var}(\epsilon^1_{nj}) = 1.64$.[4] The difference is that the coefficients will now be larger than in a probit model by a factor of $\sqrt{1.64}$. Probit coefficients can be converted to the scale of the logit coefficients by multiplying them by $\sqrt{1.64}$.[5]

Normalization with Correlated Errors

When errors are correlated over alternatives, normalizing for scale becomes more difficult. In the previous case of normalization with iid errors, we talked about setting the scale of utility. However, as we have already seen, it can be more appropriate to talk about setting the scale of utility differences. It turns out that when errors are correlated over alternatives, normalizing the variance of the error for one alternative is not sufficient to set the scale of utility differences. Train (2007, 31) provides the following four-alternative example. Suppose we have $U_{nj} = V_{nj} + \epsilon_{nj}$ and that the error vector $\epsilon_n = \{\epsilon_{n1}, \ldots, \epsilon_{n4}\}$ has 0 mean and the following covariance matrix:

$$\Omega = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{bmatrix} \tag{6}$$

Because only differences in utility matter, this model is equivalent to one in which all utilities are differenced from, say, the first alternative. Thus, we could write the model from above as: $\tilde{U}_{nj1} = \tilde{V}_{nj1} + \tilde{\epsilon}_{nj1}$ for $j = 2, 3, 4$, where $\tilde{U}_{nj1} = U_{nj} - U_{n1}$, $\tilde{V}_{nj1} = V_{nj} - V_{n1}$, and the vector of error differences is $\tilde{\epsilon}_{n1} = \langle(\epsilon_{n2} - \epsilon_{n1}), (\epsilon_{n3} - \epsilon_{n1}), (\epsilon_{n4} - \epsilon_{n1})\rangle$. Note that the variance between the first and second errors is now:

$$\begin{aligned} \text{var}(\tilde{\epsilon}_{n21}) &= \text{var}(\epsilon_{n2} - \epsilon_{n1}) \\ &= \text{var}(\epsilon_{n1}) + \text{var}(\epsilon_{n2}) - 2\text{cov}(\epsilon_{n1}, \epsilon_{n2}) \\ &= \sigma_{11} + \sigma_{22} - 2\sigma_{12} \end{aligned}$$

We can also calculate the covariance between $\tilde{\epsilon}_{21}$ (the difference between the first and second errors) and

[4] Where did this come from? Essentially, we want to set $\text{var}(\lambda\epsilon^0) = 1.64$. This is the same as saying $\lambda^2\sigma^2 = 1.64$. In this case, $\lambda = \frac{\sqrt{1.64}}{\sigma}$.

[5] Note that in previous notes, I stated that the error variances in a standard logit model are normalized to $\frac{\pi^2}{3} = 3.29$ and that you could convert probit coefficients to the scale of logit coefficients by multiplying them by$\sqrt{3.29} = 1.81$. This is true. It turns out that we are talking about two different logit models with their own parameterization. For our purposes here, it doesn't really matter in that the point is we must normalize the error variances and that this will affect how we interpret the coefficients.

$\tilde{\epsilon}_{31}$ (the difference between the first and third errors). This is:

$$\begin{aligned}
\text{cov}(\tilde{\epsilon}_{n21}, \tilde{\epsilon}_{n31}) &= E(\epsilon_{n2} - \epsilon_{n1})(\epsilon_{n3} - \epsilon_{n1}) \\
&= E(\epsilon_{n2}\epsilon_{n3} - \epsilon_{n2}\epsilon_{n1} - \epsilon_{n3}\epsilon_{n1} + \epsilon_{n1}\epsilon_{n1}) \\
&= \sigma_{23} - \sigma_{21} - \sigma_{31} + \sigma_{11}
\end{aligned}$$

The covariance matrix for the vector of error differences is:

$$\Omega = \begin{bmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{11} + \sigma_{23} - \sigma_{12} - \sigma_{13} & \sigma_{11} + \sigma_{24} - \sigma_{12} - \sigma_{14} \\ \cdot & \sigma_{11} + \sigma_{33} - 2\sigma_{13} & \sigma_{11} + \sigma_{34} - \sigma_{13} - \sigma_{14} \\ \cdot & \cdot & \sigma_{11} + \sigma_{44} - 2\sigma_{14} \end{bmatrix} \tag{7}$$

As you can see from this matrix, setting the variance of one of the original errors does not set the variance for all of the error differences. Imagine setting the variance of the first alternative to $\sigma_{11} = k$. Now the variance of the differences between the errors of the first two alternatives is $k + \sigma_{22} - 2\sigma_{12}$. An infinite number of values for $\sigma_{22} - 2\sigma_{12}$ produce equivalent models.

What the analyst typically does is set the variance of one of the error differences to some number. This sets the scale of the utility differences and therefore utility. Imagine that we set the variance of $\tilde{\epsilon}_{21} = 1$. Thus, the covariance matrix is now:

$$\tilde{\Omega}^* = \begin{bmatrix} 1 & (\sigma_{11} + \sigma_{23} - \sigma_{12} - \sigma_{13})/m & (\sigma_{11} + \sigma_{24} - \sigma_{12} - \sigma_{14})/m \\ \cdot & (\sigma_{11} + \sigma_{33} - 2\sigma_{13})m & (\sigma_{11} + \sigma_{34} - \sigma_{13} - \sigma_{14})m \\ \cdot & \cdot & (\sigma_{11} + \sigma_{44} - 2\sigma_{14})/m \end{bmatrix} \tag{8}$$

where $m = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. Utility is divided by $\sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}$ to obtain this scaling.

Let's return to the difference between iid errors and correlated errors. When you have iid errors, recall, normalizing the variance of one of the errors automatically normalizes the variance of the error differences. With iid errors, $\sigma_{ij} = \sigma_{ii}$ and $\sigma_{ij} = 0$ for $i \neq j$. Thus, when σ_{11} is normalized to k, the variance of the error difference becomes $\sigma_{11} + \sigma_{22} - 2\sigma_{12} = k + k - 0 = 2k$.

An important point to note is that normalization influence the number of parameters that can be estimated. The covariance matrix of the original errors shown in Eq. (6) has 10 elements in the four-alternative example. The covariance matrix of the error differences shown in Eq. (8) with one element normalized has 6 elements, 5 of which are parameters.

$$\tilde{\Omega}^* = \begin{bmatrix} k & \omega_{ab} & \omega_{ac} \\ \cdot & \omega_{bb} & \omega_{bc} \\ \cdot & \cdot & \omega_{cc} \end{bmatrix} \tag{9}$$

In other words, we shift from 10 parameters to 5 parameters. It turns out that a model with J alternatives has at most $J(J-1)/2 - 1$ covariance parameters after normalization.

As I noted before, the normalization of logit and nested logit models are automatic with the distributional assumptions made of the error terms. This is not the case with mixed logit and probit. As a result, you need to keep normalization issues in mind with these models.

So far, we have looked at the basic setup of discrete choice models and this information will come in extremely useful. We now turn to look at specific discrete choice models, staring with logit-based models.

2 Logit-Based Models

2.1 Deriving Logit-Based Models

We can derive logit-based models from the random utility model setup described earlier. Essentially, we start with our basic utility equation:

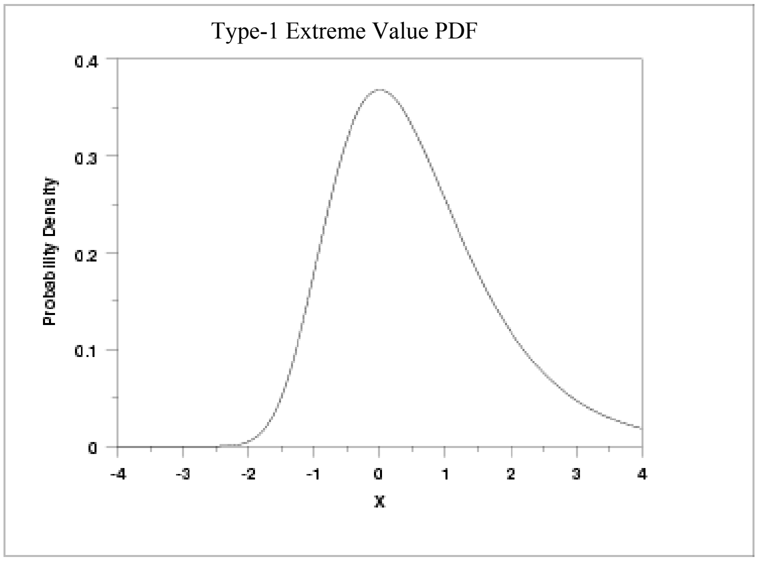
$$U_{nj} = V_{nj} + \epsilon_{nj} \tag{10}$$

The logit model is obtained by assuming that each ϵ_{nj} is distributed iid extreme value. This distribution is also called the Gumbel distribution or a Type-I extreme value distribution. The density for each unobserved component of utility is:

$$f(\epsilon_{nj}) = \text{exp}^{-\epsilon_{nj}} \text{exp}(-\text{exp}^{-\epsilon_{nj}}) \tag{11}$$

This is shown in Figure 1.

Figure 1: Type-1 Extreme Value PDF

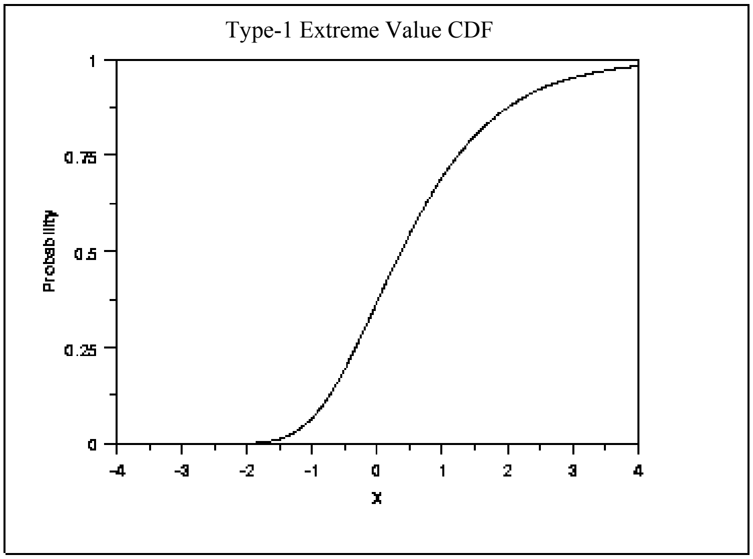


The cumulative distribution is:

$$F(\epsilon_{nj}) = \text{exp}(-\text{exp}^{-\epsilon_{nj}}) \tag{12}$$

This is shown in Figure 2.

Figure 2: Type-1 Extreme Value CDF



The variance of the distribution is $\frac{\pi^2}{6}$. By assuming this, we have normalized the utility scale.

The difference between two extreme value variables is logistic. In other words, if ϵ_{nj} and ϵ_{ni} are iid extreme value, then $\tilde{\epsilon}_{nji} = \epsilon_{nj} - \epsilon_{ni}$ has a logistic distribution i.e.

$$F(\tilde{\epsilon}_{nji}) = \frac{e^{\tilde{\epsilon}_{nji}}}{1 + e^{\tilde{\epsilon}_{nji}}} \tag{13}$$

The key assumption of the iid extreme value distribution is 'independence'. In other words, we assume that the unobserved portion of utility for one alternative is independent of the unobserved portion of utility for other alternatives. This is a somewhat restrictive assumption. How restrictive is it? One way to think about this assumption is that it is the same as assuming you have a well-specified model. Recall that ϵ_{nj} is just the difference between U_{nj} and V_{nj}. The assumption of independence simply means that the error for one alternative provides no information about the error of another alternative. In other words, we are assuming that the model is sufficiently well-specified that whatever remains in the error term is white noise. As Train (2007, 40) points out, the goal of the analyst is to specify utility well enough that the logit model is appropriate.

Anyway, once we assume that the error terms are distributed iid extreme value, we can then derive the logit choice probabilities. Recall from earlier that the probability that decision maker n chooses alternative i is just:

$$\begin{aligned} P_{ni} &= \text{Prob}(U_{ni} > U_{nj}) \\ &= \text{Prob}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}) \\ &= \text{Prob}(\epsilon_{nj} < \epsilon_{ni} + V_{ni} - V_{nj}) \end{aligned} \tag{14}$$

If ϵ_{ni} is taken as given, Eq. (14) is the cumulative probability distribution for each ϵ_{nj} evaluated at $\epsilon_{ni} + V_{ni} - V_{nj}$. According to Eq. (12), this is: $\exp(-\exp(-(\epsilon_{ni} + V_{ni} - V_{nj})))$. Due to independence, this cumulative distribution over all $j \neq i$ is the product of the individual cumulative distributions:

$$P_{ni}|\epsilon_{ni} = \prod_{j \neq i} e^{-e^{-(\epsilon_{ni}+V_{ni}-V_{nj})}} \tag{15}$$

Because ϵ_{ni} is not really given, the choice probability is the integral of $P_{ni}|\epsilon_{ni}$ over all values of ϵ_{ni} weighted by the density of ϵ_{ni}.

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\epsilon_{ni}+V_{ni}-V_{nj})}} \right) e^{-\epsilon_{ni}} e^{-e^{-\epsilon_{ni}}} d\epsilon_{ni} \tag{16}$$

It turns out that the solution to this integral is:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{17}$$

This is the equation for the logit choice probability. Typically, we model the systematic component of utility as a linear function of parameters: $V_{nj} = x_{nj}\beta$ and so the logit choice probability becomes:[6]

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \tag{19}$$

2.1.1 Scale Parameter

Let's just go back to the scale parameter for a moment. Recall that in the logit model, we assume that the unobserved component of utility is distributed iid extreme value with variance $\pi^2/6$. Think about the

[6] Note that I have derived the logit choice probability from a random utility model setup. It is possible to derive it from a basic probability framework too. Let y be a dependent variable with J nominal outcomes. Let $P(y_n = i)$ be the probability of observing outcome i in observation n.

1. Assume that $P(y_n = i)$ is a function of the linear combination, $x_{ni}\beta$.
2. To ensure non-negative probabilities, we use the exponential of $x_{ni}\beta$.
3. To make the probabilities sum to 1, we divide $e^{x_{ni}\beta}$ by $\sum_j e^{x_{ni}\beta}$ to get

$$P(y_n = i) = P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{ni}\beta}} \tag{18}$$

This is exactly the same as Eq. 22.

following model where utility is $U^*_{nj} = V_{nj} + \epsilon^*_{nj}$, where the unobserved portion has variance $\sigma^2 \times (\pi^2/6)$. Because the scale of utility is irrelevant to the choice of the decision maker, we can divide by σ without changing choices. Thus, we now have $U_{nj} = V_{nj}/\sigma + \epsilon_{nj}$, where $\epsilon_{nj} = \epsilon^*_{nj}/\sigma$. In this model, the variance of the unobserved portion is $\pi^2/6$ as in our logit setup.[7] The choice probability, though, is now:

$$P_{ni} = \frac{e^{V_{ni}/\sigma}}{\sum_j e^{V_{nj}/\sigma}} \tag{20}$$

This can be written as:

$$P(y_n = i) = P_{ni} = \frac{e^{x_{ni}(\beta^*/\sigma)}}{\sum_j e^{x_{nj}(\beta^*/\sigma)}} \tag{21}$$

The parameter σ is called the scale parameter because it scales the coefficients to reflect the variance of the unobserved portion of utility. As we saw earlier, only the ratio β^*/σ can be estimated because β^* and σ are not separately identified. In most textbooks, you will find the model expressed in its scaled form, with $\beta = \beta^*/\sigma$, and the standard logit choice probability:

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \tag{22}$$

The point to showing you all of this is to make you recognize that the coefficients you get indicate the effect of a variable relative to the variance of the unobserved factors. The larger the variance in unobserved factors, the smaller the coefficients (even if the observed factors have the same effect on utility). This is worth remembering when you evaluate the same model on two different data sets. If the variance of the unobserved factors is different across the two data sets (say two different countries or time periods), you will find that the coefficients differ across the two models. However, you should be careful not to infer from the coefficients that the effect of the variables on utility is different across the two data sets.

2.1.2 Estimation

Estimation of this model is relatively easy since the log likelihood function is globally concave. To specify the likelihood, first define $d_{ni} = 1$ if individual n chooses alternative i, $d_{ni} = 0$ otherwise. This means that there are J lots of d_{ni}, each indicating a choice. We can then use these indicators to select the appropriate terms in the likelihood function. Thus, the likelihood function for individual n is:

$$\mathcal{L}_n = P_{n1}^{d_{n1}} \times P_{n2}^{d_{n2}} \times P_{n3}^{d_{n3}} \times \ldots \times P_{nJ}^{d_{nJ}} \tag{23}$$

where P_{ni} is the probability that individual n chooses alternative i. The likelihood function for the entire sample is:

$$\mathcal{L} = \prod_{n=1}^{N} \left(P_{n1}^{d_{n1}} \times P_{n2}^{d_{n2}} \times P_{n3}^{d_{n3}} \times \ldots \times P_{nJ}^{d_{nJ}} \right) \tag{24}$$

[7] var(ϵ_{nj}) = var$(\epsilon^*_{nj}/\sigma) = (1/\sigma^2)var(\epsilon^*_{nj}) = (1/\sigma^2) \times \sigma^2 \times (\pi^2/6) = \pi^2/6$.

Thus, the log-likelihood function is just:[8]

$$\begin{aligned} \text{ln}\mathcal{L} &= \sum_{n=1}^{N}\sum_{i=1}^{J} d_{ni}\text{ln}(P_{ni}) \\ &= \sum_{n=1}^{N}\sum_{i=1}^{J} d_{ni}\text{ln}\left(\frac{e^{x_{ni}\beta}}{\sum_j e^{x_{ni}\beta}}\right) \end{aligned} \tag{25}$$

3 Pure Conditional Logit – Alternative-Specific Data

It is time for some terminology regarding different logit-based models. As we will see, there are slightly different types of logit models and it is worth thinking about what each of them mean. The model shown above is commonly referred to as a pure conditional logit model given how its subscripts are listed. Recall the basic setup.

$$\begin{aligned} U_{nj} &= V_{nj} + \epsilon_{nj} \\ &= x_{nj}\beta + \epsilon_{nj} \end{aligned} \tag{26}$$

Let's take a closer look at Eq. (26). First, x_{nj} is a matrix of the characteristics of the j^{th} alternative relative to individual n - they are choice- or alternative-specific characteristics. In other words, the independent variables don't provide information about the decision maker; rather they provide information about the decision maker *relative* to each alternative. A common example is the left-right ideological distance from individual n to some party. This means that there will be as many left-right ideological distance observations for each individual n as there are political parties. Second, β is a vector of parameters relating the relationship between individual n and the alternative x_{nj} to the individual's utility for the alternative. In other words, the parameters are not alternative specific i.e. they are not subscripted by j. This means that there will be only one coefficient on an independent variable like the left-right ideological distance between an individual and a party for *ALL* the possible alternatives (parties). To see all of this more clearly, let's imagine that our systematic component might be something like the following:[9]

$$V_{nj} = \beta_1\text{IssueDistance}_{nj} \tag{27}$$

[8] It turns out that you can also estimate this model on a subset of the alternatives. This is useful when the number of alternatives facing each decision maker is so large that estimating the model parameters is difficult. Consider a choice situation in which there are 100 alternatives. You can estimate the logit model on a subset of 10 alternatives for each sample decision maker with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99 alternatives. If all alternatives have the same chance of being selected into the subset, then this is the same as estimating the model on the full subset. The only difference is that this estimation procedure is not as efficient, which you would expect given that you have thrown away information. For more information on this, see Train (2007, 68-70).

[9] Note that there is no constant - I'll explain why in a moment.

If we had three choices, then our systematic components would be:

$$\begin{aligned} V_{n1} &= \beta_1 \text{IssueDistance}_{n1} \\ V_{n2} &= \beta_1 \text{IssueDistance}_{n2} \\ V_{n3} &= \beta_1 \text{IssueDistance}_{n3} \end{aligned} \tag{28}$$

Now substituting the systematic components into our utility equation, we have:

$$\begin{aligned} U_{n1} &= \beta_1 \text{IssueDistance}_{n1} + \epsilon_{n1} \\ U_{n2} &= \beta_1 \text{IssueDistance}_{n2} + \epsilon_{n2} \\ U_{n3} &= \beta_1 \text{IssueDistance}_{n3} + \epsilon_{n3} \end{aligned} \tag{29}$$

As you can see, the effect of each independent variable (ISSUEDISTANCE) is constant (β_1) across all three alternatives. However, the value of the independent variable varies across the alternatives.

An example of what the data will look like is shown in Table 1. Imagine that there are four individuals choosing among three alternatives. Imagine also that there is just one independent variable x_{nj}. n is the respondent's ID or individual n's ID.

Table 1: Data in a Pure Conditional Logit Model

n	Outcome j	Outcome chosen	Variable x_{nj}
1	1	0	x_{11}=1
1	2	1	x_{12}=0
1	3	0	x_{13}=7
2	1	1	x_{21}=2
2	2	0	x_{22}=1
2	3	0	x_{23}=6
3	1	1	x_{31}=1
3	2	0	x_{32}=6
3	3	0	x_{33}=9
4	1	0	x_{41}=1
4	2	0	x_{42}=3
4	3	1	x_{43}=2

As we have seen, the probability that individual n chooses alternative i in the pure conditional logit model is:

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \tag{30}$$

Note that if there were some variable (z_n) that did not vary over the choices for the individuals, then that variable would just cancel out of Eq. (32). In other words, STATA would drop it. For example, if you had a variable that measured the age of a voter, then this would not vary over the choice of parties that he gets to vote for. As a result, it would be dropped. Because the constant is essentially a variable that does not vary over the alternatives, it too is dropped. This is why you do not get a constant from a pure conditional logit model.[10]

3.1 Estimation

To estimate the pure conditional logit model in STATA you type:

```
clogit Y X, group(respid);
```

where RESPID is the name of the respondent's ID variable.

3.2 Interpretation

The results from a CL model of party choice in the Netherlands are shown in Table 2. The model examines how the respondent's probability of voting for a party is affected by the distance between the respondent and the party on the abortion, nuclear, income, and left-right issue dimensions.

Table 2: The Determinants of Party Choice in the Netherlands

Regressor	Pure CL Model
AbortionDistance	-0.24***
	(0.04)
NuclearDistance	-0.10***
	(0.04)
IncomeDifferenceDistance	-0.44***
	(0.04)
RightDistance	-0.66***
	(0.04)
Constant	
Log likelihood	-692.79
Observations	3461

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)
Standard errors are given in parentheses

3.2.1 Interpreting Coefficients

The sign of the coefficients indicate how an increased distance on a particular issue dimension between the respondent and a party affects the likelihood that the respondent will vote for that party. Thus, the fact that

[10] You probably should have a constant, but I'll show you how to put one in a conditional logit model in a moment.

all the coefficients are negative and significant means that we can be quite confident that the respondent will be less likely to vote for a party as the party moves away from the respondent on each issue dimension. In effect, this is quite strong support for a spatial model of voter choice.

3.2.2 Odds Ratios

It is possible to give the conditional logit model an odds interpretation. The odds of outcome a versus outcome b is:

$$\text{ODDS}_{ab} = \frac{P_{na}}{P_{nb}} = \frac{\frac{e^{x_{na}\beta}}{\sum_j e^{x_{nj}\beta}}}{\frac{e^{x_{nb}\beta}}{\sum_j e^{x_{nj}\beta}}} = \frac{e^{x_{na}\beta}}{e^{x_{nb}\beta}} = e^{[x_{na}-x_{nb}]\beta} \tag{31}$$

where the odds change according to the difference in the value of the x's associated with the two outcomes a and b. So say the respondent is choosing between two choices a and b. On RIGHT, INCOMEDIFFERENCE, and NUCLEAR, the two choices are the same distance from the respondent such that $x_{na} - x_{nb} = 0$. However, party a is 2 units closer on the abortion issue than party b. The odds that the respondent chooses party a over party b is $e^{-2(-0.24)}$=1.62. In other words, the respondent is 1.62 times more likely to vote for party a (the closer one on the abortion issue) than the identical party b that is 2 units further away on the abortion issue.

You can obtain exponentiated coefficients from the conditional logit model by typing:

```
clogit Y X, group(respid) or;
```

These exponentiated coefficients give you the effect of a unit change in the independent variables on the odds of any given alternative. Another useful way to obtain both the coefficients and exponentiated coefficients if you have the SPOST package is to type:

```
clogit Y X, group(respid) or;
listcoef, help;
```

3.2.3 Predicted Probabilities and First Differences

You can also calculate predicted probabilities and first differences using the following equation that we noted earlier.[11]

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \tag{32}$$

[11] As before, these predicted probabilities can be found using CLARIFY, Prvalue, or your own manual calculations.

4 Multinomial Logit Model – Case-Specific Data

As I have noted earlier, there are other logit-based models. One such model is known as the multinomial logit model. The basic utility equation for individual n choosing alternative j an MNL model is shown below.

$$U_{nj} = V_{nj} + \epsilon_{nj} \tag{33}$$

The systematic component of the utility function is given as:

$$V_{nj} = z_n \gamma_j \tag{34}$$

So, we have

$$U_{nj} = z_n \gamma_j + \epsilon_{nj} \tag{35}$$

It is important to look at Eq. (35) carefully to see how it differs from the conditional logit model shown earlier.[12] First, γ_j is a vector of alternative-specific parameters i.e. the parameters are subscripted by j. These parameters relate the characteristics of a respondent (z) to the respondent's utility for the j^{th} choice – they are individual-specific characteristics. This means that the effect of the independent variables will vary across all of the choices. In other words, there will be a separate coefficient on each independent variable for each possible outcome. For example, if the age of the individual was an independent variable, then the effect of age on choosing alternative 1 would be different to its effect on choosing alternative 2, alternative 3 etc. Second, z_n is a matrix of individual or case-specific characteristics. Note that z_n is just subscripted by n. In other words, these individual characteristics have nothing to do with the alternatives that are available. To see all of this more clearly, let's imagine that our systematic component might be something like the following:

$$V_{nj} = \gamma_{j0} + \gamma_{j1}\text{Age}_n + \gamma_{j2}\text{Education}_n + \gamma_{j3}\text{Male}_n \tag{36}$$

If we had three choices, then our systematic components would be:

$$\begin{aligned}
V_{n1} &= \gamma_{10} + \gamma_{11}\text{Age}_n + \gamma_{12}\text{Education}_n + \gamma_{13}\text{Male}_n \\
V_{i2} &= \gamma_{20} + \gamma_{21}\text{Age}_n + \gamma_{22}\text{Education}_n + \gamma_{23}\text{Male}_n \\
V_{i3} &= \gamma_{30} + \gamma_{31}\text{Age}_n + \gamma_{32}\text{Education}_n + \gamma_{33}\text{Male}_n
\end{aligned} \tag{37}$$

Now substituting the systematic components into our utility equation, we have:

$$\begin{aligned}
U_{n1} &= \gamma_{10} + \gamma_{11}\text{Age}_n + \gamma_{12}\text{Education}_n + \gamma_{13}\text{Male}_n + \epsilon_{n1} \\
U_{n2} &= \gamma_{20} + \gamma_{21}\text{Age}_n + \gamma_{22}\text{Education}_n + \gamma_{23}\text{Male}_n + \epsilon_{n2} \\
U_{n3} &= \gamma_{30} + \gamma_{31}\text{Age}_n + \gamma_{32}\text{Education}_n + \gamma_{33}\text{Male}_n + \epsilon_{n3}
\end{aligned} \tag{38}$$

As you can see, the effect of each individual characteristic (Age, Education, Male) varies across each of the three alternatives i.e. γ_{jk} is subscripted by the choice (j) and the independent variable (k).

An example of what the data will look like is shown in Table 3 below. Imagine that there are four individuals choosing among three alternatives. Individual 1 chooses alternative 2, individuals 2 and 3 choose alternative 1, and individual 4 chooses alternative 3.

[12] I use z and γ to distinguish the MNL model from the pure CL model from earlier.

Table 3: Data in an MNL Model

n	Outcome j chosen	Variable z_n
1	2	z_1=1
2	1	z_2=4
3	1	z_3=2
4	3	z_4=1

The probability that individual n chooses alternative i in the multinomial logit model is:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

$$= \frac{e^{z_n \gamma_i}}{\sum_j e^{z_n \gamma_j}} \tag{39}$$

Compare this to the probability from earlier that individual n chooses alternative i in the pure conditional logit model:

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}} \tag{40}$$

4.1 Identification – Sociodemographic Variables

Unlike with the pure conditional logit model, there is a problem as things stand – the alternative-specific coefficients $\hat{\gamma}_j$ are underidentified. This is a point we made very early on when we discussed sociodemographic variables i.e. characteristics of the decision maker. By definition, attributes of the decision maker do not vary over the alternatives. As a result, they can only enter the model in ways that create differences in utility over the alternatives. To restate the problem slightly differently, for any vector of constants q, we can get the exact same probabilities whether we use γ_j or γ^*, where $\gamma^* = \gamma_j + q$. In other words, we could add an arbitrary constant to all the coefficients in the model and we would get exactly the same probabilities. Consider the following example where we have three alternatives:

$$P_{n1} = \frac{e^{z_n \gamma_1}}{\sum_{j=1}^{3} e^{z_n \gamma_j}} \tag{41}$$

Now add a vector of constants q so that we have:

$$P_{n1} = \frac{e^{z_n(\gamma_1 + q)}}{\sum_{j=1}^{3} e^{z_n(\gamma_j + q)}} \tag{42}$$

With a little manipulation, this can be rewritten as:

$$
\begin{aligned}
P_{n1} &= \frac{e^{z_n\gamma_1}e^{z_n q}}{e^{z_n(\gamma_1+q)}+e^{z_n(\gamma_2+q)}+e^{z_n(\gamma_3+q)}} \\
&= \frac{e^{z_n\gamma_1}e^{z_n q}}{e^{z_n\gamma_1}e^{z_n q}+e^{z_n\gamma_2}e^{z_n q}+e^{z_n\gamma_3}e^{z_n q}} \\
&= \frac{e^{z_n\gamma_1}e^{z_n q}}{(\sum_{j=1}^{3}e^{z_n\gamma_j})e^{z_n q}} \\
&= \frac{e^{z_n\gamma_1}}{\sum_{j=1}^{3}e^{z_n\gamma_j}} \qquad (43)
\end{aligned}
$$

which is what we started with. Therefore, the model, as written, is underidentified.

As we noted earlier when discussing sociodemographic variables in general, a convenient normalization that solves the identification problem is to assume that one of the sets of coefficients (the coefficients for one of the choices) are all zero. As we'll see, this then becomes the reference category against which all of the results are compared. Say that we set $\gamma_1 = 0$. In other words, say we set all of the coefficients for alternative 1 to be 0. By doing this, we are assuming that the first alternative is the base category against which all others are compared.[13] Once this constraint is added, we now have:

$$P_{ni} = \frac{e^{z_n\gamma_i}}{\sum_{j=1}^{J}e^{z_n\gamma_j}} \text{ where } \gamma_1 = 0 \qquad (44)$$

Since $e^{z_n\gamma_1} = e^{z_n 0} = 1$, we can rewrite Eq. (44) as two separate equations:

$$P_{n1} = \frac{e^{z_n 0}}{e^{z_n 0}+\sum_{j=2}^{J}e^{z_n\gamma_j}} = \frac{1}{1+\sum_{j=2}^{J}e^{z_n\gamma_j}} \qquad (45)$$

and

$$P_{ni} = \frac{e^{z_n\gamma_i}}{1+\sum_{j=2}^{J}e^{z_n\gamma_j}} \text{ for } i > 1 \qquad (46)$$

[13] Note that we can arbitrarily choose any alternative as our base category. You should be very careful in that different statistical programs automatically choose different categories - some choose the lowest category, some the highest. STATA's default is to select the alternative that is chosen most often as the baseline category.

Note that this is the formulation used in Alvarez and Nagler (1998).[14] As we'll see, the way to think about the use of a base category is that the logic when it comes to interpretation is exactly the same as when you break a $K+1$ category variable into K dummy variables.

4.2 Estimation

To estimate the multinomial logit model in STATA you type:

```
mlogit Y X, base(some number);
```

where SOME NUMBER is the alternative that you want as the baseline category.

4.3 Interpretation

Here are some results from a model of party choice in the Netherlands where

$$\begin{aligned} \text{PartyChoice}_{nj} = \beta_{j0} &+ \beta_{j1}\text{Abortion}_n + \beta_{j2}\text{Nuclear}_n + \beta_{j3}\text{IncomeDifference}_n \\ &+ \beta_{j4}\text{Right}_n + \beta_{j5}\text{Male}_n + \beta_{j6}\text{Religious}_n + \beta_{j7}\text{Education}_n + \epsilon_{nj} \end{aligned} \tag{48}$$

where PARTYCHOICE has four categories (Pvda, CDA, VVD, and D66). ABORTION, NUCLEAR, INCOMEDIFFERENCE and RIGHT measure the respondents attitudes towards the various issues on a seven point scale, MALE is a dummy variable indicating gender of respondent, RELIGIOUS is a dummy variable indicating whether the respondent is religious or not, and EDUCATION indicates the respondent's level of education on a five point scale. Pvda is the reference category. Results are shown in Table 4.

[14] Note that you will see slight variants on this. For example, Charles Franklin's notes have

$$\begin{aligned} P_{n0} &= \frac{1}{1+\sum_{j=1}^{J} e^{z_n\gamma_j}} \\ P_{ni} &= \frac{e^{z_n\gamma_i}}{1+\sum_{j=1}^{J} e^{z_n\gamma_j}} \quad \text{for } i>0 \end{aligned} \tag{47}$$

The reason for the difference is that Franklin is assuming that the first alternative is actually coded 0 instead of 1. In other words, $i = 0, 1, \ldots, J$, so that there are J+1 alternatives rather than the J alternatives we have been assuming in our notes. This changes what we sum over in the denominator of our probability equations. Both formulations are obviously equally good.

Table 4: The Determinants of Party Choice in the Netherlands

	PVDA is the Reference Party		
Regressor	CDA	VVD	D66
Abortion	-0.13***	0.21***	0.16**
	(0.05)	(0.07)	(0.07)
Nuclear	-0.13***	-0.21***	-0.06
	(0.05)	(0.06)	(0.06)
IncomeDifference	-0.31***	-0.75***	-0.32***
	(0.06)	(0.08)	(0.07)
Right	1.04***	1.19***	0.47***
	(0.08)	(0.10)	(0.09)
Male	-0.60***	-0.85***	-0.57***
	(0.19)	(0.24)	(0.22)
Religious	1.59***	0.03	0.09
	(0.22)	(0.25)	(0.23)
Education	0.45***	0.77***	0.65***
	(0.09)	(0.12)	(0.11)
Constant	-3.18***	-4.04***	-3.14***
	(0.65)	(0.84)	(0.78)
Log likelihood	-1031.92	-1031.92	-1031.92
Observations	1172	1172	1172

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed).
Standard errors are given in parentheses

4.3.1 Interpreting Coefficients

As you can see, there are three sets of coefficients for each independent variable that correspond to the three different parties that are not the base category. You can interpret the signs of the coefficients in the expected manner. For example, a negative (positive) coefficient indicates that the independent variable reduces (increases) the probability of voting for a particular party compared to the baseline party. For example, it turns out that being a male reduces the probability that the individual will vote for the CDA, VVD, or the D66 when compared to the PVDA. Being religious makes it more likely that you will vote for the CDA compared to the PVDA but does not make it more or less likely that you will vote for either the VVD or the D66 compared to the PVDA.

We can also interpret the alternative-specific constants. In some sense, if our model (utility function) was complete and we had captured all the relevant components of vote choice, then the constant should be zero. Thus, a non-zero constant indicates that the respondents have some inherent propensity to vote for one party over another for reasons that are not captured in the model. The alternative-specific constants capture the average effect on utility of all factors that are not included in the model.

4.3.2 Odds Ratios

It is possible to give the multinomial logit model an odds interpretation. The odds of outcome a versus outcome b, where neither a nor b are the baseline category, is:

$$\text{ODDS}_{ab} = \frac{P_{na}}{P_{nb}} = \frac{\frac{e^{z_n\gamma_a}}{\sum_j e^{z_n\gamma_j}}}{\frac{e^{z_n\gamma_b}}{\sum_j e^{z_n\gamma_j}}} = \frac{e^{z_n\gamma_a}}{e^{z_n\gamma_b}} = e^{z_n[\gamma_a-\gamma_b]} \tag{49}$$

If you wanted to look at the odds ratio of voting for any of the parties as opposed to the baseline category (1), then this simplifies to:

$$\text{ODDS}_{a1} = e^{z_n[\gamma_a-\gamma_1]} = e^{z_n\gamma_a} \tag{50}$$

It turns out that the odds of an individual who is a 2 on ABORTION, NUCLEAR, INCOMEDIFFERENCE, and RIGHT, who is MALE, who is RELIGIOUS, and who is a 5 on EDUCATION (5) voting for the CDA rather than the Pvda (the baseline category) is:

$$\begin{aligned}\text{ODDS}_{\text{(CDA-Pvda)}} &= e^{-0.318+2\times(-0.13)+2\times(-0.13)+2\times(-0.31)+2\times(1.04)+1\times(-0.60)+1\times(1.59)+5\times(0.45)} \\ &= 2.92\end{aligned} \tag{51}$$

Thus, an individual with these characteristics is 2.92 times more likely to vote for the CDA than the Pvda.[15]

You can also think about how a change in a particular variable affects the odds of voting for one of the parties rather than the reference category. I already said that the odds of voting for party a rather than the reference category (j=1) is:

$$\text{ODDS}_{a1}(z_n) = e^{z_n\gamma_a} \tag{52}$$

It is relatively easy to show that

$$\frac{\text{ODDS}_{a1}(z_n, z_{nk}+\delta)}{\text{ODDS}_{a1}(z_n, z_{nk})} = e^{\gamma_{ka}\times\delta} \tag{53}$$

where z_{nk} is the k^{th} independent variable for individual n and γ_{ka} is the coefficient on the k^{th} independent variable for alternative a. Thus, for an increase of one unit in a respondents measure on ABORTION, the odds of voting for the VVD rather than the Pvda are expected to change by a factor of $e^{0.21\times1} = 1.24$, holding all other variables constant. In other words, an increase of one unit in an individual's ABORTION score makes them 1.24 times more likely to vote for the VVD rather than the Pvda or, equivalently, increases their likelihood of voting for the VVD rather than the Pvda by 24%. Thus, in order to know what a one unit effect of a given variable is on the odds of voting for one party over the reference category, all you have to do is exponentiate the coefficient.

It is easy to get STATA to give you this directly by typing:

```
mlogit party abortion nuclear incomedifference right male
                religious education, base(some number) rrr
```

[15] Imagine that you wanted to calculate the odds that this individual voted for the VVD rather than the CDA. Because it is easier to calculate Eq. 50 rather than Eq. 49, the easiest thing to do is reestimate the model but now make the CDA the baseline category.

4.3.3 Predicted Probabilities and First Differences

Predicted probabilities can be computed with the following equation.[16]

$$P_{ni} = \frac{e^{z_n \gamma_i}}{1 + \sum_{j=2}^{J} e^{z_n \gamma_j}} \text{ for } i > 1 \tag{55}$$

The predicted probability that an individual votes for the VVD if they have the same characteristics as before is 0.15 [0.07, 0.28]. 95% confidence intervals are shown in parentheses. The predicted probability that an individual votes for the VVD if they have exactly the same characteristics as before but their education is a 1 instead of a 5 (i.e. low education) is 0.03 [0.01, 0.06]. The effect of reducing education from a 5 to a 1 is to reduce the probability that the individual votes for the VVD by 0.13 [0.05, 0.24].[17]

Another nice way to present predicted probabilities is with the help of a graph. In Figure 3, I plot the predicted probability of voting for each of the parties across all levels of the education variable when other variables are set to particular values: abortion=2, nuclear=2, incomedifference=2, right=2, male=1, and religious=1. At any level of education, the four probabilities obviously sum to one. To get this graph, you would use the SPOST package and type:

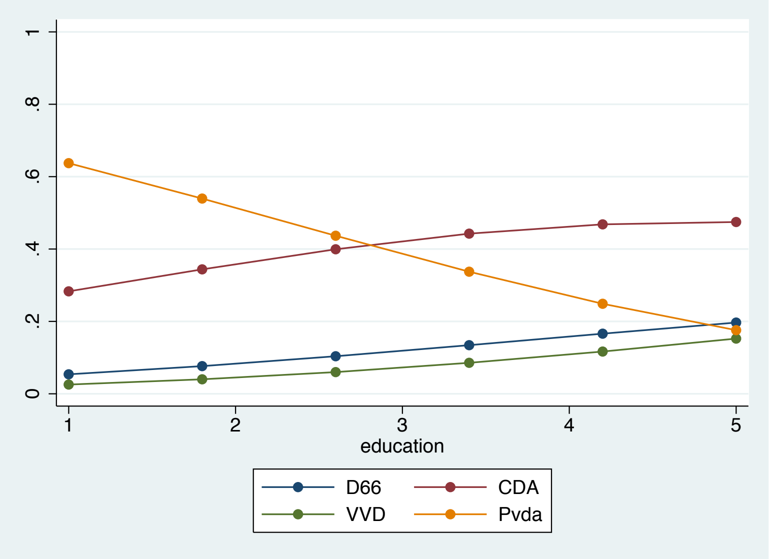
```
mlogit party abortion nuclear male religious education, base(0);
prgen education, x(abortion=2 nuclear=2 incomedifference=2
    right=2 male=1 religious=1) from(1) to (5) generate(p) ncases(6);
desc p*;
label var pp0 "Pvda";
label var pp2 "VVD";
label var pp1 "CDA";
label var pp3 "D66";
graph twoway connected pp3 pp1 pp2 pp0 px,
    yscale( range(0 1)) ylabel(0(0.2)1);
```

[16] As we saw earlier, the predicted probability for the baseline category ($j = 1$) is

$$P_{n1} = \frac{1}{1 + \sum_{j=2}^{J} e^{z_n \gamma_j}} \tag{54}$$

[17] As before, these predicted probabilities can be found using Clarify, Prvalue, or your manual calculation in STATA.

Figure 3: Predicted Probability Line Plot

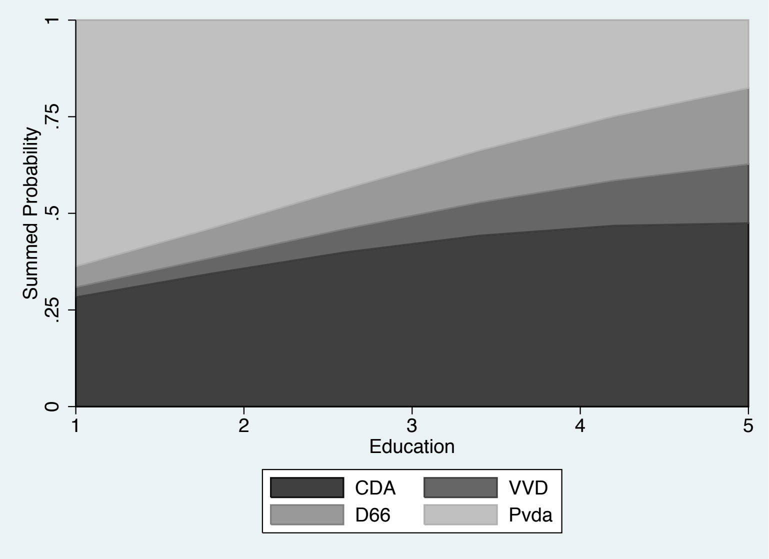


In Figure 4, I plot the same information in a slightly different way. In effect, I plot a summed predicted probability area graph across all levels of the education variable when other variables are set to the same particular values as before. To get this graph, you would use the SPOST package and type:

```
mlogit party abortion nuclear male religious education, base(0);
prgen education, x(abortion=2 nuclear=2 incomedifference=2
     right=2 male=1 religious=1) from(1) to (5) generate(p) ncases(6);
desc p*;
label var ps1 "CDA";
label var ps2 "VVD+CDA";
label var ps3 "D66+VVD+CDA";
label var ps0 "Pvda+rest";
gen zero=0;
gen one=1;
graph twoway (rarea ps1 zero px, bc(gs1))
            (rarea ps2 ps1 px, bc(gs4))
            (rarea ps3 ps2 px, bc(gs8))
            (rarea one ps3 px, bc(gs11)),
            ytitle("Summed Probability")
            legend(order(1 2 3 4)
                label(1 "CDA")
                label(2 "VVD")
                label(3 "D66")
                label(4 "Pvda"))
            xtitle("Education") xlabel(1 2 3 4 5) ylabel(0(0.25)1)
```

```
plotregion(margin(zero));
```

Figure 4: Summed Predicted Probability Area Plot



4.3.4 Marginal Effects

You can also calculate the marginal effect of an independent variable on the probability of some outcome category.

$$\begin{aligned}
\frac{\partial P_{ni}}{\partial z_{nk}} &= \frac{\partial\left(\frac{e^{z_n\gamma_i}}{\sum_{j=1}^{J} e^{z_n\gamma_j}}\right)}{\partial z_{nk}} \\
&= \frac{\beta_{ki} e^{z_n\gamma_i} \sum_{j=1}^{J} e^{z_n\gamma_j} - \sum_{j=1}^{J}[\gamma_{jk} e^{z_n\gamma_j}] e^{z_n\gamma_i}}{[\sum_{j=1}^{J} e^{z_n\gamma_j}]^2} \\
&= \gamma_{ki} \times \frac{e^{z_n\gamma_i}}{\sum_{j=1}^{J} e^{z_n\gamma_j}} - \frac{e^{z_n\gamma_i}}{\sum_{j=1}^{J} e^{z_n\gamma_j}} \times \frac{e^{z_n\gamma_j}}{\sum_{j=1}^{J} e^{z_n\gamma_j}} \times \sum_{j=1}^{J} \gamma_{jk} \\
&= P_{ni}\left[\gamma_{ki} - \sum_{j=1}^{J} \beta_{jk} P_{nj}\right] \qquad (56)
\end{aligned}$$

As always, the marginal effect is the slope of the curve relating z_{nk} to P_{ni}, holding all other variables constant. Using the following mfx command to save writing the code:

```
mfx compute,
    at(abortion=2 nuclear=2 incomedifference=2 right=2
    education=5 male=1 religious=1) predict(outcome(2))
```

we find that the marginal effect of ABORTION on the probability that an individual chooses the VVD is 0.03 [0.01, 0.05]. Note that the marginal effect depends on the values of all the other γ_{jk}s and so depends on the levels of all the other variables. As the value of z_{nk} changes, so can the sign of the marginal effect. For example, the marginal effect of some z may be positive at some point, but negative at another. Thus, you cannot infer anything about the sign of the marginal effect from the sign of the coefficient from the model just estimated.

4.4 Comparing Logit and Multinomial Logit

The multinomial logit model can be thought of as the same as simultaneously estimating binary logit models for all possible comparisons among the outcome variables. Consider the following situation with three alternatives:

$$\begin{aligned} U_{n1} &= \gamma_1 z_n + \epsilon_{n1} \\ U_{n2} &= \gamma_2 z_n + \epsilon_{n2} \\ U_{n3} &= \gamma_3 z_n + \epsilon_{n3} \end{aligned} \tag{57}$$

where U_{nj} represents the utility of the n^{th} individual for the j^{th} alternative. If the IIA assumption holds (which is assumed by both the logit and multinomial logit models), then binary logit will produce consistent estimates of the parameters because the maintained model implies that the presence of a third choice has no impact on the relative probabilities of choosing among the remaining two choices (we will see this a little later when we discuss the IIA assumption in more detail). In other words, a binary logit will generate consistent estimates of γ_2 using only individuals who choose 2 or 3. A binary logit will also generate consistent estimates of γ_1 using only individuals who choose 1 or 3. The only difference between the binary and multinomial logit models is that there is a loss of efficiency with the binary logit. This is because some information is discarded by dropping those individuals who voted for the third choice i.e. sample size is smaller. In other words, when we compare those who choose between 1 and 2, we drop those who chose 3. The multinomial logit model does not do this and produces more efficient estimates. The important point to take away is that the binary and multinomial logit models are estimating the exact same parameters; the multinomial logit model is no richer than the binary logit model (Alvarez & Nagler 1998).

Let's think about this some more. Say we had a choice between three outcomes (Bush, Clinton, Perot) and one independent variable z. We could examine the effect of z on the choice using three binary logit models.

$$\begin{aligned} \ln\left[\frac{P(\text{Clinton}|z)}{P(\text{Bush}|z)}\right] &= \gamma_{0,\text{Clinton|Bush}} + \gamma_{1,\text{Clinton|Bush}} z \\ \ln\left[\frac{P(\text{Perot}|x)}{P(\text{Bush}|z)}\right] &= \gamma_{0,\text{Perot|Bush}} + \gamma_{1,\text{Perot|Bush}} z \\ \ln\left[\frac{P(\text{Clinton}|z)}{P(\text{Perot}|z)}\right] &= \gamma_{0,\text{Clinton|Perot}} + \gamma_{1,\text{Clinton|Perot}} z \end{aligned}$$

where the subscripts to the γs indicate which comparison is being made. The three binary logit models include redundant information. Since $\ln\frac{a}{b} = \ln a - \ln b$, then it follows that:

$$\ln\left[\frac{P(\text{Clinton}|z)}{P(\text{Bush}|z)}\right] - \ln\left[\frac{P(\text{Perot}|z)}{P(\text{Bush}|z)}\right] = \ln\left[\frac{P(\text{Clinton}|z)}{P(\text{Perot}|z)}\right] \tag{58}$$

This implies that:

$$\begin{aligned} \gamma_{0,\text{Clinton|Bush}} - \gamma_{0,\text{Perot|Bush}} &= \gamma_{0,Clinton|Perot} \\ \gamma_{1,\text{Clinton|Bush}} - \gamma_{1,\text{Perot|Bush}} &= \gamma_{1,\text{Clinton|Perot}} \end{aligned} \tag{59}$$

The point here is that we only have to estimate two logit models to know what the results of the third logit will be.

4.5 STATA's MPROBIT Command

STATA's MPROBIT command is the normal error counterpart to the MLOGIT command in the same way that the PROBIT command is the normal counterpart to the LOGIT command. You have to be careful here, though, because although MPROBIT estimates a multinomial probit model, this is **NOT** the multinomial probit model that is typically referred to in the literature. The model fit by mprobit assumes that the errors are normal. With normal errors, it is possible for errors to be correlated across alternatives, thus potentially removing the IIA assumption (which we will get to in a moment). In fact, analysts typically discuss the multinomial probit model for the case where the errors are correlated since this is the only real advantage of the multinomial probit model over the multinomial logit model. However, MPROBIT assumes that the errors are uncorrelated i.e. $\text{cov}(\epsilon_j, \epsilon_i) = 0$. In other words, MPROBIT assumes IIA just like MLOGIT. If you use MLOGIT and MPROBIT on the same data, you will get almost identical predictions. Given the ease with which we can estimate MLOGIT models, there seems to be little point in estimating MPROBIT models. The important point to take away, though, is that you need to make sure you know exactly what type of multinomial probit model you are running – MPROBIT is multinomial probit with the assumption of IIA.

5 Mixed Conditional Logit Model - Alternative-Specific and Case-Specific Data

It is possible to combine the multinomial logit model with the pure conditional logit model to get what I am calling a mixed conditional logit model. This mixed model is often simply referred to as the conditional logit model, which obviously causes confusion since this is different from the pure conditional model that we described in the previous section. The mixed CL model allows you to examine how the characteristics of an individual n (MNL) – case-specific data – and the characteristics of some choice j (CL) – alternative-specific data – affect the probability that the decision maker will choose alternative i.

We start with the same basic utility equation from earlier:

$$U_{nj} = V_{nj} + \epsilon_{nj} \tag{60}$$

The systematic component of the utility function in the mixed CL model is given as:

$$V_{nj} = z_n\gamma_j + x_{nj}\beta \tag{61}$$

x_{nj} contains the values of the choice-specific variables for outcome j and individual n, while z_n contains individual-specific independent variables for individual n. So, the utility of individual n for choice j is:

$$U_{nj} = z_n\gamma_j + x_{nj}\beta + \epsilon_{nj} \tag{62}$$

To see all of this more clearly, let's imagine that our systematic component might be something like the following:

$$V_{nj} = \gamma_{j0} + \gamma_{j1}\text{Age}_n + \gamma_{j2}\text{Education}_n + \gamma_{j3}\text{Male}_n + \beta_1\text{IssueDistance}_{nj} \tag{63}$$

If we had three choices, then our systematic components would be:

$$\begin{aligned}
V_{n1} &= \gamma_{10} + \gamma_{11}\text{Age}_n + \gamma_{12}\text{Education}_n + \gamma_{13}\text{Male}_n + \beta_1\text{IssueDistance}_{n1} \\
V_{n2} &= \gamma_{20} + \gamma_{21}\text{Age}_n + \gamma_{22}\text{Education}_n + \gamma_{23}\text{Male}_n + \beta_1\text{IssueDistance}_{n2} \\
V_{n3} &= \gamma_{30} + \gamma_{31}\text{Age}_n + \gamma_{32}\text{Education}_n + \gamma_{33}\text{Male}_n + \beta_1\text{IssueDistance}_{n3}
\end{aligned} \tag{64}$$

Now substituting the systematic components into our utility equation, we have:

$$\begin{aligned}
U_{n1} &= \gamma_{10} + \gamma_{11}\text{Age}_n + \gamma_{12}\text{Education}_n + \gamma_{13}\text{Male}_n + \beta_1\text{IssueDistance}_{n1} + \epsilon_{i1} \\
U_{n2} &= \gamma_{20} + \gamma_{21}\text{Age}_n + \gamma_{22}\text{Education}_n + \gamma_{23}\text{Male}_n + \beta_1\text{IssueDistance}_{n2} + \epsilon_{n2} \\
U_{n3} &= \gamma_{30} + \beta_{31}\text{Age}_n + \gamma_{32}\text{Education}_n + \gamma_{33}\text{Male}_n + \beta_1\text{IssueDistance}_{n3} + \epsilon_{n3}
\end{aligned} \tag{65}$$

An example of what the data will look like might help. Imagine that there are four individuals choosing among three alternatives. Imagine also that there is just one choice-specific independent variable x_{nj} and one individual-specific independent variable z_n. n is the respondent's ID or individual n's ID.

Table 5: Data in a mixed CL Model

n	Outcome j	Outcome chosen	Variable x_{nj}	Variable z_n
1	1	0	z_{11}=1	x_1=1
1	2	1	z_{12}=0	x_1=1
1	3	0	z_{13}=7	x_1=1
2	1	1	z_{21}=2	x_2=4
2	2	0	z_{22}=1	x_2=4
2	3	0	z_{23}=6	x_2=4
3	1	1	z_{31}=1	x_3=2
3	2	0	z_{32}=6	x_3=2
3	3	0	z_{33}=9	x_3=2
4	1	0	z_{41}=1	x_4=1
4	2	0	z_{42}=3	x_4=1
4	3	1	z_{43}=2	x_4=1

Following the same logic as we had before, the probability that individual n chooses alternative i in the mixed CL model is:

$$P_{ni} = \frac{e^{z_n \gamma_j + x_{nj}\beta}}{\sum_j e^{z_n \gamma_j + x_{nj}\beta}} \text{ where } \gamma_1 = 0 \tag{66}$$

Note that we can rewrite this probability in terms of the systematic components of the utility function i.e.

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{67}$$

The important point here is that we can rewrite the probability that individual n chooses alternative i in exactly these same terms – in terms of the systematic components of the utility function – for the MNL model, the CL model, and the mixed CL model. In other words, the MNL model, CL model, and mixed CL model are all just forms of the exact same model with slightly different subscripts indicating whether we have alternative-specific and/or case-specific data.

5.1 Estimation

As you'll recall from earlier, I noted that it was not possible to include individual specific characteristics in a pure conditional logit model since they do not vary over the alternatives. In other words, they would cancel out of the probability equation - STATA would drop them. So, how do we get around this problem? In effect, we use a dummy variable type of trick. Essentially, we transform each individual-specific variable z_{nk} into a series of dummy variables that each take on the value of z_{nk} for each alternative, 0 otherwise. Recall that with the individual-specific MNL model, we want as many coefficients on each z_{nk} as there are alternatives. Creating the dummy variables in this way essentially produces as many coefficients on z_{nk} as there are alternatives. Obviously, when we estimate our model, we will only be able to include J-1 of these dummy variables. The omitted dummy variable will be associated with a particular alternative – this alternative will become the reference category against which all of the effects of the individual-specific variables will be compared. To see what the data set would now look like, see Table 6.

Table 6: Data in a mixed CL Model

n	Outcome j	Outcome chosen	Variable z_{nj}	Variable $z_{nk}1$	Variable $z_{nk}2$	Variable $z_{nk}3$
1	1	0	x_{11}=1	z_1=1	z_1=0	z_1=0
1	2	1	x_{12}=0	z_1=0	z_1=1	z_1=0
1	3	0	x_{13}=7	z_1=0	z_1=0	z_1=1
2	1	1	x_{21}=2	z_2=4	z_2=0	z_2=0
2	2	0	x_{22}=1	z_2=0	z_2=4	z_2=0
2	3	0	x_{23}=6	z_2=0	z_2=0	z_2=4
3	1	1	x_{31}=1	z_3=2	z_3=0	z_3=0
3	2	0	x_{32}=6	z_3=0	z_3=2	z_3=0
3	3	0	x_{33}=9	z_3=0	z_3=0	z_3=2
4	1	0	x_{41}=1	z_4=1	z_4=0	z_4=0
4	2	0	x_{42}=3	z_4=0	z_4=1	z_4=0
4	3	1	x_{43}=2	z_4=0	z_4=0	z_4=1

Let me explain how to create these dummy variables with reference to our Dutch elections data. The first thing to do is generate alternative specific dummies for our four parties.

```
generate vot1=(vote==1);
generate vot2=(vote==2);
generate vot3=(vote==3);
generate vot4=(vote==4);
```

Next we generate a series of interaction terms between our alternative specific dummies and each of our individual-specific variables. Below, I'll assume that we only have one individual-specific variable - whether the respondent is RELIGIOUS or not.

```
generate relgion1=relig*vote1;
generate relgion2=relig*vote2;
generate relgion3=relig*vote3;
generate relgion4=relig*vote4;
```

We are now ready to run our mixed CL model. Let's say that we have one individual-specific variable (RELIGION$_n$) and one choice-specific variable (DISTANCELEFT$_{nj}$) which measures the left-right distance between the individual and each alternative.

```
clogit choice vot2 vot3 vot4 religion2 religion3
                religion4 distanceleft, group(respid)
```

VOT2, VOT3, and VOT4 are the alternative-specific constant terms. Alternative 1 would be the base category against which the individual-specific effects would be evaluated.

5.2 Interpretation

The results from a mixed CL model of party choice in the Netherlands are shown in Table 7. The model examines how the respondent's probability of voting for a party is affected by several choice-specific variables (distance between the respondent and the party on the abortion, nuclear, income, and left-right issue dimensions) and several individual-specific variables (MALE, RELIGIOUS, EDUCATION).

Table 7: The Determinants of Party Choice in the Netherlands

Regressor	Mixed Conditional Logit		
AbortionDistance		-0.27***	
		(0.04)	
NuclearDistance		-0.12***	
		(0.04)	
IncomeDifferenceDistance		-0.42***	
		(0.05)	
RightDistance		-0.71***	
		(0.05)	
	Reference Party: Pvda		
	CDA	VVD	D66
Male	-0.64**	-0.88***	-0.53**
	(0.28)	(0.32)	(0.27)
Religious	1.10***	-0.35	0.26
	(0.30)	(0.31)	(0.26)
Education	0.49***	0.97***	0.56***
	(0.13)	(0.16)	(0.13)
Constant	-1.54***	-2.38***	-2.64***
	(0.47)	(0.53)	(0.44)
Log likelihood	-576.61	-576.61	-576.61
Observations	3461	3461	3461

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)
Standard errors are given in parentheses

5.2.1 Interpreting Coefficients

You interpret the coefficients in exactly the same way as you would in the MNL model for the individual-specific variables and in exactly the same way as you would in the pure CL model for the choice-specific variables. For example, we can infer that a respondent is less likely to vote for a party that is further away from him on any of the issue dimensions. We could even infer that respondents place twice as much importance on the abortion issue as the nuclear issue. We can also say that a religious respondent is more likely to vote for the CDA than the Pvda, but is no more likely to vote for the VVD or D66 than the Pvda.

5.2.2 Odds Ratios

Again, you can and should do more than simply interpret the coefficients. You can think in terms of odds ratios in exactly the same way as you did in the MNL and CL models. The odds that an individual will vote for a party a that is closer by one unit on the abortion issue than a party b is $e^{-0.27*-1}$=1.31 i.e. they will be more likely to vote for the closer party.[18] The odds that an individual voter will vote for the VVD if he were male compared to the identical voter who is female is $e^{-0.88*1}$=0.43 i.e. a male voter is less likely (57% less likely) to vote for the VVD than the Pvda all else equal than a female voter.

5.2.3 Predicted Probabilities and First Differences

We can also obtain predicted probabilities and first differences using:

$$P_{ni} = \frac{e^{z_n \gamma_j + x_{nj}\beta}}{\sum_j e^{z_n \gamma_j + x_{nj}\beta}} \text{ where } \gamma_1 = 0 \tag{68}$$

The predicted probability that a voter with the same characteristics as before (male, religious, high education) in a party system where the Pvda is two units further away from the respondent than the CDA on the nuclear issue (but the same distance away on all other issues) votes for the CDA is 0.36 [0.23, 0.50]. The predicted probability if the individual was female but with all the same characteristics is 0.45 [0.29, 0.62]. The effect of changing this individual from a male to a female (or the first difference) is to increase the predicted probability that the voter supports the CDA by 0.09 [0.03, 0.21]. Again, these results fit with our earlier inference that a woman was more likely to vote for the CDA than the Pvda compared to a man holding everything else constant.

6 Power and Limitations of Logit-Based Models

Train (2007, 46) notes that there are at least three limitations of the logit-based models that we have looked at:

1. Logit models can represent systematic taste variation (taste variation over the alternatives for observed characteristics) but not random taste variation (differences in taste that cannot be linked to observed characteristics).

2. Logit models imply proportional substitution across alternatives given the analyst's specification. In other words, logit models assume IIA and cannot capture more flexible forms of substitution across alternatives.

3. If unobserved factors are independent over time in repeated choice situations, then logit models can capture the dynamics of repeated choice. However, logit models cannot capture situations where

[18] It is -1 because the party a is closer.

unobserved factors are correlated over time.[19]

6.1 Taste Variation

The value that decision makers place on each attribute of the alternatives varies, in general, over decision makers. In other words, some people might value a certain characteristic of an alternative more than other people. The tastes of decision makers also vary for reasons that are not linked to observed demographic characteristics – different people are different. In other words, you can have two people who have the same income, same education, and so on, make different choices because they have different individual preferences and concerns. Logit models can only capture some types of taste variation. Specifically, logit models can capture tastes that vary systematically with respect to the observed variables. For example, you could include an interaction term between an alternative-specific variable and an individual-specific variable to capture the idea that the alternative-specific variable is more important for some individuals than others. A limitation of logit models, though, is that it cannot handle tastes that vary with unobserved (or unmeasured) variables or purely randomly. Consider the following example. Suppose we had the following utility equation with two alternative-specific variables, x_j and z_j:

$$U_{nj} = \alpha_n x_j + \beta_n z_j + \epsilon_{nj} \tag{69}$$

where α_n and β_n are parameters specific to individual n i.e. the effect of x_j and z_j is assumed to vary over individuals. Let's suppose that the value of x_j for each individual depends on some observed characteristic but also unobserved characteristic:

$$\alpha_n = \rho M_n + \mu_n \tag{70}$$

where M_n is some observed characteristic of individual n and μ_n is a random variable. Similarly, we can say that the importance of z_j also consists of an observed and unobserved component:

$$\beta_n = \tau N_n + \eta_n \tag{71}$$

Because μ_n and η_n are unobserved, the terms $\mu_n x_j$ and $\eta_n z_j$ become part of the unobserved component of utility:

$$U_{nj} = \rho(M_n x_j) + \tau(N_n z_j) + \tilde{\epsilon}_{nj} \tag{72}$$

where $\tilde{\epsilon}_{nj} = \mu_n x_j + \eta_n z_j + \epsilon_{nj}$. It should be obvious that the new error terms $\tilde{\epsilon}_{nj}$ cannot be distributed independently and identically as required for the logit models. Since μ_n and η_n enter each alternative, $\tilde{\epsilon}_{nj}$ is necessarily correlated over alternatives. Because x_j and z_j vary over alternatives, the variance of $\tilde{\epsilon}_{nj}$ varies over alternatives, thus violating the assumption of identically distributed errors as well.

The bottom line is that if tastes are expected to vary systematically in relation to observed variables, then taste variation can be incorporated into logit models. However, if tastes vary at least partly randomly, then a logit model is a misspecification.

[19] We will not address this last point in any detail here. However, we will come back to it when we discuss discrete time duration models in a few weeks.

6.2 Substitution Patterns and the Independence of Irrelevant Alternatives

If the attributes of some alternative improve, then the probability that this alternative is chosen goes up. But because probabilities need to sum to one, this means that an increase in the probability of one alternative necessarily means a decrease in the probability of other alternatives. The way in which these probabilities change is referred to as the pattern of substitution. All of the logit-based models that we have examined - MNL and CL models – assume a particular pattern of substitution known as the independence of irrelevant alternatives (IIA).

6.2.1 What is IIA

Recall that for any two alternatives i and k, the ratio of logit probabilities is the following:

$$\begin{aligned}\frac{P_{ni}}{P_{nk}} &= \frac{e^{V_{ni}}/\sum_j e^{V_{nj}}}{e^{V_{nk}}/\sum_j e^{V_{nj}}} \\ &= \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni}-V_{nk}} \qquad (73)\end{aligned}$$

The key thing to note here is that this ratio does not depend on any alternatives other than i and k. In other words, the relative odds of choosing i or k are the same no matter what other alternatives are in the choice set or what attributes these other alternatives have. Because the ratio is independent of alternatives other than i and k, it is said to be independent from irrelevant alternatives. As you can see, logit models have the property of IIA. If IIA does not hold, then the estimated coefficients will be inconsistent.

Let's think of a particular example. Consider an individual choosing whether to vote for the Conservatives (C) or Labor (L) in the UK during the 1980s. IIA implies that the presence of a third choice or party, the Alliance (A), should not change the relative probabilities of any single voter choosing between C and L.

$$\frac{P_n(C)|\{C,L\}}{P_n(L)|\{C,L\}} = \frac{P_n(C)|\{C,L,A\}}{P_n(L)|\{C,L,A\}} \qquad (74)$$

Say there are two parties (C and L) and say an individual voter has a probability of 0.5 of voting for C and a probability of 0.5 of voting for L. Thus, the ratio of the probability of voting for C and the probability of voting for L is 1. IIA implies that if a third party enters (A), then the voter must have a probability of 0.33 of voting for C, 0.33 of voting for L, and 0.33 of voting for L since this is the only way of keeping the ratio between the probability of voting for C and L equal to 1. IIA really becomes a problem when some of the alternatives in the choice set are seen as substitutes for each other.[20]

Consider the famous red-bus-blue-bus example where two transportation choices are exact substitutes for each other. Suppose an individual has to choose whether to go to work by car or by taking a blue bus. Let's

[20] It is important to note that IIA refers to the relative probabilities of **individual** voters choosing parties. It is possible for IIA to hold at the individual level, but not at the aggregate level. An example is given by Alvarez and Nagler (1998). You should take a look at this.

suppose that $P_c = P_{bb} = 0.5$ such that $\frac{P_c}{P_{bb}} = 1$. Now suppose that a red bus is introduced. The probability that the worker will take the red bus is the same as the probability he will take the blue bus i.e. $\frac{P_{rb}}{P_{bb}} = 1$. In the logit model, it has to be the case that $\frac{P_c}{P_{bb}}$ is the same whether or not another alternative, such as the red bus, exists or not; this ratio has to remain 1. The only probabilities for which $\frac{P_c}{P_{bb}} = 1$ and $\frac{P_{rb}}{P_{bb}} = 1$ are $P_c = P_{bb} = P_{rb} = 0.5$, which are the probabilities predicted by logit. The problem is that in reality we would expect the probability of taking the car to remain the same when a new bus that is identical to the old bus is introduced. We would also expect the original probability of taking the bus to be split between the two buses once the second bus is introduced. Thus, what we would expect is $P_c = 0.5$ and $P_{bb} = P_{rb} = 0.25$. The logit model, because of the IIA property, overestimates the probability of taking either of the buses and underestimates the probability of taking the car.

An incorrect assumption of IIA really has to do with the non-independence of the error terms. Note that even if IIA is correct in the abstract, it will appear that we have violated it if we omit a variable that is common to two choices. This is because the omitted variable is being captured in the error terms and making them appear correlated. In this sense, the problem has less to do with correlated disturbances and more to do with the fact that there is an omitted variable. Note, though, what this means. IIA is a property of a specific set of variables and choices - it is not an abstract property of decision makers. This is important because it means that we should not infer from the results of our models that, say, IIA holds in one country but not in another[21] This is because the reason why IIA appears violated in one county may have to do with the fact that their model is just not as good as it is in the other country i.e. it omits more relevant variables. A more general point here is that we don't want to interpret or infer things from error processes too much because we know that we can get rid of these errors if we had a better model.

The bottom line here is that IIA will be violated if decision makers perceive alternatives to be substitutes for each other or if we omit variables that are common to two or more alternatives.

6.2.2 Testing for IIA

One way to think about IIA is that the coefficients should be invariant to which choices are available. This allows us to conduct an eyeball test of the IIA assumption. In our Dutch example, we have 4 parties. To rule out a violation of the IIA assumption, we could check to see if the coefficients are the same when we drop either the Pvda, the CDA, D66, or VVD. Note that we should also check that the coefficients are the same if we drop combinations of these parties i.e. Pvda and CDA, Pvda and VVD, Pvda and D66, D66 and VVD, D66 and CDA, and CDA and VVD. IIA assumes that dropping any of these parties or combinations of parties will not change the coefficients. The Hausman-McFadden test of IIA is essentially a more formal version of this eyeball test.

Hausman-McFadden Test

The Hausman-McFadden test is based on the comparison of two estimators of the same parameters. One estimator is consistent and efficient if the null hypothesis is true (i.e. IIA holds), while the second estimator is consistent but inefficient. For multinomial logit and conditional logit, maximum likelihood is consistent and

[21] This is essentially what Martin, Quinn and Whitford (1999) do in their analysis.

efficient if the model is correctly specified. A consistent but inefficient estimator is obtained by estimating the model on a restricted set of outcomes. If IIA holds and the dropped choices are irrelevant, the estimates of the parameters should be the same. So the test is:

1. Estimate the full model with all J alternatives included. This produces $\hat{\beta}_F$ and $\hat{V}_F$.
2. Fit a restricted model by eliminating one or more alternatives. This produces $\hat{\beta}_R$ and $\hat{V}_R$.
3. The following test statistic is asymptotically distributed as a χ^2 random variable with k degrees of freedom where k is the number of elements in the β vector:

$$(\hat{\beta}_R - \hat{\beta}_F)'[\hat{V}_R - \hat{V}_F]^{-1}(\hat{\beta}_R - \hat{\beta}_F) \sim \chi^2_k \tag{75}$$

To do the Hausman test in STATA, we type:

```
clogit Y X, group(respid);
est store full;
clogit Y X if vote~=4, group(respid);
hausman . full;
```

This test looks to see if IIA is violated when we drop alternative 4.

Another formal test of IIA is the Small-Hsiao test. If you have SPOST, you can conduct this test by typing:

```
clogit Y X, group(respid);
mlogtest, smhsiao
```

Numerous problems arise with these tests. For example, Long (2001, 191) notes that the Hausman Test often gives inconsistent results. Similar complaints can be made about the Small-Hsiao test which requires randomly splitting the sample into two subsamples; the results of the test often depend on how the sample was divided. Another problem is that the two tests often give conflicting information on whether the IIA assumption has been violated. These practical limitations led McFadden (1973) to suggest that the assumption of IIA implies that multinomial and conditional logit models should only be used in cases where the outcome categories “can plausibly be assumed to be distinct and weighed independently in the eyes of each decision maker.” You might think of moving to discrete choice models that relax or do not require the IIA assumption. It is to these models that we now turn.

References

Alvarez, R. Michael & Jonathan Nagler. 1998. "When Politics and Models Collide: Estimating Models of Multiparty Elections." *American Journal of Political Science* 42:55–96.

Long, J. Scott. 2001. *Regression Models for Categorical and Limited Dependent Variables Using STATA*. Texas: STATA Corporation.

McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press.

Quinn, Kevin M., Andrew D. Martin & Andrew B. Whitford. 1999. "Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science* 43:1231–1247.

Train, Kenneth E. 2007. *Discrete Choice Models with Simulation*. New York: Cambridge University Press.