

Models for Unordered Outcomes

Political Science 150C/350C

Simon Jackman

Revision 134

April 30, 2007

Here we consider models for dependent variable that take on values that lack quantification at all. Variables of this type are sometimes referred to as nominal variables, and any numeral we assign to them is purely arbitrary and for the purposes of labelling or indexing only.

Classic examples come from labor economics or transport economics, where consumer choices of interest to the analyst take on the values such as {"car", "bus", "train"} or, say in a voting example {"Conservative", "Labor", "Alliance"} or {"Bush", "Clinton", "Perot", "Abstain" }

1 Random Utility Rationale

In economics and political science, models for unordered dependent variables are usually motivated via a random utility story. That is, we assume that decision-maker i faces a choice over J outcomes, which for convenience are labelled "0", "1", "2", etc, but where this labelling implies nothing about any ordering over the choices.

The utility to decision-maker i of choice j is linear in some predictors, plus a random component,

$$U_{ij} = \mathbf{x}_i \boldsymbol{\beta}_j + \varepsilon_{ij}, \quad j = 0, \dots, J.$$

2 Multinomial Logit (MNL)

A common assumption is that the ε_{ij} are from a distribution whose cumulative distribution function (CDF) is a Type-1 extreme-value distribution with functional form

$$F(\varepsilon_{ij}) = \exp[-\exp(-\varepsilon_{ij})] \tag{1}$$

and hence with density (pdf)

$$f(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \exp[-\exp(-\varepsilon_{ij})]. \tag{2}$$

The CDF and PDF for this distribution are plotted in Figure 1. Note the slight skew. This distribution is assumed identical across observations, and critically, **independent across the**

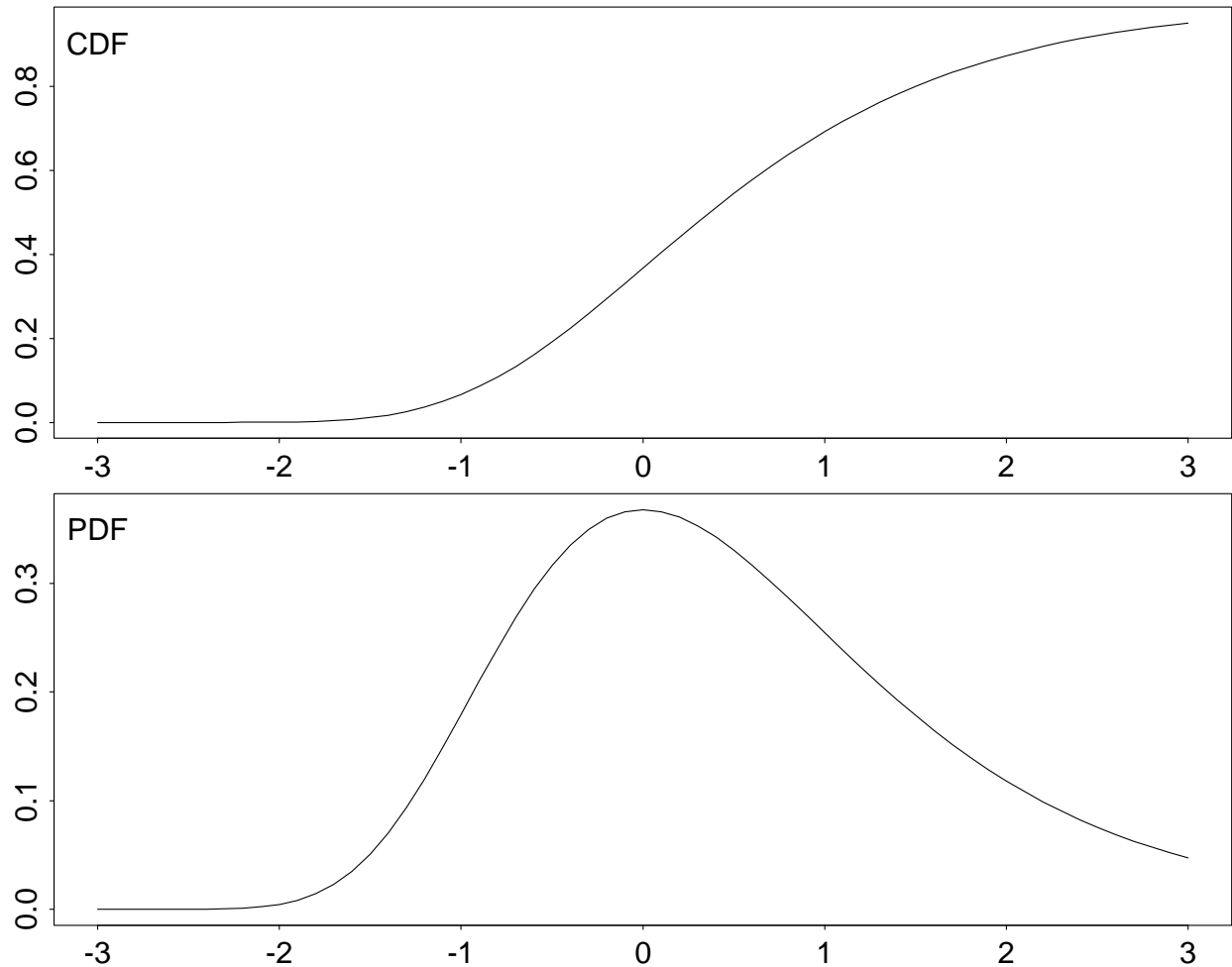


Figure 1: Type-1 Extreme Value distribution (“log-Weibull distribution”).

alternatives in the choice set. This functional form is chosen more for tractability rather than specifying something interesting about the utilities; as I show below, these distributional assumptions lead to a convenient “log-odds” interpretation of the linear combination $\mathbf{x}_i\boldsymbol{\beta}$.

Sometimes a multivariate normal distribution is posited as the joint pdf of the ε_{ij} , which would allow the possibility of covariation among the stochastic components to the utilities; see the discussion of IIA and the MNP model below.

Under the postulate of utility maximization, decision-maker i chooses j if

$$U_{ij} > U_{ik}, \quad \forall k \neq j.$$

Because the utilities contain a stochastic component ε_{ij} , choice here is probabilistic. Decision-maker i chooses option j with probability

$$P(y_i = j) = P[U_{ij} > U_{ik}], \quad \forall k \neq j.$$

Consider a choice set with 3 elements, {"0", "1", "2"}. Following [Amemiya \(1985, 297\)](#),

$$\begin{aligned}
 P(y_i = 2) &= P(U_{i2} > U_{i1}, U_{i2} > U_{i0}), \\
 &= P[\mathbf{x}_i\boldsymbol{\beta}_2 + \varepsilon_{i2} > \mathbf{x}_i\boldsymbol{\beta}_1 + \varepsilon_{i1}, \mathbf{x}_i\boldsymbol{\beta}_2 + \varepsilon_{i2} > \mathbf{x}_i\boldsymbol{\beta}_0 + \varepsilon_{i0}], \\
 &= P[\varepsilon_{i2} + \mathbf{x}_i\boldsymbol{\beta}_2 - \mathbf{x}_i\boldsymbol{\beta}_1 > \varepsilon_{i1}, \varepsilon_{i2} + \mathbf{x}_i\boldsymbol{\beta}_2 - \mathbf{x}_i\boldsymbol{\beta}_0 > \varepsilon_{i0}], \\
 &= \int_{-\infty}^{\infty} f(\varepsilon_2) \left[\int_{-\infty}^{\varepsilon_{i2} + \mathbf{x}_i\boldsymbol{\beta}_2 - \mathbf{x}_i\boldsymbol{\beta}_1} f(\varepsilon_1) d\varepsilon_1 \cdot \int_{-\infty}^{\varepsilon_{i2} + \mathbf{x}_i\boldsymbol{\beta}_2 - \mathbf{x}_i\boldsymbol{\beta}_0} f(\varepsilon_0) d\varepsilon_0 \right] d\varepsilon_2, \\
 &= \int_{-\infty}^{\infty} f(\varepsilon_2) \times \exp[-\exp(-\varepsilon_{i2} - \mathbf{x}_i\boldsymbol{\beta}_2 + \mathbf{x}_i\boldsymbol{\beta}_1)] \times \exp[-\exp(-\varepsilon_{i2} - \mathbf{x}_i\boldsymbol{\beta}_2 + \mathbf{x}_i\boldsymbol{\beta}_0)] d\varepsilon_2, \\
 &= \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_2)}{\exp(\mathbf{x}_i\boldsymbol{\beta}_0) + \exp(\mathbf{x}_i\boldsymbol{\beta}_1) + \exp(\mathbf{x}_i\boldsymbol{\beta}_2)}.
 \end{aligned}$$

Generically,

$$P(y_i = j) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_j)}{\sum_{k=0}^J \exp(\mathbf{x}_i\boldsymbol{\beta}_k)}.$$

2.1 Alternative Derivation

Here is an alternative derivation, for which I find the math slightly easier to follow, from [Maddala \(1983, 60-61\)](#). Let $y_{ij}^* = \mathbf{x}_i\boldsymbol{\beta}_j + \varepsilon_{ij} = V_{ij} + \varepsilon_{ij}$ and if the j th outcome is chosen, then (suppressing the i subscript over individuals):

$$V_j + \varepsilon_j > V_k + \varepsilon_k, \quad \forall k \neq j$$

which implies that

$$\varepsilon_k < V_j + \varepsilon_j - V_k, \quad \forall k \neq j$$

Then,

$$\begin{aligned}
 P(y_i = j) &= \int_{-\infty}^{\infty} \left[\prod_{k \neq j} F(V_j + \varepsilon_j - V_k) \right] f(\varepsilon_j) d\varepsilon_j \\
 &\text{substituting the CDF and PDF in equations (1) \& (2)} \\
 &= \int_{-\infty}^{\infty} \prod_{k \neq j} \exp \left[-e^{-(V_j + \varepsilon_j - V_k)} \right] \exp \left[-\varepsilon_j - e^{-\varepsilon_j} \right] d\varepsilon_j \\
 &= \int_{-\infty}^{\infty} \exp \left[-\varepsilon_j - \left(1 + \sum_{k \neq j} \frac{e^{V_k}}{e^{V_j}} \right) e^{-\varepsilon_j} \right] d\varepsilon_j \\
 &= \int_{-\infty}^{\infty} \exp \left[-\varepsilon_j - e^{-\varepsilon_j + \lambda_j} \right] d\varepsilon_j,
 \end{aligned}$$

where

$$\begin{aligned}\lambda_j &= \ln \sum_{k=0}^J \frac{e^{V_k}}{e^{V_j}} \\ &= \ln \left[1 + \sum_{k \neq j} \frac{e^{V_k}}{e^{V_j}} \right]\end{aligned}$$

Continuing, $\exp(-\lambda_j)$ can be brought in front of the integration:

$$\begin{aligned}P(y_i = j) &= \exp(-\lambda_j) \int_{-\infty}^{\infty} \exp[-\varepsilon_j - e^{-\varepsilon_j}] d\varepsilon_j \\ &= \exp(-\lambda_j) \\ &\quad \text{since the term after the integral is the pdf of } \varepsilon_j, \text{ and so integrates to unity} \\ &= \frac{e^{V_j}}{\sum_{k=0}^J e^{V_k}}\end{aligned}$$

or, in terms of the data and the parameters,

$$P(y_i = j) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{\sum_{k=0}^J \exp(\mathbf{x}_i \boldsymbol{\beta}_k)}$$

2.2 Log-likelihood function

We can now specify a log-likelihood function for the multinomial logit (MNL) model

$$\ln \mathcal{L} = \sum_{i=1}^N \sum_{j=0}^J Z_{ij} \ln P(y_i = j),$$

where $Z_{ij} = 1$ if $y_i = j$ and 0 otherwise. This log-likelihood is globally concave, and the values of $\boldsymbol{\beta}$ that maximize this function have the usual MLE properties.

2.3 Identification via normalizing on baseline outcome

As it stands though, the model's parameters are not identified. It is not possible to recover estimates of all $J + 1$ $\boldsymbol{\beta}_j$ parameter vectors (remembering $j = 0, \dots, J$); any of the $\boldsymbol{\beta}_j$ could be changed by some linear transformation and an identical set of estimated probabilities would result. The standard normalization employed in the literature is to make the a “baseline” category against which other alternatives are assessed, via log-odd ratios. This normalization is made by setting $\boldsymbol{\beta}_0 = 0$.

Log-odds ratios are just the log of the ratio of the probabilities of two events. This is the usual way to interpret the MNL model. The model can be said to be linear with respect to the

log-odds ratio of an outcome and the baseline category. To see this, first consider the effect of setting $\beta_0 = 0$:

$$\frac{P_{ij}}{P_{i0}} = \frac{\exp(\mathbf{x}_i \beta_j)}{\exp(\mathbf{x}_i \beta_0)} = \frac{\exp(\mathbf{x}_i \beta_j)}{\exp(0)} = \exp(\mathbf{x}_i \beta_j),$$

since if $\beta_0 = 0$, $\exp(\mathbf{x}_i \beta_0) = \exp(0) = 1$. Accordingly,

$$\ln \left(\frac{P_{ij}}{P_{i0}} \right) = \mathbf{x}_i \beta_j.$$

After this normalization, maximizing the log-likelihood results in estimates of J k by 1 vectors of parameters. The probabilities of observing the respective choices are now

$$\Pr(y_i = j) = \frac{\exp(\mathbf{x}_i \beta_j)}{1 + \sum_{k=1}^J \exp(\mathbf{x}_i \beta_k)} \quad \forall j = 1, 2, \dots, J \quad (3)$$

$$\Pr(y_i = 0) = \frac{1}{1 + \sum_{k=1}^J \exp(\mathbf{x}_i \beta_k)} \quad (4)$$

The parallels between MNL and the logit model for binary outcomes is no mere coincidence. The model amounts to a series of binary choices, considered simultaneously, but independently. Note that only the β_j terms appear in the expression for $P(y_i = j)$ after the normalization on a baseline outcome (but see below).

2.4 Marginal Effects

As is common in discrete choice models, care must be taken when interpreting the coefficients. [Greene \(2003\)](#) notes that “it is tempting to associate β_j with the j th outcome, but this would be misleading”. Differentiating equation (3) shows that each β_j enters into a consideration of the effects of x on a particular $P(y_i = j)$:

$$\delta_j = \frac{\partial P_j}{\partial \mathbf{x}_i} = P_j \left[\beta_j - \sum_{k=0}^J P_k \beta_k \right] = P_j [\beta_j - \bar{\beta}] \quad (5)$$

where $\bar{\beta}$ is a weighted average of the β_k , $k = 0, \dots, J$. Accordingly, the effect of a change in x on the probability that $y_i = j$ is non-linear not only with respect to β_j , but also depends on the predicted probabilities for the other categories.

2.5 Goodness of Fit

As for binary choice models, there are no firm ideas as to how to assess goodness-of-fit. Two widely used proposals are to compare the model’s performance against a null model containing only an intercept. A version of McFadden’s pseudo- r^2 might be used (based on the difference in the likelihoods for the full model versus the intercept-only model). A table of hits and misses is also a useful summary, using the prediction rule $\hat{y}_i = j \iff \hat{P}(y_i = j) > \hat{P}(y_i = k) \forall j \neq k$.

2.6 Conditional Logit vs Multinomial Logit

Econometricians distinguish between two types of predictor variables: data specific to individuals *and* choices, and data specific to individuals but constant over choices (e.g., objective social-structural features of the choice-makers). We write the first type of regressor as \mathbf{x}_{ij} , where i indexes individuals and j indexes choices. The second type of regressor is \mathbf{x}_i , picking up only the i subscript since it is constant with respect to choices for a given individual.

A model which uses just the former type of regressors (choice-specific predictors) is called a “conditional logit” model. Note that the data used here can vary over individuals *and* choices, or simply over choices (e.g., features of the choices that are constant over individuals). Conditional logit models have the feature that the predictors vary over choices (and possibly individuals too), but the parameters do not, and so the conditional model is

$$\Pr(y_i = j) = \frac{\exp(\mathbf{z}_{ij}\boldsymbol{\beta})}{\sum_{j=0}^J \exp(\mathbf{z}_{ij}\boldsymbol{\beta})} \quad (6)$$

where \mathbf{z}_{ij} is the vector of regressors associated with the i th individual on the j th choice. In the transportation choice example these variables comprised a series of choice-specific constant terms, plus data on the costs associated with each choice. It is important to note that the conditional logit model contains no normalizations on a baseline outcome, and so no intercept is estimated for this model.

The constraint that the coefficients are constant over choices means that it is differences on the \mathbf{x}_j (or \mathbf{x}_{ij}) that account for the different choices observed. That is, for the conditional logit model, the log-odds of choosing option j over k is

$$\ln \left(\frac{p_{ij}}{p_{ik}} \right) = (\mathbf{x}_{ij} - \mathbf{x}_{ik})\boldsymbol{\beta}$$

which contrasts with the multinomial logit log-odds

$$\ln \left(\frac{p_{ij}}{p_{ik}} \right) = \mathbf{x}_i(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)$$

with \mathbf{x}_i constant across choices and differences in the $\boldsymbol{\beta}$ accounting for the differences in observed behavior.

Marginal effects are again a little tricky to work out, but have the form [Greene \(2003, 723\)](#)

$$\frac{\partial P_j}{\partial \mathbf{x}_k} = [P_j(\mathbf{1}(j = k) - P_k)] \boldsymbol{\beta}, \quad (7)$$

where $\mathbf{1}(j = k)$ is a dummy variable, equal to 1 if $j = k$ and 0 otherwise. Importantly, attributes of *all* choices enter into this marginal effect of any x_k on any P_j .

It should be noted that any MNL model can be written as a conditional logit model, and vice-versa; a simple example is given in [Long \(1997, 180-1\)](#).

2.7 Multinomial/Conditional Logits

A general mixed or hybrid model can be obtained by combining the multinomial and conditional logit models, say where we have available data that is constant over choices (e.g., choice-makers' social-structural characteristics) and data that varies over choices (and possibly individuals as well), such as costs/benefits of choices or individual-specific perceptions of those costs/benefits. The more general model can be written as

$$U_{ij} = \mathbf{x}_i \boldsymbol{\beta}_j + \mathbf{z}_{ij} \boldsymbol{\gamma} + \varepsilon_{ij}$$

where \mathbf{x}_i are characteristics of individuals that are constant across choices, and \mathbf{z}_{ij} are characteristics that vary across individuals and choices. See Powers and Xie (2000, §7.6.2) for more details.

Note that it is also possible to estimate models where the \mathbf{z}_{ij} pick up coefficients specific to each choice: e.g.,

$$U_{ij} = \alpha_j + \mathbf{z}_{ij} \boldsymbol{\gamma}_j + \varepsilon_{ij}$$

is estimable, provided we impose an identifying restriction that one of the $\alpha_j = 0$.

2.8 Independence of Irrelevant Alternatives - IIA

This form of the discrete choice model assumes that the stochastic component of the utilities for each alternative are independent. This has the virtue of leading to the relatively simply looking log-likelihood function for the model (above) and the relatively straightforward interpretation of the parameters estimates in terms of log-odds. Nonetheless, these features of the model rest on what is often a dubious assumption, that the stochastic component of the utility associated with each outcome is uncorrelated with stochastic components for the other outcomes. In practice this is often not the case.

A useful thought experiment to see if this property is valid in a given setting is to ask whether the relative probabilities (i.e., the the log-odds ratio) for alternatives j and k remains the same whether or not another alternative l is in the choice-set. This is sometimes referred to as the “red bus - blue bus” problem, since it is often illustrated with reference to the following problem:

An analyst observes people making transportation choices from the set {“car”, “red bus”} with the probabilities { 3/5, 2/5}. The odds-ratio is 3/2. With the introduction to the choice set of the seeming innocuous alternative “blue bus” the probabilities change to { 3/5, 1/5, 1/5} (people are indifferent about color) and the odds-ratio between “car” and “red bus” changes to 3.

Amemiya (1985) also puts the problem this way:

Using McFadden’s famous example, suppose that the three alternatives...consist of car, red bus, and blue bus.... In such a case, the independence between ε_1 and ε_2 is a clearly unreasonable assumption because a high (low) utility for red bus should generally imply a high (low) utility for a blue bus. The probability $P_0 = P(U_0 > U_1, U_0 > U_2)$ calculated under the independence assumption would

underestimate the true probability in this case because the assumption ignores the fact that the event $U_0 > U_1$ make the event $U_0 > U_2$ more likely (p298).

That is, the model specifies the relative probabilities between a pair of alternatives without regard for the the third alternative. For instance, the probability of choosing j or k is specified without any regard at all for the probability of choosing any other outcome $r \neq j, k$. We can see this by simply noting that given that

$$P(y_i = j) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{\sum_{k=0}^J \exp(\mathbf{x}_i \boldsymbol{\beta}_k)}$$

then

$$\begin{aligned} P(y_i = j | y_i = j \text{ or } y_i = k) &= \frac{P(y_i = j)}{P(y_i = j) + P(y_i = k)} \\ &= \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j) [\exp(\mathbf{x}_i \boldsymbol{\beta}_j) + \exp(\mathbf{x}_i \boldsymbol{\beta}_k) + \exp(\mathbf{x}_i \boldsymbol{\beta}_r)]^{-1}}{[\exp(\mathbf{x}_i \boldsymbol{\beta}_j) + \exp(\mathbf{x}_i \boldsymbol{\beta}_k)] [\exp(\mathbf{x}_i \boldsymbol{\beta}_j) + \exp(\mathbf{x}_i \boldsymbol{\beta}_k) + \exp(\mathbf{x}_i \boldsymbol{\beta}_r)]^{-1}} \\ &= \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{\exp(\mathbf{x}_i \boldsymbol{\beta}_j) + \exp(\mathbf{x}_i \boldsymbol{\beta}_k)} \end{aligned}$$

and so the relative probability of j vis-a-vis k is independent of the r , or even the presence or absence of r in the choice set!

Now it may be that we have done a bad job of modeling preferences here or posed the question badly (if people are indifferent between bus color then in what sense is the distinction between red and blue buses substantively interesting?). Nonetheless, one can imagine political science examples in which this is a more pressing problem, say thinking about how the odds of voting for Clinton over Bush change with the absence or presence of Perot in the choice set, and at different times in 1992 this was a real concern (Perot entering, leaving, and re-entering the race). [Whitten and Palmer \(1996\)](#) use the MNL model to analyze vote choice in multi-party Dutch and British elections.

If IIA is violated, then the MNL estimates of $\boldsymbol{\beta}$ are inconsistent. [Hausman and McFadden \(1984\)](#) propose a test of IIA, a specialized version of Hausman's (1978) more general test for model specification.¹ Taking the Bush-Clinton-Perot vote choice example, a full, unrestricted MNL would yields $2k$ by 1 vectors of parameters, $\hat{\boldsymbol{\beta}}_1$ for the Clinton/Bush odds ratios, and $\hat{\boldsymbol{\beta}}_2$ for the Perot/Bush odds ratios (setting Bush as the baseline category), and a $2k$ by $2k$ variance-covariance matrix of the parameters, with the upper-left k by k sub-matrix V_f containing the variances and covariances of the parameter estimates for the Clinton/Bush comparison.

A restricted model in this case amounts to a binary logit (logistic regression), ignoring the Perot voters, with a vote for Clinton as the $y_i = 1$ outcome, and a vote for Bush as the $y_i = 0$ outcome. This restricted model yields just one k by 1 parameter vector, corresponding to the $\hat{\boldsymbol{\beta}}_1$ in the unrestricted MNL, and an accompanying k by k variance-covariance matrix I will

¹This is a very general test: "The only requirement of the test is that we have an estimator, usually a maximum likelihood estimator, that is asymptotically efficient under the null hypothesis but loses consistency under an alternative hypothesis and another estimator that is asymptotically less efficient than the first under the null hypothesis but remains consistent under an alternative hypothesis" ([Amemiya 1985](#), 145).

designate as V_r . The Hausman-McFadden test statistic for the irrelevance of the Perot choice to the Clinton/Bush choice is

$$(\hat{\beta}_r - \hat{\beta}_f)' [V_r - V_f]^{-1} (\hat{\beta}_r - \hat{\beta}_f)$$

which is distributed χ^2 with k degrees of freedom. The null hypothesis is that the two parameters vectors are equal, and that the Perot option is irrelevant to the Clinton/Bush choice.

Alvarez and Nagler (1998) critique the Whitten and Palmer (1996) analysis of British and Dutch elections on these grounds. Whitten and Palmer do not appear to test for IIA, but nonetheless argue that if one believes that IIA is violated then nested logit might be a preferable alternative. I discuss nested logit below, along with another alternative to MNL, the multinomial probit model.

3 Multinomial Probit (MNP)

MNP specifies a $J + 1$ multivariate Normal density for the disturbances in the utility functions. The off-diagonal elements of the $J + 1$ by $J + 1$ variance-covariance matrix of the disturbances explicitly operationalize the interdependence of the choices. Rather than do MNL and then test for IIA, MNP embeds a test of independence in the model itself, via the resulting estimates of the covariance parameters.

Formally, we have a random utility setup,

$$U_{ij} = \mathbf{x}_i \beta_j + \varepsilon_{ij}, \quad j = 0, 1, \dots, J,$$

but a joint, multivariate density on the ε_{ij} ,

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

with variances $\sigma_1^2 = \sigma_{11}, \sigma_2^2 = \sigma_{22}, \dots, \sigma_{J+1}^2 = \sigma_{J+1, J+1}$ on the leading diagonal of $\boldsymbol{\Sigma}$, and covariances σ_{12}, \dots as the off-diagonals.

The probability of observing outcome j is

$$\begin{aligned} P(y_i = j) &= P(U_{ij} > U_{ik}), \quad \forall k \neq j \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{U_{ij}} \dots \int_{-\infty}^{U_{ij}} f(U_0, U_1, \dots, U_j) dU_0 dU_1 \dots dU_j \end{aligned}$$

where f is the $J + 1$ -variate Normal density defined above. In the one-dimensional case encountered for binary or ordered probit this integral is simply $\Phi(\cdot)$. Evaluating integrals of multivariate Normal densities is a reasonably complex problem and only a few software packages support routines for tri-variate Normal densities. Higher dimensional problems can be extremely computationally burdensome, especially when embedded in an optimization problem (i.e., MLE) and require specialized programming on a case-by-case basis. For this reason the MNP model has not found wide application beyond the case of three or, at most, four outcome choice problems.

In general a m outcome problem can be reduced to a $m - 1$ outcome problem by normalizing one of the coefficient vectors and collapsing one dimension of the m -variate density.

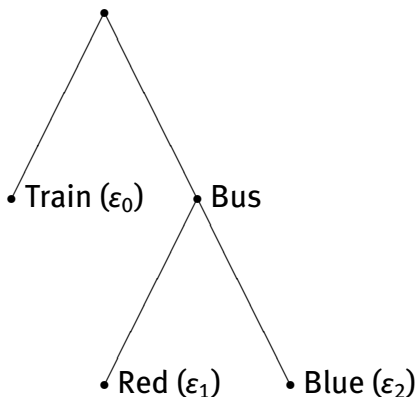
Another standard normalization is to set the variances of the disturbance terms to 1, which is equivalent to an assumption of homoskedasticity in the disturbances across the choices. This normalization also means that the off-diagonal elements of Σ are interpretable as correlations instead of covariances (Keane 1992; McCulloch, Polson and Rossi 1998).

Even with these normalizations, applied work has mixed findings with the MNP model. Alvarez and Nagler (1995) report that in even a reasonably large sample ($N=1000$) reliable estimates of the covariance terms are hard to come by, even in a relatively straightforward three outcome model. In their work on the 1992 U.S. presidential election, they estimate small correlations between the errors in the estimated utilities for the three candidates, all of which are swamped by their standard errors: $\hat{\sigma}_{BC} = -.08(.28)$, $\hat{\sigma}_{BP} = .27(.54)$, $\hat{\sigma}_{CP} = -.07(.26)$, where B, C, P stand for Bush, Clinton, Perot, respectively.

Some recent work with MNP in political science exploits the fact that we can now approximate the high-dimensional integrals via Monte Carlo methods, exploiting ideas from Bayesian statistics. This is a separate topic in its own right, but I refer you to my own piece on this (Jackman 2000), and a similar application by Quinn and Martin (1998). See also the MNP library in R, and the article by Imai and van Dyk (2005).

4 Nested Logit

Sometimes alternatives in a choice set can be grouped together. Reconsidering the “red bus, blue bus” problem, it seems sensible to consider that these “bus” options should go together, while the “train” or “car” options can be considered separately. Having chosen “bus” over “train” or “car”, the traveler’s choice becomes one between the two colors of bus. In this sense can the choice between the red and the blue buses be considered as “nesting” in a larger choice setting:



Many will recognize the parallels between this and a sequential game tree. While utilities attach to each terminal node (and I have labelled these with the stochastic component of those utilities, $\varepsilon_j, j = 0, \dots, m$), there is no notion of strategy here. The “nesting” here is a convenience for grouping interdependent elements of a choice set, rather than an attempt to characterize a strategic choice setting or necessarily reflecting a temporally-ordered sequence of choices.

The “nested logit” model is the usual way one of estimating choices over alternatives that can be grouped this way. This involves modifying the standard MNL setup with a parameter tapping any interdependence between options. In this simplest case (above), with three outcomes, two of which can be grouped, a bivariate generalization of the extreme-value distribution sometimes called “Gumbel’s Type B bivariate extreme-value distribution” (Amemiya 1985, 300) is used to model the joint distribution of the (correlated) stochastic components of the utilities associated with the two bus options,

$$F(\varepsilon_1, \varepsilon_2) = \exp \left[-[\exp(-\rho^{-1}\varepsilon_1) + \exp(-\rho^{-1}\varepsilon_2)]^\rho \right], \quad 0 < \rho \leq 1,$$

while the standard extreme-value distribution is used for ε_0 ,

$$F(\varepsilon_0) = \exp[-\exp(-\varepsilon_0)],$$

and I have dropped the i subscript for convenience.

When $\rho = 1$ the MNL model results, with no correlation between the two outcomes, in which case the bivariate distribution above reduces to simply the product of the two univariate extreme-value distributions for ε_0 and ε_1 . In this way ρ does not look like a correlation as we are used to seeing them, since it equals 1 in the case of independence; the actual correlation between the ε_1 and ε_2 is $1 - \rho^2$.

With this joint CDF for $(\varepsilon_1, \varepsilon_2)$,

$$\begin{aligned} P(y_i = 0) &= P(U_0 > U_1, U_0 > U_2), \\ &= P(\mu_0 + \varepsilon_0 > \mu_1 + \varepsilon_1, \mu_0 + \varepsilon_0 > \mu_2 + \varepsilon_2), \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\varepsilon_0 + \mu_0 - \mu_1} \left[\int_{-\infty}^{\varepsilon_0 + \mu_0 - \mu_2} e^{-\varepsilon_0} \cdot \exp[-e^{-\varepsilon_0}] f(\varepsilon_1, \varepsilon_2) d\varepsilon_2 \right] d\varepsilon_1 \right\} d\varepsilon_0, \\ &= \int_{-\infty}^{\infty} \exp(-\varepsilon_0) \exp[-\exp(-\varepsilon_0)] \times \\ &\quad \exp \left(-\{ \exp[-\rho^{-1}(\varepsilon_0 + \mu_0 - \mu_1)] + \exp[-\rho^{-1}(\varepsilon_0 + \mu_0 - \mu_2)] \}^\rho \right) d\varepsilon_0, \\ &= \int_{-\infty}^{\infty} \exp(-\varepsilon_0) \exp[-\alpha \exp(-\varepsilon_0)] d\varepsilon_0, \\ &= \alpha^{-1}, \end{aligned}$$

where $\alpha = 1 + \exp(-\mu_0)[\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)]^\rho$, and μ_j is the right-hand side of the utility functions, $\mathbf{x}_i\boldsymbol{\beta}_j$, again dropping the i subscript through the derivation.

Simple re-arranging yields

$$P(\mathbf{y} = 0) = \frac{\exp(\mu_0)}{\exp(\mu_0) + [\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)]^\rho}.$$

The other probabilities in the model follow from first noting that

$$\begin{aligned} P(y_i = 1 | y_i \neq 0) &= P(U_1 > U_2 | [U_1 > U_0 \text{ or } U_2 > U_0]), \\ &= P(U_1 > U_2), \end{aligned}$$

because of the assumption that ε_1 and ε_2 are correlated, but independent of ε_0 . Expanding this expression yields

$$\begin{aligned}
 P(U_1 > U_2) &= P(\mu_1 + \varepsilon_1 > \mu_2 + \varepsilon_2), \\
 &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\varepsilon_1 + \mu_1 - \mu_2} f(\varepsilon_1, \varepsilon_2) d\varepsilon_2 \right] d\varepsilon_1, \\
 &= \int_{-\infty}^{\infty} \exp(-\varepsilon_1) (1 + \exp[-\rho^{-1}(\mu_1 - \mu_2)])^{\rho-1} \\
 &\quad \times \exp(-\exp(-\varepsilon_1)) \\
 &\quad \times (1 + \exp[-\rho^{-1}(\mu_1 - \mu_2)])^{\rho} d\varepsilon_1, \\
 &= (1 + \exp[-\rho^{-1}(\mu_1 - \mu_2)])^{-1}, \\
 &= (1 + \exp[\rho^{-1}\mu_2 - \rho^{-1}\mu_1])^{-1}, \\
 &= \left(1 + \frac{\exp(\rho^{-1}\mu_2)}{\exp(\rho^{-1}\mu_1)} \right)^{-1}, \\
 &= \left(\frac{\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)}{\exp(\rho^{-1}\mu_1)} \right)^{-1}, \\
 &= \frac{\exp(\rho^{-1}\mu_1)}{\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)},
 \end{aligned}$$

and a symmetric version of this derivation yields an expression for $P(y_i = 2 \mid y_i \neq 0)$.

These probabilities are the key element of the the log-likelihood function for this model

$$\ln \mathcal{L} = \sum_{i=1}^N \sum_{j=0}^m Z_{ij} \ln P_{ij}$$

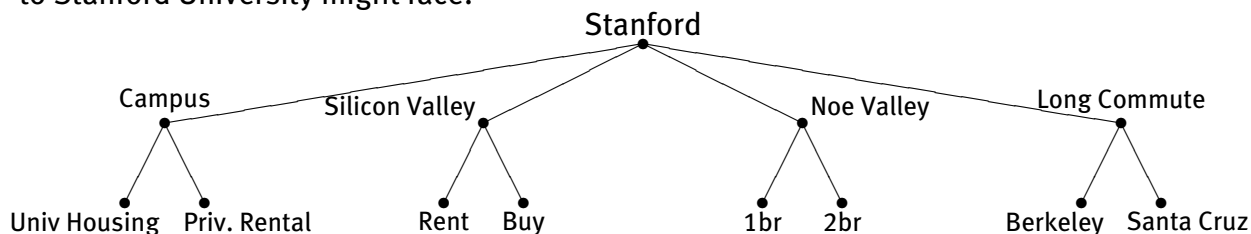
where $Z_{ij} = 1$ if $y_i = j$ and 0 otherwise, and $P_{ij} = P(y_i = j)$. Maximizing this function with respect to the β_j and the ρ parameters yields MLEs of those parameters, with the usual properties. Note that a test of the hypothesis $\rho = 1$ amounts to a test of IIA. One might say that the nested logit model “nests” MNL/IIA within it; MNL is simply a special case of this nested model, when $\rho = 1$.

Political science examples are rare but many come to mind. Consider the 1992 U.S. presidential election again. Are voters’ choices between Clinton and Perot interdependent, with the Bush alternative independent? Or perhaps the Bush-Perot choice is the pair of alternatives that are correlated, with the Clinton option being the distinct alternative? This absence of nested logit setups in the published literature is a little puzzling, since the model seems a relatively straightforward alternative to MNL when IIA is thought to be implausible, but doesn’t require us to go all the way to MNP.²

²Whitten and Palmer say they estimated nested logit models, but do not report the results in their (1996) paper. A quick of the Web turned up the following, from a poster session at last year’s APSA meetings: “Strategic Voting in Multiparty Systems: Estimating Coalition Voting Effects with Multiattributive Nested Logit Models”, by Paul W. Thurner, Universitat Mannheim.

4.1 Generalizing the Nested Logit Model

However the case considered above is quite special in its own right. It is easy to imagine choice structures more complicated than one with three outcomes, with the pairing constituting the “nesting”. As a step towards generalizing this model, consider models with two levels in the choice tree, but with possibly many alternatives (or “branches”) at each level. An example appears below, a stylized representation of the location choice a newcomer to Stanford University might face:³



Several of the options seem to “go together” and can be nested together under the same branch. *Within* each cluster of terminal nodes the IIA hypothesis is worth considering, as it is *among* the clusters. To parameterize these two types of interdependence (within group, between group) we can formalize the scheme in the above tree as follows.

Let $\mathcal{M} = \{\text{“Campus”, “Silicon Valley”, “Noe Valley”, “Long Commute”}\}$ be the set of (unordered) alternatives. Generically, $\mathcal{M} = \{0, \dots, m\}$. The clustering into branches depicted in the tree corresponds to a partition of $\mathcal{M} = B_1 \cup B_2 \cup \dots \cup B_S = \bigcup_{s=1}^S B_s$.

The CDF of the joint distribution of the $\varepsilon_j, j = 0, \dots, m$, is

$$F(\varepsilon_0, \varepsilon_1, \dots, \varepsilon_m) = \exp \left\{ - \sum_{s=1}^S a_s \left[\sum_{j \in B_s} \exp(-\rho^{-1} \varepsilon_j) \right]^{\rho_s} \right\},$$

where $a_s > 0 \forall s$ parameterizes the relative weight of each group or cluster of alternatives, and ρ_s parameterizes the correlation of utilities *within* each group.

Direct MLE of this model is normally not tractable as a two-step procedure. While the log-likelihood is familiar and simple enough,

$$\ln \mathcal{L} = \sum_{i=1}^N \sum_{j=0}^m Z_{ij} \ln P_{ij},$$

the P_{ij} here have a special form, due to the nesting structure:

$$P(y_i = j) = \frac{P(y_i = j | j \in B_s)}{P(j \in B_s)} = \frac{P(y_i = j | j \in B_s)}{\sum_{j \in B_s} P(y_i = j)}.$$

This suggests two steps to estimating a two-level tree (and h steps to estimating a h -level tree).

³Location choice is the application considered in McFadden’s seminal (1978) paper.

The “within” group part of the model can be dealt with first, corresponding to the conditional part of the expression for P_{ij} above. Within each group we have a MNL with the possibility of correlated utilities; implicit here is the assumption that correlations within groups are constant between all alternatives within the group, i.e., $\text{cor}(\varepsilon_p, \varepsilon_q)$ corresponds to $\rho_s, \forall p, q \in B_s$. If this restriction seems unlikely then another partition of \mathcal{M} is warranted, or perhaps another level of nesting (see below). The probability of a specific outcome j conditional on j being in a group B_s so defined is

$$P(y_i = j | j \in B_s) = \frac{\exp(\rho_s^{-1} \mu_{ij})}{\sum_{k \in B_s} \exp(\rho_s^{-1} \mu_{ik})},$$

$s = 1, 2, \dots, S$.

Each of these S MNL problems can be estimated separately to yield estimates of the combination $\rho_s^{-1} \boldsymbol{\beta}$, where here $\mu_{ij} = X'_{ij} \boldsymbol{\beta}$.⁴

These estimates are then used in solving the estimation problem at the “previous” level in the tree; here the problem shifts to the “between” groups part of the two-level tree considered here, and the quantity of interest is

$$\sum_{j \in B_s} P_{ij} = \frac{a_s \left[\sum_{j \in B_s} \exp(\rho_s^{-1} \mu_{ij}) \right]^{\rho_s}}{\sum_{\tau=1}^S a_\tau \left[\sum_{j \in B_\tau} \exp(\rho_\tau^{-1} \mu_{ij}) \right]^{\rho_\tau a_\tau}}.$$

The $\widehat{\rho_s^{-1} \boldsymbol{\beta}}$ from the first stage of the problem can be inserted into this one can find estimates of a_s and ρ_s , which in turn allow one to recover the structural parameters $\boldsymbol{\beta}$ from the reduced form $\widehat{\rho_s^{-1} \boldsymbol{\beta}}$. One of the a_s is set to a constant to identify the model, or the a_s are sometimes assumed constant and ignored altogether.

A generalization to arbitrarily deep choice trees is provided in [McFadden \(1981\)](#), along with details for recovering asymptotically correct standard errors for the model parameters from the two- (or multi-) stage estimation procedure.

4.2 Generalized Extreme-Value Model

The idea behind a generalization to arbitrarily higher levels of nesting is to parameterize dependencies among the choice utilities at each stage of the tree. All terminal nodes below a

⁴Note that this formulation has the parameters constant over alternatives and the data varying (at least in part) by choice alternative. In models of a detailed choice situation it is often the case that some of the regressors measure characteristics of the *choices*, in addition or perhaps even to the exclusion of regressors measuring characteristics of the *choosers*. When the right-hand side variables contain only information about *choices*, or are collected at the level of the choices (“grouped data”, or “choice-based sampling”) then obviously it is impossible to recover a vectors of parameter estimates for each of the $m + 1$ alternatives. The usual formulation in this case is to estimate just one $\boldsymbol{\beta}$ for the whole sample and accordingly the j subscript on $\boldsymbol{\beta}$ is lost, and one introduced for \mathbf{x} . Designs with the two types of regressors seem the more likely route in the social sciences when survey data and contextual data are likely to be available. Depending on the design and the sparseness of the data, parts of $\boldsymbol{\beta}$ may be free to vary across choices, as in the usual MNL setup, while parameters corresponding to information about the choice alternatives can not take on unique values across all alternatives; the exact identification restrictions on choice-specific parameters required will vary from case to case. In any event, a compact notation for this possible mixing of data types is elusive.

given node share some characteristics; parameters specific to each level of branching capture these dependencies. The joint distribution of the ε_j incorporates these parameters, plus the structural parameters. [McFadden \(1981\)](#) provides the following general form for the CDF of the joint density which define nested logit models (of arbitrarily-high levels of nesting):

$$F(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m) = \exp\{-G[\exp(-\varepsilon_1), \exp(-\varepsilon_2), \dots, \exp(-\varepsilon_m)]\},$$

where G satisfies

1. $G(u_1, u_2, \dots, u_m) \geq 0, u_1, u_2, \dots, u_m \geq 0$;
2. $G(\alpha u_1, \alpha u_2, \dots, \alpha u_m) = \alpha G(u_1, u_2, \dots, u_m)$;
3. The mixed partial derivatives of G exist and are continuous, with non-positive even and nonnegative odd mixed partial derivatives.

If the nesting parameterized via G satisfies these criteria then $F(\cdot)$ is a multivariate extreme-value distribution, or a generalized extreme-value distribution.

This formulation is at a high level of generality, but it should be noted that the nested logit models considered here are just special cases of the GEV model. In turn, MNL is a special case of the GEV when the grouping parameters all support IIA and the multivariate GEV reduces to the product of the m univariate extreme-value distributions.

5 Example: Transportation Choice

I now walk through a slightly non-standard application (interleaving some R code), reporting the results of a conditional logit model that appears in [Greene \(2003, §21.7.8\)](#). This is a study of transportation choice in Australia. Two hundred and ten respondents were surveyed about their travel choice between Sydney and Melbourne; the four modes of travel are air, train, bus and car.

R Code

```

1 > #####
2 > ## read in Greene data and do conditional logit
3 > ##
4 > ## travel between Sydney and Melbourne
5 > ## 4 records per observation
6 > ##
7 > ## Mode chosen = Air, Train, Bus, Car (1 if chosen, 0 otherwise)
8 > ## Ttme = terminal waiting time, 0 for car
9 > ## Invc = in vechicle cost
10 > ## Invt = travel time, in vehicle
11 > ## GC = generalized cost (sum of in-vehicle cost)
12 > ## Hinc = household income (fixed over choices)
13 > ## Psize = party size in mode chosen
14 > ##
15 > ## model fitted comes from Greene, _Econometric Analysis_, 5th
16 > ## edition, 21.7.8
17 > #####
18 >
19 > sydney <- read.table(file="sydney.asc",
20 +                       header=TRUE)
21 > attach(sydney)
22 > n <- dim(sydney)[1]/4           ## number of observations

```

```

23 > obs <- rep(1:210,each=4)          ## observation identifier
24 > indx <- rep(1:4,210)             ## record identifier
25 > choice <- rep(1:4,n)
26 > choice <- factor(choice,labels=c("Air","Train","Bus","Car"))
27 > chosen <- Mode==1                ## T if i chose choice j, o'wise F
28 > y <- choice[chosen]
29 > table(y)

```

```

Y
  Air Train  Bus  Car
58   63   30  59

```

Data on the “generalized cost” of each travel option (GC) and terminal waiting time (Ttme) are available at the level of each respondent, as well as a variable measuring household income (Hinc) summarized below (reported quantities are means, with the range appearing in brackets):

	GC	Ttme	Hinc
Air (All Respondents)	102.6 [56 - 197]	61.01 [5 - 99]	34.55 [2 - 72]
Air (Choosing Air, $n=58$)	113.6 [58 - 197]	46.53 [2 - 72]	41.72 [4 - 70]
Train (All Respondents)	130.2 [42 - 269]	35.69 [1 - 99]	34.55 [2 - 72]
Train (Choosing Train, $n=63$)	106.6 [42 - 211]	28.52 [2 - 72]	23.06 [4 - 72]
Bus (All Respondents)	115.3 [45 - 222]	41.66 [5 - 60]	34.55 [2 - 72]
Bus (Choosing Bus, $n=30$)	108.0 [45 - 222]	25.2 [5 - 60]	29.7 [2 - 60]
Car (All Respondents)	95.41 [30 - 238]	0 [0 - 0]	34.55 [2 - 72]
Car (Choosing Car, $n=59$)	89.08 [30 - 238]	0 [0 - 0]	42.22 [4 - 70]

The specified model is

$$U_{ij} = \alpha_j + \beta_{GC} GC_{ij} + \beta_{Ttme} Ttme_{ij} + \gamma_{Hinc} Hinc_i \times \mathcal{I}(j = \text{"air"}) + \varepsilon_{ij}$$

where

- j indexes the choices air, train, bus, car
- $\alpha_{car} = 0$ so as to normalize the unobserved utilities

- $\mathcal{I}(j = \text{"air"})$ is an indicator, set to 1 if j equals “air” and zero otherwise, so that household income only impacts the choice of air over the other options, but has no impact on the choice among car, bus and train
- the ε_{ij} have independent Type-1 extreme value distributions.

With these assumptions and definitions, let

$$\mu_{ij} = \alpha_j + \beta_{GC}GC_{ij} + \beta_{Tme}Ttme_{ij} + \gamma_{Hinc}Hinc_i \times I(j = \text{air}).$$

Then the probability that person i chooses outcome j is

$$p_{ij} = \frac{\exp(\mu_{ij})}{\sum_k \exp(\mu_{ik})}$$

where $k \in \{\text{“air”}, \text{“train”}, \text{“bus”}, \text{“car”}\}$.

We write a function to compute the log-likelihood function, and then pass that function to the optimizer in R, called `optim`.

R Code

```

1 > #####
2 > ## form matrix of predictors, 840 by k
3 > ## to add or drop predictors, add columns to X here
4 > #####
5 > X <- cbind(as.numeric(choic=="Air"),      ## intercept for air
6 +           as.numeric(choic=="Train"),    ## intercept for train
7 +           as.numeric(choic=="Bus"),      ## intercept for bus
8 +           GC,                             ## generalized cost
9 +           Ttme,                           ## terminal time
10 +          Hinc*as.numeric(choic=="Air"))  ## interaction income and air
11 > k <- dim(X)[2]                            ## number of predictors
12 > llhfunc <- function(beta){                ## a function to compute the log-likelihood
13 +   mu <- X%*%beta                          ## X'beta (systematic component of utilities)
14 +   eta <- exp(mu)                          ## exponentiate
15 +   denom <- tapply(eta,obs,sum)           ## sum within each observation (4 per indiv)
16 +   prob <- eta[chosen]/denom              ## probs of the choices actually made
17 +   llh <- sum(log(prob))                  ## sum the log probabilities
18 +   llh                                     ## return this as the log-likelihood
19 + }
20 > beta <- rep(0,k)                          ## start values for maximum likelihood
21 > foo <- optim(par=beta,                    ## starting with parameter values beta
22 +           fn=llhfunc,                   ## optimize the llhfunc function
23 +           control=list(trace=TRUE,
24 +             fnscale=-1),                ## do maximization
25 +           method="BFGS",               ## popular algorithm
26 +           hessian=TRUE)                 ## compute a Hessian matrix

```

```

initial value 291.121816
iter 10 value 199.128576
final value 199.128370
converged

```

R Code

```

1 > se <- sqrt(diag(solve(-foo$hessian))) ## extract standard errors
2 > results <- cbind(foo$par,              ## make pretty results
3 +               se)
4 > results <- cbind(results,results[,1]/results[,2])
5 > dimnames(results) <- list(NULL,
6 +               c("Estimate","Std. Error","z-stat"))
7 > print(signif(results,3))

```

	Estimate	Std. Error	z-stat
[1,]	5.2100	0.77900	6.68
[2,]	3.8700	0.44300	8.73
[3,]	3.1600	0.45000	7.03
[4,]	-0.0155	0.00441	-3.52
[5,]	-0.0961	0.01040	-9.21
[6,]	0.0133	0.01030	1.29

Maximum likelihood estimates of the model parameters and asymptotic standard errors appear in the following table:

	Estimate	Std. Error
α_{air}	5.21	0.78
α_{bus}	3.87	0.44
α_{train}	3.16	0.45
β_{GC}	-0.016	0.004
β_{Ttme}	-0.096	0.010
γ_{Hinc}	0.013	0.010

The negative coefficients on GC and Ttme indicate sensitivity to cost, while the positive coefficient for household income indicates the higher likelihood of flying among wealthier respondents (although this coefficient is not distinguishable from zero at conventional levels of statistical significance). The relatively high constant term for air indicates that respondents derive considerably more utility from flying than from other forms of transport, net of considerations of cost and waiting time. The model correctly predicts 69% of travel choices: 71% of air and bus choices, 77% of train choices and 61% of car choices.

I examine the sensitivity of travel choice to costs by computing predicted probabilities under some hypothetical scenarios. For each mode, I let its generalized cost (GC) vary over its observed range, holding the costs and terminal waiting times (Ttme) of the other travel modes at their observed means, and household income at its mean. The results of this exercise are reported in Figure 2; as costs increase each mode is replaced by car as the more likely outcome, although for car, the most likely substitute is train. Air loses to car as a substitute mode largely due to the fact that car has zero terminal waiting time and the lowest average generalized costs of any mode.

The necessary R code appears here:

```

R Code
1 > ## what happens as cost of each alternative varies (holding other costs constant)
2 > beta <- results[,1]
3 > modelabs <- c("Air", "Train", "Bus", "Car")
4 > par(mfrow=c(2,2), las=1)
5 > cols <- c("black", "orange", "blue", "green")
6 > for (j in 1:4){
7 +   xseq <- seq(from=min(GC[indx==j]), ## loop over modes
8 +               to=max(GC[indx==j]), ## hypothetical values for GC, this mode
9 +               length=100)
10 +   mu <- matrix(0,100,4)
11 +   phat <- matrix(NA,100,4)
12 +   for(z in 1:3){
13 +     mu[,z] <- beta[z] ## intercepts
14 +   }
15 +   mu[,j] <- mu[,j] + beta[4]*xseq ## j-th mode gets hypothetical GC effect
16 +   nonj <- (1:4)[-j]
17 +   for(z in nonj){ ## other modes get their usual (mean) GC effect
18 +     mu[,z] <- mu[,z] + beta[4]*mean(GC[indx==z])

```

```

19 + }
20 + for(z in 1:4){ ## all modes get their (mean) Ttme effect
21 +   mu[,z] <- mu[,z] + beta[5]*mean(Ttme[indx==z])
22 + }
23 + mu[,1] <- mu[,1] + beta[6]*mean(Hinc) ## air gets mean Hinc effect
24 + eta <- exp(mu) ## exponentiate
25 + phat[,1] <- eta[,1]/apply(eta,1,sum) ## convert to probabilities
26 + phat[,2] <- eta[,2]/apply(eta,1,sum)
27 + phat[,3] <- eta[,3]/apply(eta,1,sum)
28 + phat[,4] <- eta[,4]/apply(eta,1,sum)
29 +
30 + plot(xseq, ## plot for this mode
31 +   phat[,j],
32 +   ylim=c(0,.8),
33 +   xlim=range(GC),
34 +   type="l",
35 +   lwd=3,
36 +   xlab=paste("GC of",modelabs[j]),
37 +   ylab="Probability")
38 + title(modelabs[j])
39 + for(z in 1:3){ ## overlay for other modes
40 +   lines(xseq,
41 +     phat[,nonj[z]],
42 +     lty=1,
43 +     col=cols[1+z],
44 +     lwd=2)
45 + }
46 + legend(x="topright", ## legend
47 +   bty="n",
48 +   legend=c(modelabs[j],modelabs[nonj]),
49 +   lty=rep(1,4),
50 +   lwd=c(3,1,1,1),
51 +   cex=.85,
52 +   col=cols)
53 + } ## close loop over modes

```

References

- Alvarez, R. Michael and Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714--44.
- Alvarez, R. Michael and Jonathan Nagler. 1998. "When Politics and Models Collide: Estimating Models of Multiparty Elections." *American Journal of Political Science* 42:55--96.
- Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge: Harvard University Press.
- Greene, William H. 2003. *Econometric Analysis*. Fifth ed. Upper Saddle River, New Jersey: Prentice Hall.
- Hausman, J. 1978. "Specification Tests in Econometrics." *Econometrica* 46:1251--1271.
- Hausman, J. and D. McFadden. 1984. "A Specification Test for the Multinomial Logit Model." *Econometrica* 52.
- Imai, Kosuke and David A. van Dyk. 2005. "A Bayesian Analysis of the Multinomial Probit Model Using the Data Augmentation." *Journal of Econometrics* 124:311--334.

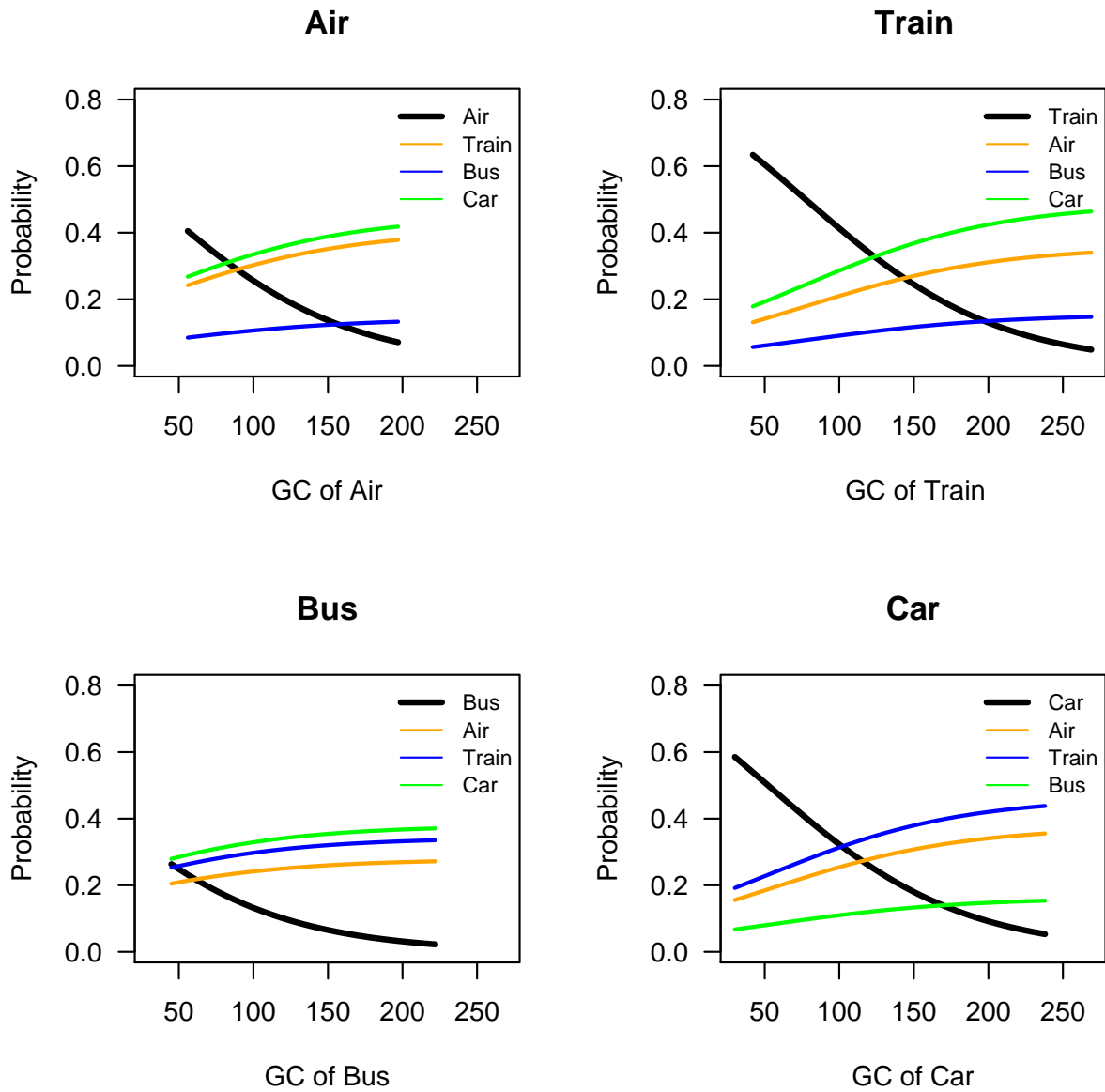


Figure 2: Estimated Sensitivity of Travel Choice to Generalized Cost

- Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44:375--404.
- Keane, Michael P. 1992. "A Note on Identification in the Multinomial Probit Model." *Journal of Economics and Business Statistics* 10:193--200.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Number 7 in *Advanced Quantitative Techniques in the Social Sciences*. Thousand Oaks, California: Sage.
- Maddala, G. S. 1983. *Limited-dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.
- McCulloch, Robert E., Nicholas G. Polson and Peter E. Rossi. 1998. "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters." Typescript. Graduate School of Business, University of Chicago.
- McFadden, Daniel. 1978. "Modeling the Choice of Residential Location." In *Spatial Interaction Theory and Planning Models*, ed. A. Karlqvist. Amsterdam: North-Holland.
- McFadden, Daniel. 1981. "Econometric Models of Probabilistic Choice." In *Structural Analysis of Discrete Data with Econometric Applications*, ed. Charles F. Manski and Daniel McFadden. Cambridge, Mass: MIT Press.
- Powers, Daniel A. and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press.
- Quinn, Kevin M. and Andrew D. Martin. 1998. "Operationalizing and Testing Spatial Theories of Voting." Presented to the meetings of the MidWestern Political Science Association, Chicago, Illinois.
- Whitten, Guy D. and Harvey D. Palmer. 1996. "Heightening Comparativist's Concern for Model Choice: Voting Behavior in Great Britain and the Netherlands." *American Journal of Political Science* 40:231--60.