SPECIFYING ECONOMETRIC MODELS

The target of an econometric analysis is the data generation process (DGP) that maps explanatory variables x into a dependent variable y, with unobserved elements making the mapping stochastic. Write such a mapping as $y = m^*(x,\varepsilon)$, where ε denotes an unobserved effect or state of nature that has a cumulative distribution function $G^*(\varepsilon)$. One might equivalently describe the mapping by the conditional cumulative distribution function $F^*(y|x)$ of y given x. A "*" signifies the functions are not known to the econometrician.

Examples:

y = GDP, x = aggregate resources and other macroeconomic characteristics

y = food expenditures, x = income, family size, education, age

y = indicator for college education, x = ability, gender, parents' income and education

Relationships: Starting from m* and G*, define $\nu = G^*(\varepsilon)$ and $m^{**}(x,\nu) \equiv m^*(x,G^{*-1}(\nu))$. Then, ν has a uniform distribution and m* and G* can always be redefined so that G* is uniform.

- **$F^*(y|x) = G^*(\{\varepsilon|m^*(x,\varepsilon) \le y\})$.**
- **$m^*(x,\nu) = F^{*-1}(\nu|x)$, where $F^{*-1}(\nu|x) = \inf\{y|F^*(y|x) \ge \nu\}$.**
- **Some normalization on $m^*(x,\varepsilon)$ and $G^*(\varepsilon)$, such as ε uniform or some restrictions on the way ε enters m, are necessary for identification, and the normalization can be picked to simplify subsequent analysis; see Matzkin (1999).**

The task of econometrics is to specify models $m(x,\varepsilon)$ and $G(\varepsilon)$, or $F(y|x)$, that approximate the real mappings set by nature. Finding these approximations requires a *specification step,* in which one restricts attention to a class of candidate functions believed to contain the real mappings, or something close, and an *estimation step* that picks out one candidate mapping from the class that by some criterion seems to be closest to the true mapping. The choice of the candidate class will be governed by what is believed to be true about the real mapping from prior research and from economic and statistical theory, and by practical considerations. For example, economic theory may specify what x variables influence y, and justify assuming invariance of $f(y|x)$ under policy interventions. Practical considerations may justify limiting the class of candidate functions to a finite-parameter family, or to a linear regression model, or may only justify limiting the class of candidate functions to those with some mild smoothness and shape properties. How the specification and estimation stages are done depends on how the analysis is to be used, and the approximation accuracy it requires. Typical tasks range from describing empirical features of the mapping, such as the conditional mean $M^*(x) = \int yF^*(dy|x) = \int m^*(x,\varepsilon)G^*(d\varepsilon)$ and other moments or quantiles, to testing economic hypotheses about $F^*(y|x)$, to predicting the conditional distribution of y following policy interventions that alter the distribution of x. The usefulness of the analysis will depend on the quality of the model specification as an approximation to reality, and the validity of assumptions of invariance under policy interventions. The conditional mean M^* is important, but not the whole story of the mapping F^*.

DISCRETE RESPONSE MODELS

When economic behavior is expressed as a continuous variable, a linear regression model is often adequate to describe the impact of economic factors on this behavior, or to predict this behavior in altered circumstances. For example, a study of food expenditures as a function of price indices for commodity groups and income, using households from the Consumer Expenditure Survey, can start by modeling indirect utility as a translog function and from this derive a linear in logs regression equation for food expenditures that does a reasonable job of describing behavior. This situation remains true even when the behavioral response is limited in range (e.g., food consumption of households is non-negative) or integer-valued (e.g., number of times per year eat outside home), provided these departures from a unrestricted continuous variable are not conspicuous in the data (e.g., food consumption is observed over a range where the non-negativity restriction is clearly not binding; the count of meals outside the home is in the hundreds, so that round-off of the dependent variable to an integer is negligible relative to other random elements in the model). However, there are a variety of economic behaviors where the continuous approximation is not a good one.

Examples:

(1) For individuals: Whether to attend college; whether to marry; choice of occupation; number of children; whether to buy a house; what brand of automobile to purchase; whether to migrate, and if so where; where to go on vacation.

(2) For firms: Whether to build a plant, and if so, at what location; what commodities to produce; whether to shut down, merge or acquire other firms; whether to go public or private; whether to accept union demands or take a strike.

For sound econometric analysis, one needs probability models that approximate the true data generation process. To find these, it is necessary to think carefully about the economic behavior, and about the places where random factors enter this behavior. For simplicity, we initially concentrate on a single binomial (Yes/No) response. An example illustrates the process:

> **Yellowstone National Park has been overcrowded in recent years, and large user fees to control demand are under consideration. The National Park Service would like to know the elasticity of demand with respect to user fees, and the impact of a specified fee increase on the total number of visitors and on the visitors by income bracket. The results of a large household survey are available giving household characteristics (income, number of children, etc.), choice of vacation site, and times and costs associated with vacations at alternative sites. Each vacation is treated as an observation.**

Start with the assumption that households are utility maximizers. Then, each household will have an indirect utility function, *conditioned* on vacation site, that gives the payoff to choosing this particular site and then optimizing consumption in light of this choice. This indirect utility function will depend on commodity prices and on household income net of expenditures mandated by the vacation site choice. It may also contain factors such as household tastes and perceptions, and unmeasured attributes of sites, that are, from the standpoint of the analyst, random. (Some of what appears to be random to the analyst may just be heterogeneity in tastes and perceptions over the population.)

Now consider the *difference* between the indirect utility of a Yellowstone vacation and the *maximum* indirect utilities of alternative uses of leisure. This is a function $y^* = f(z,\zeta)$ of observed variables z and unobserved variables ζ. We put a "*" on the utility difference y to indicate that is *latent* rather than observed directly. Included in z are variables such as household income, wage rate, family characteristics, travel time and cost to Yellowstone, and so forth. The form of this function will be governed by the nature of indirect utility functions and the sources of ζ. In some applications, it makes sense to parameterize the initial indirect utility functions tightly, and then take *f* to be the function implied by this. Often, it is more convenient to take *f* to be a form that is flexibly parameterized and convenient for analysis, subject only to the generic properties that a difference of indirect utility functions should have. In particular, it is almost always possible to approximate *f* closely by a function that is linear in parameters, with an additive disturbance: $f(z,\zeta) \approx x\beta - \varepsilon$, where β is a k×1 vector of unknown parameters, x is a 1×k vector of transformations of z, and $\varepsilon = -f(z,\zeta) + Ef(z,\zeta)$ is the deviation of *f* from its expected value in the population. Such an approximation might come, for example, from a Taylor's expansion of E*f* in powers of (transformed) observed variables z.

Suppose the gain in utility from vacationing in Yellowstone rather than at an alternative site is indeed given by $y^* = x\beta - \varepsilon$. Suppose the disturbance ε is known to the household and unknown to the econometrician, but the cumulative distribution function (CDF) of ε is a function $F(\varepsilon)$ that is known up to a finite parameter vector. The utility-maximizing household will then choose Yellowstone if $y^* > 0$, or $\varepsilon < x\beta$. The probability that this occurs, given x, is

$$P(\varepsilon < x\beta) = F(x\beta).$$

Define y = 1 if Yellowstone is chosen, y = -1 otherwise; then, y is an (observed) indicator for the event $y^* > 0$. The probability law governing observed behavior is then, in summary,

$$P(y|x\beta) = \begin{cases} F(x\beta) & if\, y = 1 \\ 1 - F(x\beta) & if\, y = -1 \end{cases} .$$

Assume that the distribution of ε is symmetric about zero, so that $F(\varepsilon) = 1 - F(-\varepsilon)$; this is not essential, but it simplifies notation. The probability law then has an even more compact form,

$$P(y|x\beta) = F(yx\beta) .$$

How can you estimate the parameters β? An obvious approach is maximum likelihood. The log likelihood of an observation is

$$\boldsymbol{l(\beta | y,x) = \log P(y | x\beta) \equiv \log F(yx\beta)\ .}$$

If you have a random sample with observations t = 1,...,T, then the sample log likelihood is

$$\mathbf{L_T(\beta) = \sum_{t=1}^{T} \log F(y_t x_t \beta)\ .}$$

The associated score and hessian of the log likelihood are

$$\mathbf{\nabla_\beta L_T(\beta) = \sum_{t=1}^{T} y_t x_t{}' F'(y_t x_t \beta)/F(y_t x_t \beta)}$$

$$\mathbf{\nabla_{\beta\beta} L_T(\beta) = \sum_{t=1}^{T} x_t{}' x_t \{F''(y_t x_t \beta)/F(y_t x_t \beta) - [F'(y_t x_t \beta)/F(y_t x_t \beta)]^2\}}$$

A maximum likelihood program will either ask you to provide these formula, or will calculate them for you analytically or numerically. If the program converges, then it will then find a value of β (and other parameters upon which F depends) that are (at least) a local maximum of L_T. It can fail to converge to a global maximum if no maximum exists or if there are numerical problems in the evaluation of expressions or in the iterative optimization. The estimates obtained at convergence will have the usual large-sample properties of MLE, provided the usual regularity conditions are met, as discussed later.

It is sometimes useful to write the score and hessian in a slightly different way. Let d = (y+1)/2; then d = 1 for Yellowstone, d = 0 otherwise, and d is an indicator for a Yellowstone trip. Then, we can write

$$\boldsymbol{l(y|x,\beta) = d\cdot\log F(x\beta) + (1-d)\cdot\log F(-x\beta).}$$

Differentiate, and noting that $F'(x\beta) = F'(-x\beta)$, to get

$$\boldsymbol{\nabla_\beta l = xF'(x\beta)\{d/F(x\beta) - (1-d)/F(-x\beta)\} = w(x\beta)\cdot x\cdot[d - F(x\beta)],}$$

where $w(x\beta) = F'(x\beta)/F(x\beta)F(-x\beta)$. The sample score is then

$$\boldsymbol{\nabla_\beta L_T(\beta) = \sum_{t=1}^{T} w(x_t\beta)\cdot x_t{}'\cdot[d_t - F(x_t\beta)]\ .}$$

The MLE condition that the sample score equal zero can be interpreted as a weighted *orthogonality condition* between a residual [d - F(xβ)] and the explanatory variables x. Put another way, a weighted non-linear least squares (NLLS) regression $d_t = F(x_t\beta) + \eta_t$, with observation t weighted by $w(x_t\beta)^{1/2}$, will be equivalent to MLE.

The hessian can also be rewritten using d rather than y: $\nabla_{\beta\beta}l$ = -x′x·s(xβ), where

$$\mathbf{s(x\beta) =} \quad \frac{F'(x\beta)^2}{F(x\beta)F(-x\beta)}$$

$$\mathbf{- [d - F(x\beta)]} \left\{ \frac{F''(x\beta)}{F(x\beta)F(-x\beta)} - \frac{F'(x\beta)^2(1-2F(x\beta))}{F(x\beta)^2F(-x\beta)^2} \right\} .$$

The expectation of s(xβ) at the true β_o is $\dfrac{F'(x\beta_o)^2}{F(x\beta_o)F(-x\beta_o)}$ **> 0, so that the sample sum of the hessians of the observations in sufficiently large samples is eventually almost surely negative definite in a neighborhood of β_o.**

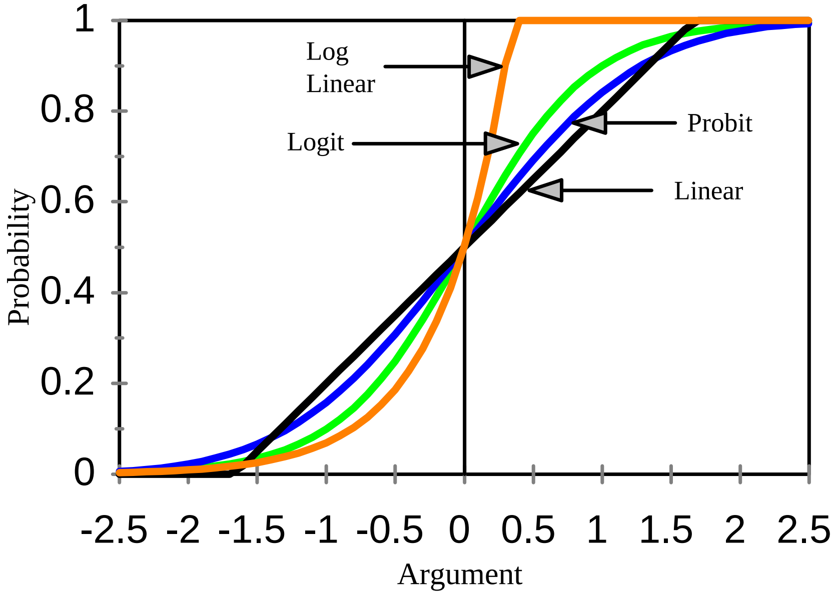
It should be clear from the sample score, or the analogous NLLS regression, that the distribution function F enters the likelihood function in an intrinsic way. Unlike linear regression, there is no simple estimator of β that rests only on assumptions about the first two moments of the disturbance distribution.

FUNCTIONAL FORMS AND ESTIMATORS

In principle, the CDF F(ε) will have a form deduced from the application. Often, F would naturally be conditioned on the observed explanatory variables. However, an almost universal practice is to assume that F(ε) has one of the following standard distributions that are not conditioned on x:

(1) *Probit*: F is standard normal.
(2) *Logit*: $F(\varepsilon) = 1/(1+e^{-\varepsilon})$, the standard logistic CDF.
(3) *Linear*: $F(\varepsilon) = \varepsilon$, for $0 \le \varepsilon \le 1$, the standard uniform distribution.
(4) *Log-Linear*: $F(\varepsilon) = e^{\varepsilon}$, for $\varepsilon \le 0$, a standard exponential CDF.

Binomial Response Models



There are many canned computer programs to fit models (1) or (2). Model (3) can be fit by linear regression, although heteroscedasticity is an issue. Model (4) is not usually a canned program when one is dealing with individual observations, but for repeated observations at each configuration of x it is a special case of the *discrete analysis of variance* model that is widely used in biostatistics and can be fitted using ANOVA or regression methods. Each of the distributions above has the property that the function s(xβ) that appears in the hessian is globally positive, so that the log likelihood function is globally concave. This is convenient in that any local maximum is the global maximum, and any stable hill-climbing algorithm will always get to the global maximum. The linear and log-linear distributions are limited in range. This is typically not a problem if the range of x is such that the probabilities are bounded well away from zero and one, but can be a serious problem when some probabilities are near or at the extremes, particularly when the model is used for forecasting.

ALTERNATIVES TO MLE

Recall that MLE chooses the parameter vector β to achieve orthogonality between the explanatory variables x, and residuals d - $F(x\beta)$, with weights $w(x\beta)$. When the explanatory variables are grouped, or for other reasons there are multiple responses observed for the same x, there is another estimation procedure that is useful. Let j = 1,...,J index the possible x configurations, m_j denote the number of responses observed at configuration x_j, and s_j denote the number of "successes" among these responses (i.e., the number with d = 1). Let $p_j = F(x_j\beta_o)$ denote the true probability of a success at configuration x_j. Invert the CDF to obtain $c_j = F^{-1}(p_j) = x_j\beta$. Note that $p = F(c)$ implies $\partial c/\partial p = 1/F'(c)$ and $\partial^2 c/\partial p^2 = - F''(c)/F'(c)^3$.

A Taylor's expansion of $F^{-1}(s_j/m_j)$ about p_j gives

$$F^{-1}\left(\frac{s_j}{m_j}\right) = F^{-1}(p_j) + \frac{s_j/m_j - p_j}{F'(F^{-1}(p_j))} - \frac{(s_j/m_j - p_j)^2}{2}\cdot\frac{F''(F^{-1}(q_j))}{F'(F^{-1}(q_j))^3}$$

$$= x_j\beta + \nu_j + \xi_j\,,$$

where q_j is a point between p_j and s_j/m_j,

$$\nu_j = (s_j/m_j - p_j)/F'(F^{-1}(p_j))$$

is a disturbance that has expectation zero and a variance proportional to $p_j(1-p_j)/m_j$, and ξ_j is a disturbance that goes to zero in probability relative to ν_j. Then, when the m_j are all large (the rule-of-thumb is $s_j \geq 5$ and m_j-$s_j \geq 5$), the regression

$$F^{-1}(s_j/m_j) = x_j\beta + \nu_j$$

gives consistent estimates of β. This is called *Berkson's method*. It can be made asymptotically equivalent to MLE if a FGLS transformation for heteroscedasticity is made. Note however that in general this transformation is not even defined unless s_j is bounded away from zero and m_j, so it does not work well when some x's are continuous and cell counts are small.

Berkson's transformation in the case of probit is $\Phi^{-1}(s_j/m_j)$; in the case of logit is $\log(s_j/(m_j-s_j))$; in the case of linear is s_j; and in the case of the exponential model is $\log(s_j/m_j)$. It is a fairly general proposition that the asymptotic approximation is improved by using the transformation $F^{-1}((s_j+0.5)/(m_j+1))$ rather than $F^{-1}(s_j/m_j)$ as the dependent variable in the regression; for logit, this minimizes the variance of the second-order error.

There is an interesting connection between the logit model and a technique called *normal linear discriminant analysis*. Suppose that the conditional distributions of x, given d = 1 or given d = 0, are both multivariate normal with respective mean vectors μ_1 and μ_0, and a *common* covariance matrix Ω. Note that these assumptions are not necessarily very plausible, certainly not if some of the x variables are limited or discrete. If the assumptions hold, then the means μ_0 and μ_1 and the covariance matrix Ω can be estimated from sample averages, and by Bayes law the conditional distribution of d given x when a proportion q_1 of the population has state d = 1 has a logit form

$$\mathbf{P(d=1|x)} = \frac{q_1 n(x-\mu_1,\Omega)}{q_0 n(x-\mu_0,\Omega) + q_1 n(x-\mu_1,\Omega)}$$

$$= \frac{1}{1 + \exp(-\alpha - x\beta)} \;,$$

where $\beta = \Omega^{-1}(\mu_1-\mu_0)$ and $\alpha = \mu_1{}'\Omega^{-1}\mu_1 - \mu_0{}'\Omega^{-1}\mu_0 + \log(q_1/q_0)$. This approach produces a fairly robust (although perhaps inconsistent) estimator of the logit parameters, even when the normality assumptions are obviously wrong.

3. STATISTICAL PROPERTIES OF MLE

The MLE estimator for most binomial response models is a special case of the general setup treated in the statistical theory of MLE, so that the incantation "consistent and asymptotically normal (CAN) under standard regularity conditions" is true. This is a simple enough application so that it is fairly straightforward to see what these "regularity" conditions mean, and verify that they are satisfied. This is a thought exercise worth going through whenever you are applying the maximum likelihood method. First, here is a list of fairly general sufficient conditions for MLE to be CAN in discrete response models; these are taken from McFadden "Quantal Response Models", Handbook of Econometrics, Vol. 2, p. 1407. Commentaries on the assumptions are given in italics.

(1) The support of the explanatory variables is a closed set X with a measurable probability p(x). ***This just means that the explanatory variables have a well-defined distribution. It certainly holds if p is a continuous density on closed*** **X.**

(2) The parameter space is a subset of $\mathbb{R}^k$, and the true parameter vector is in the interior of this space. ***This says you have a finite-dimensional parametric problem. This assumption does*** **not** ***require that the parameter space be bounded, in contrast to many sets of assumptions used to conclude that MLE are CAN. The restriction that the true parameter vector be in the interior excludes some cases where CAN breaks down. This is not a restrictive assumption in most applications, but it is for some. For example, suppose a parameter in the probit model is restricted (by economic theory) to be non-negative, and that this parameter is in truth zero. Then, its asymptotic distribution will be the (non-normal) mixture of a half-normal and a point mass.***

(3) The response model is measurable in x, and for almost all x is continuous in the parameters. ***The standard models such as probit, logit, and the linear probability model are all continuous in their argument*** **and** ***in*** **x*****, so that the assumption holds. Only "pathological" applications in which a parameter determines a "trigger level" will violate this assumption.***

(4) The model satisfies a global identification condition (that guarantees that there is at most one global maximum; see McFadden, *ibid*, p. 1407). ***The concavity of the log likelihood of an observation for probit, logit, linear, and log linear models guarantees global identification, requiring only that the x's are not linearly dependent.***

(5) The model is once differentiable in the parameters in some neighborhood of the true values. ***This is satisfied by the four CDF from Section 2 (provided parameters do not give observations on the boundary in the linear or log linear models where probabilities are zero or one), and by most applications. This assumption is weaker than most general MLE theorems, which assume the log likelihood is twice or three times continuously differentiable.***

(6) The log likelihood and its derivative have bounds independent of the parameters in some neighborhood of the true parameter values. The first derivative has a Lipschitz property in this neighborhood. ***This property is satisfied by the four CDF, and any CDF that are continuously differentiable.***

(7) The information matrix, equal to the expectation of the outer product of the score of an observation, is nonsingular at the true parameters. ***This is satisfied automatically by the four CDF in Section 2, provided the*** **x's** ***are not linearly dependent.***

The result that conditions (1)-(7) guarantee that MLE estimates of β are CAN is carried out essentially by linearizing the first-order condition for the estimator using a Taylor's expansion, and arguing that higher-order terms than the linear term are asymptotically negligible. With lots of differentiability and uniform bounds, this is an easy argument. A few extra tricks are needed to carry this argument through under the weaker smoothness conditions contained in (1)-(7).

4. EXTENSIONS OF MAXIMUM LIKELIHOOD PRINCIPLE

The assumptions under which the maximum likelihood criterion produces CAN estimates include, critically, the condition (2) that the parametric family of likelihoods that are being maximized include the true data generation process. There are several reasons that this assumption can fail. First, you may have been mistaken in your assumption that the model you have written down includes the truth. This might happen in regression analysis because some variable that you think does not influence the dependent variable or is uncorrelated with the included variables actually does belong in the regression. Or, in modeling a binomial discrete response, you may assume that the disturbance in the model $y^* = x\beta - \varepsilon$ is standard normal when it is in truth logistic. Second, you may deliberately write down a model you suspect is incorrect, simply because it is convenient for computation or reduces data collection problems.

For example, you might write down a model that assumes observations are independent even though you suspect they are not. This might happen in discrete response analysis where you observe several responses from each economic agent, and suspect there are unobserved factors such as tastes that influence all the responses of this agent.

What are the statistical consequences of this model misspecification? The answer is that this will generally cause the CAN property to fail, but in some cases the failure is less disastrous than one might think. The most benign situation arises when you write down a likelihood function that fails to use all the available data in the most efficient way, but is otherwise consistent with the true likelihood function. For example, if you have several dependent variables, such as binomial responses on different dates, you may write down a model that correctly characterizes the marginal likelihood of each response, but fails to characterize the dependence between the responses. This setup is called *quasi-maximum likelihood* estimation. What may happen in this situation is that not all the parameters in the model will be identified, but those that are identified are estimated CAN, although not necessarily with maximum efficiency. In the example, it will be parameters characterizing the correlations across responses that are not identified.

Also fairly benign is a method called *pseudo-maximum likelihood* estimation, where you write down a likelihood function with the property that the resulting maximum likelihood estimates are in fact functions only of selected moments of the data. A classic example is the normal regression model, where the maximum likelihood estimates depend only on first and second moments of the data. Then the estimates that come out of this criterion will be CAN even if the pseudo-likelihood function is misspecified, so long as the true likelihood function and the pseudo-likelihood function coincide for the moments that the estimators actually use.

More tricky is the situation where the likelihood you write down is not consistent with the true likelihood function. In this case, the parameters in the model you estimate will not necessarily match up, even in dimension, with the parameters of the true model, and there is no real hope that you will get reasonable estimates of these true parameters. However, even here there is an interesting result. Under quite general conditions, it is possible to talk about the "*asymptotically least misspecified model*", defined as the model in your misspecified family that asymptotically has the highest log likelihood. To set notation, suppose $f(y|x)$ is the true data generation process, and $g(y|x,\beta)$ is the family of misspecified models you consider.

Define β_1 to be the parameters that maximize

$$\mathbf{E}_{y,x}\, f(y|x)\cdot\log g(y|x,\beta).$$

Then, β_1 determines the least misspecified model. While β_1 does not characterize the true data generation process, and the parameters as such may even be misleading in describing this process, what is true is that β_1 characterizes the model *g* that in a "likelihood metric" is as close an approximation as one can reach to the true data generation process when one restricts the analysis to the *g* family. Now, what is interesting is that the maximum likelihood estimates *b* from the misspecified model are CAN for β_1 under mild regularity conditions. A colloquial way of putting this is that MLE estimates are usually CAN for whatever it is they converge to in probability, even if the likelihood function is misspecified.

All of the estimation procedures just described, quasi-likelihood maximization, pseudo-likelihood maximization, and maximization of a misspecified likelihood function, can be interpreted as special cases of a general class of estimators called *generalized method of moment estimators*. One of the important features of these estimators is that they have asymptotic covariance matrices of the form $\Gamma^{-1}\Sigma\Gamma'^{-1}$, where Γ comes from the hessian of the criterion function, and Σ comes from the expectation of the outer product of the gradient of the criterion function. For true maximum likelihood estimation, this form reduces to Σ^{-1}, but more generally the full form $\Gamma^{-1}\Sigma\Gamma'^{-1}$ is required.

One important family of quasi-maximum likelihood estimators arises when an application has a likelihood function in two sub-vectors of parameters, and it is convenient to obtain preliminary CAN estimates of one sub-vector, perhaps by maximizing a conditional likelihood function. Then, the likelihood is maximized in the second sub-vector of parameters after plugging in the preliminary estimates of the first sub-vector. This will be a CAN procedure under general conditions, but it is necessary to use a formula of the form $\Gamma^{-1}\Sigma\Gamma'^{-1}$ for its asymptotic covariance matrix, where Σ includes a contribution from the variance in the preliminary estimates of the first sub-vector. The exact formulas and estimators for the terms in the covariance matrix are given in the lecture notes on generalized method of moments.

5. TESTING HYPOTHESES

It is useful to see how the general theory of large sample hypothesis testing plays out in the discrete response application. For motivation, return to the example of travel to Yellowstone Park. The basic model might be binomial logit,

$$\mathbf{P(y|x\beta) = F(yx\beta) = 1/(1 + exp(-yx\beta)),}$$

where x includes travel time and travel cost to Yellowstone, and family income, all appearing linearly:

$$\mathbf{x\beta = TT{\cdot}\beta_1 + TC{\cdot}\beta_2 + I{\cdot}\beta_3 + \beta_4,}$$

with TT = travel time, TC = travel cost, I = income. The parameter β_4 is an intercept term that captures the "average" desirability of Yellowstone relative to alternatives after travel factors have been taken into account. The Park Service is particularly concerned that an increase in Park entry fees, which would increase overall travel cost, will have a particularly adverse effect on low income families, and asks you to test the hypothesis that sensitivity to travel cost increases as income falls. This suggests the alternative model

$$\mathbf{x\beta = TT{\cdot}\beta_1 + TC{\cdot}\beta_2 + I{\cdot}\beta_3 + \beta_4 + \beta_5{\cdot}TC/I,}$$

with the null hypothesis that $\beta_5 = 0$.

This hypothesis can be tested by estimating the model without the null hypothesis imposed, so that β_5 is estimated. The Wald test statistic is the quadratic form (b_5 - 0)$'$V(b_5)$^{-1}$(b_5 - 0); it is just the square of the T-statistic for this one-dimensional hypothesis, and it is asymptotically chi-square distributed with one degree of freedom when the null hypothesis is true. When the null hypothesis is non-linear or of higher dimension, the Wald statistic requires retrieving the covariance matrix of the unrestricted estimators, and forming the matrix of derivatives of the constraint functions evaluated at *b*. An alternative that is computationally easier when both the unrestricted and restricted models are easy to estimate is to form the Likelihood Ratio statistic 2[L_T(*b*) - L_T(*b)], where *b* and *b** are the estimates obtained without the null hypothesis and with the null hypothesis imposed, respectively, and L_T is the sample log likelihood. This statistic is asymptotically equivalent to the Wald statistic. Finally, the Lagrange Multiplier statistic is obtained by estimating the model under the null hypothesis, evaluating the score of the unrestricted model at the restricted estimates, and then testing whether this score is zero. In our example, there is a slick way to do this. Regress a normalized residual [d_t - F(x_t*b*)]/[F(x*b*)F(-x *b*)]$^{1/2}$ from the restricted model on the weighted explanatory variables xF$'$(x*b*)/[F(x*b*)F(-x*b*)]$^{1/2}$. that appear in the unrestricted model. The F-test for the significance of the explanatory variables in this regression is asymptotically equivalent to the Lagrange Multiplier test. The reason this trick works is that the Lagrange Multiplier test is a test of orthogonality between the normalized residual and the weighted variables in the unrestricted model.**

6. MULTINOMIAL RESPONSE

Conceptually, it is straightforward to move from modeling binomial response to modeling multinomial response. When consumers or firms choose among multiple, mutually exclusive alternatives, such as choice of brand of automobile, occupation, or plant location, it is natural to introduce the economic agent's objective function (utility for consumers, profit for firms), and assume that choice maximizes this objective function. Factors unobserved by the analyst, particularly heterogeneity in tastes or opportunities, can be interpreted as random components in the objective functions, and choice probabilities derived as the probabilities that these unobserved factors are configured so as to make the respective alternatives optimal.

Suppose there are J alternatives, indexed C = {1,...,J}, and suppose the economic agent seeks to maximize an objective function $U(z_i,s,\nu_i)$, where z_i are observed attributes of alternative i, s are characteristics of the decision maker, and ν_i summarizes all the unobserved factors that influence the attractiveness of alternative i.

Then, the multinomial response probability is

$$P_C(i|z,s) = \mathrm{Prob}(\{\nu|U(z_i,s,\nu_i) > U(z_j,s,\nu_j) \text{ for } j \neq i\}),$$

where $z = (z_1,...,z_J)$. For example, if $C = \{1,...,J\}$ is the set of automobile brands, with z_i the attributes of brand i including price, size, horsepower, fuel efficiency, etc., then this model can be used to explain brand choice, or to predict the shares of brands as the result of changing prices or new model introductions. If one of the alternatives in C is the "no purchase" alternative, the model can describe the demand for cars as well as brand choice. If C includes both new and used alternatives, then it can explain replacement behavior. If $i \in C$ identifies a portfolio of two brands, or one brand plus a "no purchase", it can explain the holdings of two-car families.

Placing U in a parametric family and making ν a random vector with a parametric probability distribution produces a parametric probability law for the observations. However, it is difficult to do this in a way that leads to simple algebraic forms that do not require multivariate integration. Consequently, the development of multinomial response models has tended to be controlled by computational issues, which may not accommodate some features that might seem sensible given the economic application, such as correlation of unobservables across alternative portfolios that have common elements.

■ For notational shorthand, associate with alternative i in a feasible set C a "payoff" $u_i = z_i\beta + \varepsilon_i$, which in the case of consumer choice may be the indirect utility attached to alternative i and in the case of firm choice may be profit from alternative i. The z_i are observed explanatory variables, and the ε_i are unobserved disturbances. Observed choice is assumed to maximize payoff: $y_i = 1(u_i \geq u_j \text{ for } j \in C)$.

■ One form of this model is a <u>random</u> <u>coefficients</u> formulation $u_i = z_i\alpha$, $E\alpha = \beta$, $\varepsilon_i = z_i(\alpha - \beta)$, implying $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = z_i{\cdot}\mathrm{Cov}(\alpha){\cdot}z_j'$. ■ For C = {1,...,J}, define u, z, ε, and y to be J×1 vectors with components u_j, z_j, ε_j, y_j, respectively. Define a (J-1)×J matrix Δ_i by starting from the J×J identity matrix, deleting row i, and then replacing column i with the vector (-1,...,-1). For example if J = 4,

$$\Delta_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} .$$

Then alternative i is chosen if $\Delta_i u \leq 0$. The probability of this event is

$$P_i(z,\theta) = \Pr(\Delta_i u \leq 0 | z,\theta) \equiv \int_{\Delta_i u \leq 0} f(u|z,\theta)du,$$

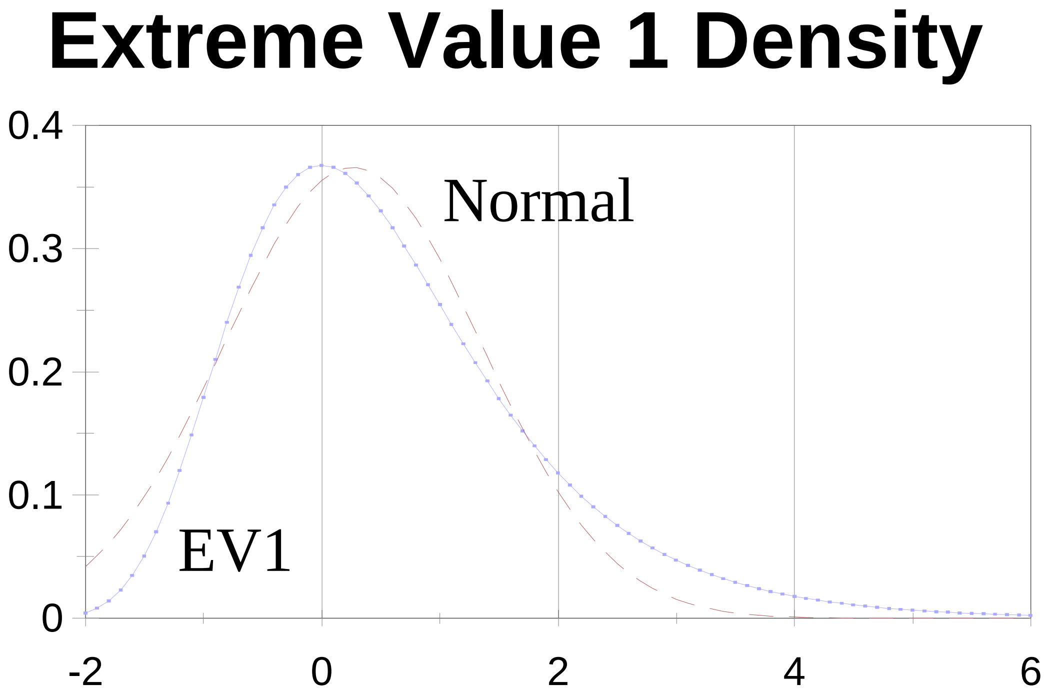
where $f(u|z,\theta)$ is the conditional density of u given z. The parameters θ include the slope parameters β and any additional parameters characterizing the distribution of the disturbances ε.

MULTINOMIAL LOGIT

A special case where $P_i(z,\theta)$ has a simple closed form is the multinomial logit (MNL) model

$$P_i(z,\beta) = \exp(v_i)/\textstyle\sum_{j\in C} \exp(v_j)$$

with $v_i = z_i\beta$. This is derived from $u_i = z_i\beta + \varepsilon_i$ with the disturbances ε_i being independently, identically distributed with a distribution called Extreme Value Type 1, which has $\mathrm{Prob}(\varepsilon_i \le c) \equiv F(c) = \exp(-e^{-c})$ and density $f(c) = e^{-c}\cdot\exp(-e^{-c})$. This distribution is bell-shaped and skewed to the right; the density is plotted in the figure below.



Derivation:

$$P_i(z,\beta) = \text{Prob}(v_i+\varepsilon_i \geq v_j+\varepsilon_j \text{ for } j \in C)$$

$$= \int_{\Delta_i\varepsilon \leq -\Delta_i v} f(\varepsilon_1)\cdot\ldots\cdot f(\varepsilon_J)d\varepsilon_1\ldots d\varepsilon_J$$

$$= \int_{\varepsilon_i=-\infty}^{+\infty} f(\varepsilon_i)\{\textstyle\prod_{j\neq i} F(\varepsilon_i+v_i-v_j)\}d\varepsilon_i$$

$$= \int_{c=-\infty}^{+\infty} e^{-c}\ \{\textstyle\prod_j \exp(-\exp(-c-v_{i+}v_j))\}dc$$

$$= \int_{c=-\infty}^{+\infty} e^{-c}\cdot\exp(-\textstyle\sum_j \exp(-c-v_{i+}v_j))dc$$

$$= A \int_{c=-\infty}^{+\infty} A^{-1}\ e^{-c}\cdot\exp(-e^{-c}/A)dc$$

$$= A[F(\infty/A) - F(-\infty/A)] = A$$

with $A^{-1} = \sum_j \exp(-v_{i+}v_j)$, or $A = \exp(v_i)/\sum_{j\in C} \exp(v_j)$.

The reason that the somewhat unusual EV1 distribution is linked to the closed MNL formula is that the EV1 family is closed under the operation of maximization. (Compare with the normal family, which is closed under the operation of addition.)

The MNL discrete response probabilities with suitably articulated v_i's are often reasonable approximations to true response probabilities, for essentially the same reason that linear regression models are often reasonable approximations a true data generating process – the coefficients can compensate to some degree for failures of the specification. However, when v_i is specified to be a function solely of attributes z_i of alternative i, the MNL model satisfies a very powerful and very restrictive property called *Independence from Irrelevant Alternatives* (IIA). This property says that the relative probabilities of responses i and j depend only on v_i and v_j, and not on the attractiveness, or even the presence or absence, of additional alternatives. From the MNL formula,

$$P_i(z,\beta)/P_j(z,\beta) = \exp((z_i-z_j)\beta) \equiv \exp(v_i)/\exp(v_j).$$

When the true responses l satisfy IIA, this is extremely useful. One can predict multinomial response by studying binomial responses, predict responses when new alternatives are added, and analyze responses as if the set of feasible alternatives were a proper subset of the true choice set (next lecture). However, when the true responses do not satisfy IIA, predictions from a MNL approximation can be very misleading.

Red Bus / Blue Bus Problem

Two alternatives originally, car and blue bus, with $v_c = z_c\beta = z_{bb}\beta = v_{bb}$. Then the MNL choice probabilities are $P_c = P_{bb} = 1/2$.

Suppose a third alternative is added, a red bus that is identical or nearly identical to the blue bus except for color (which does not matter to the consumer), so that $z_{rb} \approx z_{bb}$. What one expects in reality is that if $v_{rb} = v_{bb}$, then those who preferred car to bb will prefer car to both rb and bb, whereas those who preferred bb to car will continue to choose bus, and divide their patronage evenly between rb and bb, leading to $P_c = 1/2$, $P_{rb} = P_{bb} = 1/4$. Further, if $v_{rb} > v_{bb}$ and the two buses are nearly identical in terms of unobserved characteristics, one expects in reality that $P_c \approx 1/2$, $P_{rb} \approx 1/2$, and $P_{bb} \approx 0$.

The MNL model estimated using data on choices between c and bb will have the IIA property, and will predict that $P_{bb}/P_c = 1$when the rb is added, just as when it was absent. Thus, when $v_c = v_{bb} = v_{rb}$, the MNL model predicts $P_c = P_{rb} = P_{bb} = 1/3$. This contradicts reality, where rb gets its patronage solely from previous bb users.

Comparing the sensitivity of P_{rb}/P_{bb} and P_c/P_{bb} to service attributes, the former ratio is much more sensitive because differences in unobserved attributes are unimportant, while the latter ratio is much less sensitive because differences in unobserved attributes (ε's) are important and will induce many decision-makers to stay with their initial choice even when there is some variation in observed attributes. The validity of IIA in an application is an empirical question. The elevated sensitivity to observed attributes between alternatives that are similar in unobserved attributes, compared to alternatives with independent unobserved attributes, can be used as a basis for a test of the validity of the IIA property.

Tests of IIA

1. **Hausman-McFadden Test**
   Intuition: If IIA holds, then one should get approximately the same parameter estimates using the full choice set or using the observations that fall in a subset.

2. **McFadden omitted variables tests**
   Intuition: Failures of IIA are usually associated with sharper discrimination within a subset of alternatives than otherwise. This can be detected by the coefficients on added variables that are zero for alternatives outside the subset. Chosen appropriately, these additional variables give tests that are equivalent to the Hausman-McFadden test, or to a test against the *nested* MNL model.

1. Hausman-McFadden Test on a subset of alternatives.

- **Estimate logit model twice:**
  a. **on full set of alternatives**
  b. **on a specified subset of alternatives (and the subsample with choices from this subset)**

- **If IIA holds, the two sets of estimates should not be statistically different: Let β_b denote the estimates obtained from setup b. above, and Ω_b denote their estimated covariance matrix. Let β_a denote the estimates of the same parameters obtained from setup a. above, and Ω_a denote their estimated covariance matrix. (Some parameters that can be estimated in setup a. may not be identified in setup b, in which case β_a refers to estimates under setup a. of the subvector of parameters that are identified in both setups.) Then, the quadratic form**

$$(\beta_a - \beta_b)'(\Omega_b - \Omega_a)^{-1}(\beta_a - \beta_b)$$

has a chi-square distribution when IIA is true. In calculating this test, one must be careful to restrict the comparison of parameters, dropping components as necessary, to get $\Omega_b - \Omega_a$ non-singular. When this is done, the degrees of freedom of the chi-square test equals the rank of $\Omega_b - \Omega_a$.

Reference: Hausman-McFadden, Econometrica, 1984.

2. **McFadden omitted variables test.**

- **Estimate a MNL model, using all the observations. Suppose A is a specified subset of alternatives. Create new variables in one of the following two forms:**

a. If x_i are the variables in the basic logit model, define new auxiliary variables

$$z_i = \begin{cases} x_i - \sum_{j\in A} P_{j|A} x_j & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$$

where $P_{j|A} = P_j/\sum_{j\in A} P_j$ is the conditional probability of choice of j given choice from A , calculated from the basic estimated model. The variables z_i can be written in abbreviated form as $z_i = \delta_{iA}(x_i - x_A)$, where $\delta_{iA} = 1$ iff $i \in A$ and $x_A = \sum_{j\in A} P_{j|A} x_j$.

b. If $V_i = x_i\beta$ is the representative utility from the basic model, calculated at the basic model estimated parameters, define the new variable

$$z_i = \begin{cases} V_i - \sum_{j\in A} P_{j|A} V_j & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$$

or more compactly, $z_i = \delta_{iA}(V_i - V_A)$.

The auxiliary variable in b. equals the vector of auxiliary variables in a., multiplied by the MNL parameter vector β. The auxiliary variable in b. can also be written as

$$z_i = \begin{cases} -\log(P_i) + \sum_{j \in A} P_{j|A} \log(P_j) & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$$

where P_i is calculated using basic model estimates.

- **To carry out an IIA test, estimate an expanded model that contains the basic model variables plus the new variables z_i, and carry out a LR test that the coefficients of z_i are zero:**

  LR = 2[(Log likelihood with z's) - (Log likelihood without z's)]

  If IIA holds, then this statistic has a chi-square distribution with degrees of freedom equal to the number of non-redundant auxiliary variables added.

Properties:

- **The test using variables of type a. is equivalent to the Hausman-McFadden test for the subset of alternatives A.**

- **The test using variables of type b. is equivalent to a one-degree-of-freedom Hausman-McFadden test focused in the direction determined by the parameters β. It is likely to have greater power than the previous test if there is substantial variation in the V's across A. This test is also equivalent to a Lagrange Multiplier test of the basic MNL model against a <u>nested</u> MNL model in which subjects discriminate more sharply between alternatives within A than they do between alternatives that are not both in A. One plus the coefficient of the variable can be interpreted as a preliminary estimate of a nested logit model inclusive value coefficient for the nest A.**

- **The tests described above are for a single specified subset A (which can vary by observation, although the power of the test is generally highest when A is fixed across observations). It is trivial to test the MNL model against several nests at once, simply by introducing an auxiliary variable for each suspected nest, and testing jointly that the coefficients of these omitted variables are zero. Alternative nests in these tests can be nested or overlapping. The coefficients on the auxiliary variables provide some guide to choice of nesting structure if the IIA hypothesis fails.**

- **If there are subset-A-specific dummy variables, appearing alone and in interaction with some generic variables in the basic model, then some of the auxiliary type a. variables duplicate these variables, and cannot be used in the testing procedure. In the extreme case that all the original model variables appear in a form equivalent to allowing all interactions of generic variables and a subset-A-specific dummy, it is impossible to distinguish IIA failure from the effect of these variables on choice, and the IIA test has zero power.**

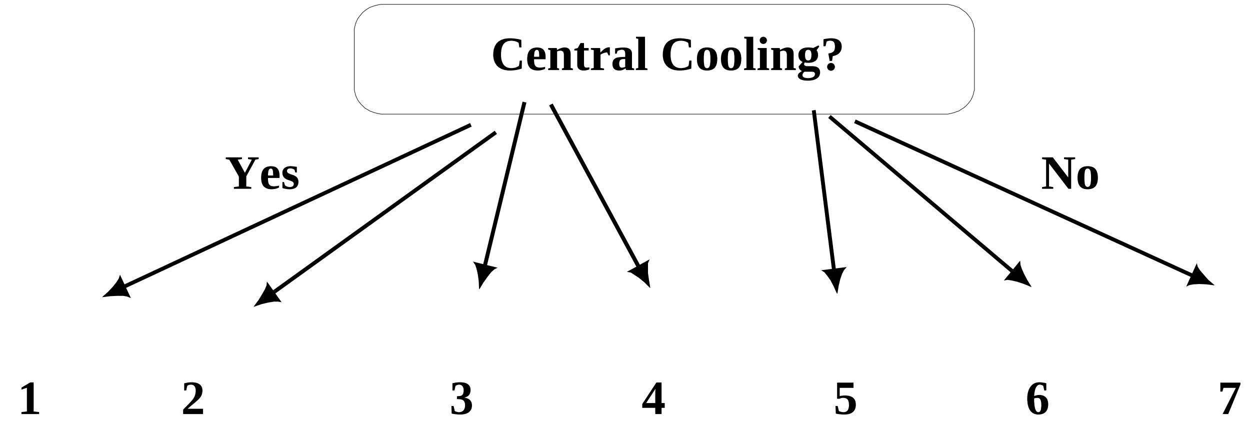
- **One may get a rejection of the null hypothesis <u>either</u> if IIA is false, <u>or</u> if there is some other problem with the model specification, such as omitted variables or a failure of the logit form due, say, to asymmetry or to fat tails in the disturbances.**

- **Rejection of the IIA test will often occur when IIA is false, even if the nest A does not correctly represent the pattern of nesting. However, the test will typically have greatest power when A is a nest for which an IIA failure occurs.**

Reference: D. McFadden, "Regression based specification tests for the multinomial logit model" <u>Journal</u> <u>of</u> <u>Econometrics</u>, 1987.

EXTENSIONS OF MNL: NESTED LOGIT

Example: Choice of Heating Type and Central Cooling



Alternatives:

1. **Central cooling and gas central heating**
2. **Central cooling and electric central heating**
3. **Central cooling and electric room heating**
4. **Heat pump (central cooling and heating)**
5. **No central cooling, gas central heating**
6. **No central cooling, electric central heating**
7. **No central cooling, electric room heating**

Model **$U_i = V_i + \varepsilon_i$, $i = 1,...,7$ $\varepsilon_i \sim$ Generalized Extreme Value**
with: correlation among ε_1 , ε_2 , ε_3 , and ε_4 ;
correlation among ε_5 , ε_6 , and ε_7 ;
no correlation between ε_i , $i = 1,2,3,4$ and ε_j , $j = 5,6,7$;
and 1 - λ is a measure of correlation.

Then:

$$i = 1,...,4 \qquad P_i = \frac{e^{V_i/\lambda}\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda-1}}{\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda} + \left(\sum_{j=5}^{7} e^{V_j/\lambda}\right)^{\lambda}}$$

$$i = 5,...,7 \qquad P_i = \frac{e^{V_i/\lambda}\left(\sum_{j=5}^{7} e^{V_j/\lambda}\right)^{\lambda-1}}{\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda} + \left(\sum_{j=5}^{7} e^{V_j/\lambda}\right)^{\lambda}}$$

IIA holds within nests but not across nests:

$$\frac{P_1}{P_2} = \frac{e^{V_1/\lambda}\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda-1}}{e^{V_2/\lambda}\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda-1}} = \frac{e^{V_1/\lambda}}{e^{V_2/\lambda}}$$

- depends on V_1 and V_2 only.

$$\frac{P_1}{P_5} = \frac{e^{V_1/\lambda}\left(\sum_{j=1}^{4} e^{V_j/\lambda}\right)^{\lambda-1}}{e^{V_5/\lambda}\left(\sum_{j=5}^{7} e^{V_j/\lambda}\right)^{\lambda-1}}$$

- depends on all V_1 ,..., V_7.

- **An improvement in the attributes of one alternative draws proportionately from other alternatives in the nest, but less than proportionately from alternatives outside the nest.**

THE ISSUES IN SPECIFIYING ALTERNATIVES TO MNL

■ **The multivariate integral defining $P_i(z,\theta)$,**

$$P_i(z,\theta) = \Pr(\Delta_i u \le 0 \mid z,\theta) \equiv \int_{\Delta_i u \le 0} f(u \mid z,\theta)du,$$

can be calculated analytically in special cases, notably multinomial logit and its generalizations. However, for most densities the integral is analytically intractable, and for dimensions much larger than J = 5 is also intractable to evaluate with adequate precision using standard numerical integration methods. (Numerical integration works by forming a weighted average of values of the integrand at judiciously selected points. A typical procedure called Gaussian quaditure can get acceptable precision for most problems with about 10 evaluation points per dimension, so that the total number of function evaluations required is about 10^{J-1}. This count rises too rapidly with J to be feasible for J much above 5.) Then, the four practical methods of working with random utility models for complex applications are (1) use of nested multinomial logit and related specializations of Generalized Extreme Value (GEV) models, (2) use of multinomial probit with special factor-analytic structure to provide feasible numerical integration; (3) use of multinomial probit with simulation estimators that handle high dimensions; and (4) use of mixed (random coefficients) multinomial logit, with simulation procedures for the coefficients.

GEV Models

■ Let C = {1,...,J}. Let 1_j denote the unit vectors for $j \in C$, and for $A \subseteq C$, let 1_A denote a vector with components that are one for the elements of A, zero otherwise. Assume that the indirect utility of i can be written

$$u_i = x_i\beta + \varepsilon_i,$$

where x_i is a vector of attributes of alternative i, β is a parameter vector, and ε_i is a part that varies randomly across consumers. Let $v_i = x_i\beta$ index the desirability of alternative i.

Define a *GEV generating function* $H(w_1,...,w_J)$ on $w = (w_1,...,w_J) \geq 0$ to have the properties that it is non-negative, homogeneous of degree one, and differentiable, with its mixed partial derivatives for j = 1,...,J satisfying $(-1)^j\partial^jH/\partial w_1...\partial w_j \leq 0$. A GEV generating function H is *proper* with respect to a subset A of C if $H(1_j) > 0$ for $j \in A$ and $H(1_{C\backslash A}) = 0$. Let $\mathcal{H}$ denote the family of GEV generating functions, and let $\mathcal{H}(A)$ denote the subfamily that is proper with respect to A. Let $\gamma = 0.5772156649$ denote Euler's constant.

Theorem. **If a random vector $U = (U_1,...,U_J)$ has a GEV distribution $F(u) = Prob(U \leq u)$, then this distribution has the form**

[1] $$F(u) = \exp(-H(\exp(-u_1 + v_1),...,\exp(-u_J + v_J))),$$

where $(v_1,...,v_J)$ are location parameters and $H(w_1,...,w_J)$ is a non-negative function of $w \geq 0$ which is homogeneous of degree one and satisfies $H(1_j) > 0$ for $j \in C$. Conversely, a sufficient condition for the function [1] to be a GEV distribution is that $H \in \mathcal{H}(C)$. GEV distributions have the properties:

A. $f(u) = \partial^J F(u)/\partial w_1...\partial w_J \geq 0$, $F(u) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_J} f(u)du$, and $0 \leq F(u) \leq 1$.

B. The U_j for $j = 1,...,J$ are EV1 with common variance $\pi^2/6\mu^2$, means $v_j + \mu^{-1} \log H(1_j) + \gamma/\mu$, and moment generating functions $\exp(tv_j)H(1_j)^{t/\mu}\Gamma(1-t/\mu)$.

C. $U_0 = \max_{i=1,...,J} U_i$ is EV1 with variance $\pi^2/6\mu^2$, mean $(\log H(\exp(v_1),...,\exp(v_J))) + \gamma)/\mu$, and moment generating function $H(\exp(v_1),...,\exp(v_J))^{t/\mu}\Gamma(1-t/\mu)$.

D. Letting $H_j(w) = \partial H(w)/\partial w_j$, the probability P_j that $U_j = \max_{i=1,...,J} U_i$ satisfies

[2] $$P_j = \exp(v_j)\cdot H_j(\exp(v_1),...,\exp(v_J))/\mu H(\exp(v_1),...,\exp(v_J)).$$

The linear function $H(w) = w_1 + ... + w_J$ is a GEV generating function; the vector U with the distribution function [1] for this H has independent extreme value distributed components. The choice probabilities [2] for this case have a *multinomial logit* (MNL) form,

[3] $$P_j = \exp(v_j)/\sum_{i\in C}\exp(v_i).$$

The next result gives operations on GEV generating functions that can be applied recursively to generate additional GEV generating functions.

Lemma 2. The family $\mathcal{H}$ of GEV generating functions is closed under the following operations:

A. If $H(w_1,...,w_J) \in \mathcal{H}(A)$, then $H(\alpha_1 w_1,...,\alpha_J w_J) \in \mathcal{H}(A)$ for $\alpha_1,...,\alpha_J > 0$.
B. If $H(w_1,...,w_J) \in \mathcal{H}(A)$ and $s \geq 1$, then $H(w_1^s,...,w_J^s)^{1/s} \in \mathcal{H}(A)$.
C. If $H^A(w_1,...,w_J) \in \mathcal{H}(A)$ and $H^B(w_1,...,w_J) \in \mathcal{H}(B)$, where A and B are subsets of C, not necessarily disjoint, then
$$H^A(w_1,...,w_J) + H^B(w_1,...,w_J) \in \mathcal{H}(A\cup B).$$

A three-level nested MNL model is generated by a function H of the form

$$\mathbf{H} = \sum_{m=1}^{M} \left[\sum_{k=1}^{K} \left[\sum_{i \in A_{mk}} w_i^{s_m^{/} s_k} \right]^{\frac{1}{s_k}} \right]^{\frac{1}{s_m^{/}}} \quad ,$$

where the A_{mk} partition {1,...,J} and s_k,s_m’ ≥ 1. This form corresponds to a tree: m indexes major branches, k indexes limbs from each branch, and i indexes the final twigs. The larger s_k or s_m’, the more substitutable the alternatives in A_{mk}. If s_k = s_m’ = 1, this model reduces to the MNL model.

If the utility index v_i is linear in income, with a coefficient α, then the expected change in utility in moving from one environment to another, measured in income units, is

$$\mathbf{WTP} = \frac{1}{\alpha} \cdot \left\{ \log H(e^{v''_1},...,e^{v''_J}) - \log H(e^{v'_1},...,e^{v'_J}) \right\}$$

This is the "log sum" formula first developed by Ben-Akiva (1972), McFadden (1973), and Domencich and McFadden (1975) for the multinomial logit model, and by McFadden (1978, 1981) for the nested logit model. This formula is valid *only* when the indirect utility function is linear in income.

The MNP Model

■ A density that is relatively natural for capturing unobserved effects, and the patterns of correlation of these effects across alternatives, is the multivariate normal distribution with a flexible covariance matrix. This is termed the *multinomial probit* model.

■ If $\varepsilon = z\xi$, where ξ is interpreted as a random variation in "taste" weights across observations with $\xi \sim N(0,\Omega)$, then the transformed variable $w = \Delta_i u$ is multivariate normal of dimension J-1 with mean $\Delta_i z\beta$ and covariance $\Delta_i z\Omega z'\Delta_i'$. Unless $J \leq 5$ or dimensionality can be reduced because ξ has a factorial covariance structure, the resulting MNP response probabilities are impractical to calculate by numerical integration. The method of simulated moments was initially developed to handle this model; see McFadden (1989).

- **The log likelihood of an observation is**

$$\boldsymbol{l}(\theta) = \sum_{i\in C} \mathbf{d_i \cdot \log P_i(z,\theta)} \ ,$$

where d_i is an indicator for the event that i is chosen. The *score* of an observation is then

$$\boldsymbol{s}(\theta) = \sum_{i\in C} \mathbf{d_i \cdot \nabla_\theta \log P_i(z,\theta)} \equiv \sum_{i\in C} \mathbf{[d_i - P_i(z,\theta)] \cdot \nabla_\theta \log P_i(z,\theta)} \ ,$$

with the second form holding because $0 \equiv \sum_{i\in C} \nabla_\theta P_i(z,\theta)$.

This score can be adapted to Generalized Method of Simulated Moments (GMSM) or Method of Simulated Scores (MSS) estimation when $P_i(z,\theta)$ is intractable by conventional analysis. Simulators are required for $P_i(z,\theta)$ and $\nabla_\theta \log P_i(z,\theta)$.

■ **Consider the problem of approximating the multinomial probit (MNP)**

$$\text{(1)} \qquad \mathbf{P} \equiv \mathbf{P(B;\mu,\Omega)} = \boldsymbol{n}\mathbf{(v\text{-}\mu,\Omega)dv} \equiv \mathbf{E_V 1(V\in B)},$$

where V is a m-dimension normal random vector with mean μ, covariance matrix Ω, with density denoted by *n*(v - μ,Ω), and 1(V∈B) is an indicator for B = {V| a < V < b}.

■ **The derivatives of (1) with respect to μ and Ω are**

$$\text{(2)} \qquad \nabla_{\mu}\mathbf{P(B;\mu,\Omega)} = \Omega \int_{-\infty}^{+\infty} \mathbf{1(v\in B)(v\text{-}\mu)}\boldsymbol{n}\mathbf{(v\text{-}\mu,\Omega)dv}$$

$$\equiv \Omega^{-1}\mathbf{E_V\ 1(V\in B)(V\text{-}\mu)},$$

$$\nabla_{\Omega}\mathbf{P(B;\mu,\Omega)} = \frac{\Omega^{-1}}{2} \cdot \int_{-\infty}^{+\infty} \mathbf{1(v\in B)[(v\text{-}\mu)(v\text{-}\mu)'\text{-}\Omega]}\boldsymbol{n}\mathbf{(v\text{-}\mu,\Omega)dv}\cdot\Omega^{-1}$$

$$\equiv \mathbf{(1/2)}\Omega^{-1}\mathbf{E_V\ 1(V\in B)[(V\text{-}\mu)(V\text{-}\mu)'\text{-}\Omega]}\cdot\Omega^{-1}.$$

■ For statistical inference, it is often unnecessary to achieve high numerical accuracy in evaluation of (1) and (2). For example, simulating P by the frequency of the event 1(v∈B) in a number of Monte Carlo draws comparable to sample size will tend to produce statistics in which the variance introduced by simulation is at worst of the same magnitude as the variance due to the observed data. Further, when probabilities appear linearly across observations in an estimation criterion, independent unbiased simulation errors are averaged out. Then, a small, fixed number of draws per probability to be evaluated will be sufficient with increasing sample size to reduce simulation noise at the same rate as noise from the observed data.

MONTE CARLO METHODS

■ Crude Frequency Sampling. The random vector V can be written $V = \mu + \Gamma\eta$, where η is an independent standard normal vector of dimension m and Γ is an lower triangular Cholesky factor of Ω, so $\Omega = \Gamma\Gamma'$. Make repeated Monte Carlo draws of η, and fix these throughout the iteration. Calculate $V = \mu + \Gamma\eta$ for trial parameters (μ,Γ) and form empirical analogs of the expectations (1) and (2). Advantage: very fast, unbiased. Disadvantages: Discontinuous simulator, relative error large for small probabilities.

■ Importance Sampling. Consider the generic integral $H = \int_{-\infty}^{+\infty} 1(v\in B)\cdot h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)dv$, where h is an array of polynomials in v; integrals (1)-(2) have this form. Let g(v) be a density with support B chosen by the analyst. Then,

$$H = \int_{-\infty}^{+\infty} \{h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)/g(v)\}\cdot g(v)dv$$

and a smooth unbiased simulator of H is obtained by drawing from g, fixing these draws, and then for (μ,Ω) averaging $\{h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)/g(v)\}$ over these draws.

■ Advantages: Smoothness, unbiased, and positiveness for simulated P, aid iteration to estimates. Fast if g(v) is an easy density to draw from. Disadvantages: can be inaccurate unless mass of g is concentrated near mass of normal, simulator can exceed one.

■ <u>Geweke-Hajivassiliou-Keane</u> <u>Simulator</u> (GHK) This is an importance sampling simulator that has performed well in comparison with many other simulators. It is based on sampling from recursive truncated normals after a Cholesky transformation. The approach was suggested by Geweke (1986), and has been developed by Hajivassiliou, who proposed the weighting used here. Keene (1988) independently developed a weighting scheme of essentially the same form for a problem of estimating transition probabilities.

Let $v = \mu + \Gamma\eta$, where Γ is the Cholesky factor of Ω. The indicator $1(v\in B)$ is then transformed to $1(\mu+\Gamma\eta\in B)$, which can be written recursively as the product of indicators of the events $B_j(\eta_{<j})$ defined by $(a_j-\mu_j-\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj} < \eta_j < (b_j-\mu_j-\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj}$ for $j = 1,...,m$; $\eta_{<j}$ denotes the subvector of η containing the components below the jth. Define $\phi(\eta_j|B_j(\eta_{<j}) = \phi(\eta_j)1(\eta_j\in B_j(\eta_{<j}))/\Phi(B_j(\eta_{<j}))$, the conditional distribution of η_j given the event $B_j(\eta_{<j})$. Define a weight $\omega(\eta) = \Pi_j\Phi(B_j(\eta_{<j}))$, with j ranging from 1 to M. Then

$$H = \int h(\mu+\Gamma\eta)\omega(\eta)\,\Pi_j\phi(\eta_j|B_j(\eta_{<j}))d\eta.$$

The GHK simulator is obtained by drawing and fixing uniform [0,1] variates ζ, then for (μ,Ω) calculating variates

$$\eta_j = \Phi^{-1}(\zeta_j\Phi((a_j-\mu_j-\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj}) + (1-\zeta_j)\Phi((b_j-\mu_j-\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj})),$$

and then averaging $h(\mu+\Gamma\eta)\omega(\eta)$ over these variates.

■ Advantages: For a broad spectrum of applications, this importance sampling density is concentrated near the ideal truncated multivariate normal density, giving low variance simulators for both probabilities and derivatives that have small relative error even when P is small. Disadvantages: The recursive loops with multiple evaluations of standard normal CDFs and inverse CDFs are computationally costly and may introduce additional approximation errors.

Applications

■ **A number of applications of MNP using simulation have appeared in the literature; some examples:**

Berkovec and Stern (1991) "Job Exit Behavior of Older Men", ***Econometrica,*** **59, 189-210.**

Bolduc, D. (1992) "Generalized autoregressive errors in the multinomial probit model", ***Transportation Research B,*** **26, 155-170.**

Borsch-Supan, A., V. Hajivassiliou, L. Kotlikoff, J. Morris (1992) "Health, Children, and Elderly Living Arrangements", ***Topics in the Economics of Aging*** **(D. Wise, ed.), Univ. of Chicago Press.**

■ **MNP simulation by MSLE using the GHK simulator is available in GAUSS. Code from Hajivassiliou- McFadden-Ruud for GHK and other simulators for probabilities and derivatives, in GAUSS and FORTRAN, is available from Berkeley's Econometrics Laboratory Software Archive (ELSA) on the World Wide Web. These programs are practical for up to 25 alternatives without covariance matrix restrictions, but memory and speed are likely to be problems for larger applications.**

■ **For dynamic applications (e.g., multiperiod binomial probit with autocorrelation), and other applications with large dimension, alternatives to A GHK setup with an unrestricted covariance matrix may perform better. McFadden (1984, 1989) suggests a "factor analytic" MNP with a components of variance structure, starting from**

$$u_i = z_i\beta + \sum_{k=1}^{K} \lambda_{ik}\xi_k + \sigma_i\nu_i ,$$

where $\xi_1,...,\xi_K,\nu_1,...,\nu_J$ are independent standard normal, with the ξ_k interpreted as levels of unobserved factors and the λ_{ik} as the loading of factor k on alternative i. The λ's are identified by normalization and exclusion restrictions.

■ **The choice probabilities for this specification:**

$$P_i(z,\theta) = \int_{\xi=-\infty}^{+\infty} \phi(\nu_i)\cdot \prod_{k=1}^{K} \phi(\xi_k)$$

$$\times \prod^{j\neq i} \Phi\left(\frac{(z_j - z_i)\beta + \sum_k[\Lambda_{jk} - \Lambda_{ik}]\cdot\xi_k + \sigma_i\nu_i}{\sigma_j}\right) \cdot d\nu_i d\xi_1\cdots d\xi_K$$

Mixed MNL (MMNL)

Mixed MNL is a generalization of standard MNL that shares many of the advantages of MNP, allowing a broad range of substitution patterns. Train and McFadden (1999) show that any regular random utility model can be approximated as closely as one wants by a MMNL model. Assume $u_i = z_i\alpha + \varepsilon_i$, with the ε_i independently identically Extreme Value I distributed, and α random with density $f(\alpha;\theta)$, where θ is a vector of parameters. Conditioned on α,

$$L_i(z|\alpha) = \exp(x_i\alpha) / \sum_{j\in C} \exp(x_j\alpha).$$

Unconditioning on α,

$$P_i(z|\theta) = \int_\alpha L_i(z|\alpha)\cdot f(\alpha;\theta)\cdot d\alpha \, .$$

This model can be estimated by sampling randomly from $f(\alpha;\theta)$, approximating $P_i(z|\theta)$ by an average in this Monte Carlo sample, and varying θ to maximize the likelihood of the observations. Care must be taken to avoid chatter in the draws when θ varies.

Example: $\alpha = \beta + \Gamma\nu$, where $\theta = (\beta,\Gamma)$ are the parameters and ν is a standard normal vector that is drawn in a Monte Carlo sample and then fixed during iteration to estimate the parameters.

The MMNL model has proved computationally practical and flexible in applications. It can approximate MNP models well, and provides one convenient route to specification of models with flexibility comparable to that provided by MNP.

A LM Test for MNL Against the Mixed MNL Model

The mixed MNL family is very flexible and can approximate any well-behaved discrete response data generation process that is consistent with utility maximization. However, because the MMNL model requires the use of simulation methods for estimation, it is very useful to have a specification test that can indicate whether mixing is needed. The next result describes a Lagrange Multiplier test for this purpose. This test has the pivotal property that its asymptotic distribution under the null hypothesis that the correct specification is MNL does not depend on the parameterization of the mixing distribution under the alternative.

Theorem. Consider choice from a set **C = {1,...,J}.** ***Let*** **$\mathrm{x_i}$** ***be a*** **1×K** ***vector of attributes of alternative*** **i.** ***From a random sample*** **n = 1,...,N,** ***estimate the parameter*** **α** ***in the simple MNL model*** **$L_C(\mathrm{i};x,\alpha)$ = $e^{x_i\alpha}/\sum_{j\in C} e^{x_j\alpha}$** ***, using maximum likelihood; construct artificial variables***

$$z_{ti} = ½(x_{ti} - x_{tC})^2 \quad \textit{with} \quad x_{tC} = \sum_{j\in C} x_{tj}\cdot L_C(\mathrm{j;x},\hat{\alpha})$$

for selected components **t** ***of*** **x_i*****, and use a Wald or Likelihood Ratio test for the hypothesis that the artificial variables*** **$\mathrm{z_{ti}}$** ***should be omitted from the MNL model. This test is asymptotically equivalent to a Lagrange multiplier test of the hypothesis of no mixing against the alternative of a MMNL model*** **$P_C(\mathrm{i}|x,\theta) = \int L_C(\mathrm{i};x,\alpha)\cdot G(\mathrm{d}\alpha;\theta)$** ***with mixing in the selected components*** **t** ***of*** **α*****. The degrees of freedom equals the number of artificial variables*** **$\mathrm{z_{ti}}$** ***that are linearly independent of x.***

To examine the operating characteristics of the test, consider two simple Monte Carlo experiments for choice among three alternatives, with random utility functions $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. The disturbances ε_i were i.i.d. Extreme Value Type I. In the first experiment, the covariate were distributed as described below:

Variable	**Alternative 1**	**Alternative 2**	**Alternative 3**
x_1	**±½ w.p. ½**	**0**	**0**
x_2	**±½ w.p. ½**	**±½ w.p. ½**	**0**

The parameter $\alpha_2 = 1$ under both the null and the alternative. The parameter $\alpha_1 = 0.5$ under the null hypothesis, and under the alternative $\alpha_1 = 0.5 \pm 1$ w.p. 1/2. We carried out 1000 repetitions of the test procedure for a sample of size N = 1000 and choices generated alternately under the null hypothesis and under the alternative just described, using likelihood ratio tests for the omitted variable z_{1i}.

Nominal Significance Level	**Actual Significance Level**	**Power Against the Alternative**
10%	**8.2%**	**15.6%**
5%	**5.0%**	**8.2%**

The nominal and actual significance levels of the test agree well. The power of the test is low, and an examination of the estimated coefficients reveals that the degree of heterogeneity in tastes present in this experiment gives estimated coefficients close to their expected values. Put another way, this pattern of heterogeneity is difficult to distinguish from added extreme value noise.

In the second experiment, the covariates are distributed:

Variable	Alternative 1	Alternative 2	Alternative 3
x_1	±½ w.p. ½	±½ w.p. ½	0
x_2	±½ w.p. ½	±½ w.p. ½	0

The utility function is again $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. Under the null hypothesis, $\alpha_1 = \alpha_2 = 1$, while under the alternative $(\alpha_1,\alpha_2) = (2,0)$ w.p. ½ and (0,2) w.p. ½. Again, 1000 repetitions of the tests are made for N = 1000 under the null and the alternative:

Nominal Significance Level	Actual Significance Level	Power Against the Alternative
10%	9.7%	52.4%
5%	3.9%	39.8%

In this case where mixing is across utility functions of different variables, the test is moderately powerful. It remains the case in this example that the estimated coefficients in the MNL model without mixing are close to their expected values.