Econometrics 240B

SECOND HALF READER

Daniel McFadden

University of California
Berkeley

CHAPTER 1. DISCRETE RESPONSE MODELS

1. INTRODUCTION

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 good 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., college enrollment by state), 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; college enrollments are in the thousands, 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. Here are some 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 $\varepsilon = -f(z,\zeta) + \mathbf{E}f(z,\zeta)$ 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 $\mathbf{E}f$ 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,

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

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

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

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

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

The associated score and hessian of the log likelihood are

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

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

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 any additional parameters upon which F depends) that are (at least) a local maximum of L_T. It can fail to converge to a 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

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

Differentiating this expression, and noting that $F'(x\beta) = F'(-x\beta)$, we get

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

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

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

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.[1]

The hessian can also be rewritten using d rather than y: $\nabla_{\beta\beta} l = -x' x \cdot s(x\beta)$, where $s(x\beta) = \frac{F'(x\beta)^2}{F(x\beta)F(-x\beta)} - [d - F(x\beta)] \left\{ \frac{F''(x\beta)}{F(x\beta)F(-x\beta)} - \frac{F'(x\beta)^2(1-2F(x\beta))}{F(x\beta)^2 F(-x\beta)^2} \right\}$. The expectation of s(xβ) at

[1] To be precise, iterated NLLS, with the β appearing in the weighting function replaced by the last iterate, will converge to the MLE estimator; a single NLLS *without weighting* provides estimates of β that are consistent and asymptotically normal, but not asymptotically efficient; and *one* iterate with weights calculated from a consistent estimator of β will be asymptotically equivalent to MLE.

the true value 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.[2]

2. FUNCTIONAL FORMS AND ESTIMATORS

In principle, the CDF $F(\varepsilon)$ will have a form deduced from the application; in many cases, this form would naturally be conditioned on the observed explanatory variables. However, an almost universal practice is to assume that $F(\varepsilon)$ 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.

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 $s(x\beta)$ 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.

The remainder of this section deals with some alternatives to maximum likelihood estimation, and can be skipped on first reading. 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

[2] We will see later that there are some more robust estimators, not as simple, that avoid having to place F in a parametric family, or use a non-parametric estimate of F. Sometimes assumptions on F are sufficiently problematic so this extra complexity is worth the trouble.

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 $\partial^2 c/\partial p^2 = - F''(c)/F'(c)^3$. Then, a Taylor's expansion of $F^{-1}(s_j/m_j)$ about p_j gives

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

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

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. Note that 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

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

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

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

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

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

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

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

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 $[d_t - F(x_t b)]/\sqrt{F(xb)F(-xb)}$ from the restricted model on the weighted explanatory variables $x\cdot F'(xb)/\sqrt{F(xb)F(-xb)}$. 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

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

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.

The simplest multinomial response model is *multinomial logit* (MNL), which has a closed form

$$P_C(i|z,s) = \exp(x_i\beta) / \sum_{j \in C} \exp(x_j\beta),$$

where x_i is a vector of known functions of z_i and s. This model is derived from the maximizing framework above by assuming $U(z_i,s,\nu_i) = x_i\beta + \varepsilon_i$, with the ε_i independently identically distributed with the special CDF $\exp(-e^{-\varepsilon_i})$, termed the *Type I extreme value distribution*.

The *likelihood* of observation n from a MNL model for choice from C is

$$l_n = \sum_{i \in C} d_{in} \cdot \log(P_{Cn}(i)),$$

where $P_{Cn}(i) = e^{x_{in}\beta} / \sum_{k \in C} e^{x_{kn}\beta}$, and $d_{in} = 1$ indicates choice and $d_{jn} = 0$ for non-chosen alternatives. The gradient, or *score*, is

$$s_n = \nabla_\beta l_n = \sum_{i \in C} d_{in} \cdot [x_{in} - \sum_{k \in C} x_{kn} \cdot P_{Cn}(k)]$$

$$= \sum_{i \in C} [d_{in} - P_{Cn}(i)] \cdot x_{in}. = \sum_{i \in C} [d_{in} - P_{cn}(I)] \cdot x_{iCn}$$

$$\text{where } x_{Cn} = \sum_{i \in C} P_{Cn}(i) \cdot x_{in}.$$

The score has the interpretation of requiring orthogonality in the sample between the explanatory variables x_{in} and the residuals $d_{in} - P_{Cn}(i)$. The hessian, or *information matrix*, is

$$H_n = -\nabla_{\beta\beta} l_n = \sum_{i \in C} P_{Cn}(i) \cdot [x_{in} - x_{Cn}] \cdot [x_{in} - x_{Cn}]',$$

The matrix H_n is positive semi-definite, and the expectation of H_n will be positive definite so long as the x_{in} are not linearly dependent. This assures that the log likelihood function is concave.

Consider the sample log likelihood $L_N = \sum_{n=1}^{N} l_n$. Any parameter vector that sets the sample score to zero will also be a global maximum, and standard iterative maximization by a procedure like Newton-Raphson will converge to a global maximum.[3] The Newton-Raphson iterative adjustment in parameters will be

$$\Delta\beta = \left(\sum_{n=1}^{N} H_n \right)^{-1} \sum_{n=1}^{N} s_n \equiv \left(\sum_{n=1}^{N} \sum_{i \in C} P_{Cn}(i) \cdot x_{iCn} x_{iCn}' \right)^{-1} \sum_{n=1}^{N} \sum_{i \in C} x_{iCn} \cdot [d_{in} - P_{Cn}(i)],$$

where $x_{iCn} = x_{in} - x_{Cn} \equiv x_{in} - \sum_{i \in C} P_{Cn}(i) \cdot x_{in}$. The adjustment $\Delta\beta$ can also be interpreted as the estimates of the coefficients from a linear regression of $[d_{in} - P_{Cn}(i)] / \sqrt{P_{Cn}(i)}$ on the variables $\sqrt{P_{Cn}(i)} \cdot x_{iCn}$. This has the same form as a Lagrange Multiplier test statistic, and one can write down a criterion for convergence that is identical to a LM test of whether the last iterate of the parameter vector is the true parameter vector. (One would want to accept the hypothesis and stop iterating only if there is very little probability of a type II error, accepting a false hypothesis.

[3] A step size adjustment may speed convergence or avoid "overshooting" that could interfere with convergence.

Therefore, the convergence criterion should use this LM statistic with a very *large* type I error, say 99.9%.)

One implication of the MNL model is that the ratio of the probabilities of two alternatives i and j depends only on x_i and x_j, and not on the presence or properties of other alternatives; i.e., $P_{Cn}(i)/P_{Cn}(j) = e^{(x_{in}-x_{jn})\beta}$. This is called the *Independence from Irrelevant Alternatives* (IIA) property. This is a very restrictive property when x_{in} depends only on attributes of alternative i for each i. It implies patterns of cross-elasticities of substitution that are implausible for many applications. For example, a MNL model of the multinomial choice of school for graduate study in economics makes no allowance for the possibility that there may be unobserved factors shared by several schools (e.g., the Northern California location of Berkeley and Stanford), so that discrimination within this class (which we might call the "blue department" and the "red department") is likely to be sharper than it is between one of these departments and an East Coast department such as Princeton. The IIA property is a powerful restriction which if true can greatly simplify estimation and forecasting, and if false produces a misspecified model that can give misleading estimates and forecasts. The IIA property is not on its face particularly plausible, and what is remarkable about the MNL model is that it often performs well in forecasting situations even when IIA does not appear to be reasonable. However, it is important to understand the consequences of the IIA property of MNL, and to develop models for discrete response that can be used when IIA is clearly invalid.

7. ALTERNATIVES TO THE MNL MODEL FOR MULTINOMIAL RESPONSE

As in the derivation of the MNL model, 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: $y_i = \mathbf{1}(u_i \geq u_j \text{ for } j \in C)$. One form of this model is a random coefficients formulation $u_i = z_i\alpha$, $\mathbf{E}\alpha = \beta$, $\varepsilon_i = z_i(\alpha - \beta)$, implying $\text{cov}(\varepsilon_i,\varepsilon_j) = z_i \cdot \text{Cov}(\alpha) \cdot z_j'$. 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, letting $\mathbf{1}_{J-1}$ denote a (J-1)×1 vector of ones and $\mathbf{I}_{J-1}$ denote an identity matrix of dimension J-1, one has

$$\Delta_1 = [-\mathbf{1}_{J-1} \;\; \mathbf{I}_{J-1}].$$

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

Assume that the indirect utility of i can be written $u_i = v_i + \varepsilon_i$ with ε_i a disturbance and v_i the systematic part of utility, depending on observed variables and unknown parameters. For example, one might have $v_i = \alpha(y-t_i) + \gamma x_i$, where y is income, t_i is the cost of alternative i (including costs of time), and ε_i is a part that varies randomly across consumers. The terms α, γ are parameters.

The ε's have a joint CDF of *generalized extreme value* (GEV) form if

$$F(\varepsilon_1,...,\varepsilon_J) = \exp(-H(e^{-\varepsilon_1},...,e^{-\varepsilon_J})),$$

where (i) $H(w_1,...,w_J)$ is a non-negative linear homogeneous function of $w \geq 0$, satisfying (ii) if any argument goes to $+\infty$, then H goes to $+\infty$; and (iii) the mixed partial derivatives of H exist, are continuous, and alternate in sign, with non-negative odd mixed derivatives. A function H with properties (i) - (iii) will be termed a *GEV generating function*.

Theorem 1. Suppose H(w) for $w = (w_1,...,w_J)$ is a GEV generating function. Then, $F(\varepsilon)$ is a CDF with Extreme Value Type I univariate marginals. Further the random utility model $u_i = v_i + \varepsilon_i$ with ε distributed $F(\varepsilon)$ satisfies

$$\mathbf{E} \max_i u_i = \log H(e^{v_1},...,e^{v_J}) + E,$$

where $E = 0.5772156649$ is Euler's constant, and the choice probabilities satisfy

$$P_i = e^{v_i} \cdot H_i (e^{v_1},...,e^{v_J}) / H(e^{v_1},...,e^{v_J}).$$

The linear function $H = \sum_{i=1}^{J} w_i$ is a GEV generating function which yields the multinomial logit (MNL) model. The following result can be used to build up complex choice models. In this theorem, the sets A and B are not required to be mutually exclusive.

Theorem 2. If sets A,B satisfy $A \cup B = \{1,...,J\}$, $H^A(w_A)$ and $H^B(w_B)$ are GEV generating functions in w_A and w_B, respectively, and if $s \geq 1$, then $H(w) = H^A(w_A^s)^{1/s} + H^B(w_B)$ is a GEV generating function in $(w_1,...,w_J)$.

One can use this theorem to show that a three-level nested MNL model is generated by a function H of the form

$$H = \sum_{m=1}^{M} \left[\sum_{k=1}^{K} \left[\sum_{i \in A_{mk}} w_i^{s_m s'_k} \right]^{\frac{1}{s_m}} \right]^{\frac{1}{s'_k}} ,$$

where the A_{mk} partition {1,...,J} and $s'_k, s_m \geq 1$. 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. The GEV model is most efficiently estimated by MLE, but a convenient (and numerically relatively stable) method of getting preliminary estimates is to proceed sequentially, starting at the innermost nests. At each level of nesting, choice can be represented by a MNL model, which will however depend on parameters estimated from deeper levels of nesting. Details of this estimation procedure are given in McFadden (1984).

One interesting feature of GEV models is that they provide a convenient computational formula for the exact consumers' surplus associated with a policy that changes the attributes of alternatives. Let $v_i' = \alpha(y-t_i) + \gamma x_i'$ and $v_i'' = \alpha(y-t_i) + \gamma x_i''$, where x_i' is the vector of original attributes and x_i'' is the vector of improved attributes. Then, the willingness-to-pay for the change from x' to x'' is

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

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).

For dynamic applications (e.g., multiperiod binomial probit with autocorrelation), and other applications with large dimension, alternatives to simulation of the MNP model with a 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 normalizations and exclusion restrictions. The choice probabilities for this specification are

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

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

Numerical integration (when K+1 < 5) or simulation methods can be used to approximate this function and its derivatives for purposes of approximate maximum likelihood estimation. If simulation is used, two important rules should be followed: First, the Monte Carlo draws used for simulation should be made once and then frozen over the course of iterative search for parameters. This avoids "chatter" that can destroy the statistical properties of simulation-based estimators. Second, the number of simulation draws per observation should rise faster than the square root of sample size. This will assure that the simulation is asymptotically negligible, and cannot interfere with the CAN properties of MLE.

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,

$$L_i(z|\alpha) = e^{z_i\alpha} / \sum_{j\varepsilon C} e^{z_j\alpha} .$$

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. 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.

8. TESTS FOR THE IIA PROPERTY OF MNL

Alternatives to the MNL model may be derived from random utility models in which subsets of alternatives have disturbances ε_{in} that are correlated, perhaps because of common unobserved attributes. Common components of disturbances cancel out of the determination of choice within such a subset. As a result, discrimination of differences in observed attributes

is sharper in a subset than overall; there is less random noise to blur discrimination. Tests for the presence of sharper discrimination in subsets is then a test of the IIA property of the MNL model.

For any discrete response model, including but not limited to MNL, let s_n denote the score of an observation, and H_n the negative of the hessian for an observation. A Taylor's expansion of the sample score about the maximum likelihood estimator establishes that in large samples

$$b - \beta_o = \left(\sum_{n=1}^{N} H_n \right)^{-1} \left(\sum_{n=1}^{N} s_n \right) + O(N^{-1/2}),$$

and the covariance matrix of $b - \beta_o$ is approximately $\Omega = \left(\sum_{n=1}^{N} H_n \right)^{-1}$, where all expressions are evaluated at β_o. In sufficiently large samples, b is approximately normally distributed with mean β_o and covariance matrix Ω, and the quadratic form

$$(b - \beta_o)' \Omega_C^{-1} (b - \beta_o) = \left(\sum_{n=1}^{N} s_n \right)' \left(\sum_{n=1}^{N} H_n \right)^{-1} \left(\sum_{n=1}^{N} s_n \right)$$

is approximately chi-squared distributed with degrees of freedom equal to the dimension of β_o. This is a Wald test statistic for the null hypothesis that $\beta = \beta_o$. It can also be applied to a subvector of β, with the commensurate submatrix of Ω_C^{-1} in the center of the quadratic form, to test the null hypothesis that this subvector takes on specified values.

We describe a series of hypothesis testing procedures that can be interpreted as tests of the IIA property of MNL. We will show a connection between these statistics and conventional test statistics for omitted variables.

Hausman-McFadden IIA Test:[4]

Estimate the MNL model twice, once on a full set of alternatives C, and second on a specified subset of alternatives A and the subsample with choices from this subset. If IIA holds, the two estimates should not be statistically different. If IIA fails, then there may be sharper discrimination within the subset A, so that the estimates from the second setup will be larger in magnitude than the estimates from the full set of alternatives. Let β_A denote the estimates obtained from the second setup, and Ω_A denote their estimated covariance matrix. Let β_C denote the estimates of the same parameters obtained from the full choice set, and Ω_C denote their estimated covariance matrix. (Some parameters that can be estimated from the full choice set may not be identified in the second setup, in which case β_C refers to estimates of the subvector of parameters that are identified in both setups.) Consider the quadratic form

$$(\beta_C - \beta_A)' (\Omega_A - \Omega_C)^{-1} (\beta_C - \beta_A).$$

This 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 $\Omega_A - \Omega_C$

[4] Hausman-McFadden, Econometrica, 1984.

non-singular. When this is done, the degrees of freedom of the chi-square test equals the rank of $\Omega_A - \Omega_C$. The simple form of the covariance matrix for the parameter difference arises because β_C is the efficient estimator for the problem.

McFadden omitted variables test.[5]

Estimate the basic MNL model, using all the observations; let $P_{in} = P_{Cn}(i)$ denote the fitted model. Suppose A is a specified subset of alternatives. Create new variables in one of the following three forms:

a. If x_{in} are the variables in the basic logit model, define new variables

$$z_{in} = \begin{cases} x_{in} - (\sum_{j \in A} P_{jn} x_{jn}) / (\sum_{j \in A} P_{jn}) & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases},$$

The variables z_{in} can be written in abbreviated form as $z_{in} = \delta_{iA}(x_{in} - x_{An})$, where $\delta_{iA} = 1$ iff $i \in A$ and $x_{An} = \sum_{j \in A} P_{jn|A} x_j$ and $P_{jn|A}$ is the conditional probability of choice of j given choice from A, calculated from the base model.

b. If $V_{in} = x_{in}\beta$ is the representative utility from the basic model, calculated at basic model estimated parameters, define the new variable

$$z_{in} = \begin{cases} V_{in} - (\sum_{j \in A} P_{jn} V_{jn}) / (\sum_{j \in A} P_{jn}) & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases},$$

or more compactly, $z_{in} = \delta_{iA}(V_{in} - V_{An})$.

c. Define the new variable

$$z_{in} = \begin{cases} \log(P_{in|A}) - \sum_{k \in A} P_{kn|A} \log(P_{kn|A}) & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases},$$

where $P_{in|A}$ is calculated using the basic model estimates.

☐ The constructions b. and c. are the <u>same</u>. The denominators of the probabilities in the expression $-\log(P_{in|A})$ that appears in the type c. variable drop out, leaving the terms in the construction b.

☐ Estimate an expanded MNL model that contains the basic model variables plus the new variables z_{in}. Then test whether these added variables are significant. If there is a single added

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

variable, as in the construction b., then the T-statistic for this added variable is the appropriate test statistic. More generally, one can form a likelihood ratio statistic

$$LR = 2\left[\begin{pmatrix} Log\ Likelihood \\ with\ z's \end{pmatrix} - \begin{pmatrix} Log\ Likelihood \\ without\ z's \end{pmatrix}\right]$$

If IIA holds, this likelihood ratio statistic has a chi-square distribution with degrees of freedom equal to the number of added z variables (after eliminating any that are linearly dependent).

Properties:

☐ The test using variables of type a. is statistically asymptotically 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 will have greater power than the previous test if there is substantial variation in the V's across A. It is also asymptotically equivalent to a *score* or *Lagrange Multiplier* test of the basic MNL model against a nested MNL model in which subjects discriminate more sharply between alternatives within A than they do between alternatives that are not both in A. One minus the coefficient of the variable can be interpreted as a preliminary estimate of the inclusive value coefficient for the nest A.

☐ If there are subset-A-specific dummy variables in the basic model, then some of the omitted type a. variables are linearly dependent upon these variables, and cannot be used in the testing procedure. Put another way, subset-A-specific dummy variables can mimic the effects of increased discrimination within A due to common unobserved components.

☐ One may get a rejection of the null hypothesis either if IIA is false, or 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.

☐ The tests described above are for a single specified subset A. However, it is trivial to test the MNL model against several nests at once, simply by introducing an omitted variable for each suspected nest, and testing jointly that the coefficients of these omitted variables are zero. Alternative nests in the test can be overlapping and/or nested. The coefficients on the omitted variables and their T-statistics provide some guide to choice of nesting structure if the IIA hypothesis fails.

CHAPTER 2. SAMPLING AND SELECTION

1. INTRODUCTION

Economic survey data are often obtained from sampling protocols that involve stratification, censoring, or selection. Econometric estimators designed for random samples may be inconsistent or inefficient when applied to these samples. Several strands in the econometrics literature have investigated estimators appropriate to such data: seminal papers of Heckman (1974) on sample selection, and Manski and Lerman (1977) on choice-based sampling; further work on endogenous stratification by Hausman and Wise (1977), Manski and McFadden (1981), Cosslett (1981), and Hsieh, Manski, and McFadden (1984); and related work on switching regression by Goldfeld and Quandt (1973, 1975), Madalla and Nelson (1974), and Lee and Porter (1984). This chapter synthesizes this literature, and provides machinery that can be used to crank out estimators for a variety of biased sampling problems.

When the econometrician can influence sample design, then the use of stratified sampling protocols combined with appropriate estimators can be a powerful tool for maximizing the useful information on structural parameters obtainable within a data collection budget.[6]

The estimation problem facing an econometrician can be described, schematically, in terms of a contingency table relating a vector of exogenous variables z and a vector of endogenous variables y, as in the table below where each column and row corresponds to different values for the vector of variables. The joint distribution of (z,y) in the population is a probability

$$(1) \qquad p(z,y) \equiv P(y|z,\beta_o)p(z) \equiv Q(z|y)q(y),$$

where $P(y|z,\beta_o)$ is the *conditional probability* of the endogenous vector y, given the exogenous vector z, defined as a member of a parametric family with true parameter vector β_o; p(z) is the *marginal distribution* of the exogenous variables, obtained by a row sum in the table; q(y) is the *marginal distribution* of y, obtained by a column sum in the table; and $Q(z|y)$ is the *conditional distribution* of z given y, defined by Bayes law in equation (1).[7] We identify $P(y|z,\beta_o)$ as the *structural* model of econometric interest; where by "structural" we mean that this conditional probability law is *invariant* in different populations or policy environments where the marginal distribution of z is altered. A structural model will result if there is a *stable causal relationship* from

[6] Stratification may in itself be economical, permitting the contacting and interviewing of subjects at reduced cost. In addition, stratification may concentrate observations in areas yielding high information on the behavior of economic interest.

[7] In this chapter, we will treat the data vector (z,y) as discrete. There is no fundamental change if some components of (z,y) are continuous; it is merely necessary to replace summations with integrals with respect to appropriate continuous or counting measures. There are additional technical assumptions required to assure measurability and integrability when some components are continuous.

z to y, with no contemporaneous feedback from y to z. One would expect this to be the case if z describes the environment of an economic agent (e.g., prices, income) and y describes the agent's behavioral response (e.g., occupation choice, hours of labor supplied). However, there are many economic applications where it is a reasonable approximation for policy analysis to assume $P(y|z,\beta_o)$ is a "reduced form" with the needed invariance property, without invoking strict assumptions on causality.

	y_1	y_2	.....	y_J	
z_1	$P(y_1\|z_1,\beta_o)p(z_1)$	$P(y_2\|z_1,\beta_o)p(z_1)$	.....	$P(y_J\|z_1,\beta_o)p(z_1)$	$p(z_1)$
z_2	$P(y_1\|z_2,\beta_o)p(z_2)$	$P(y_2\|z_2,\beta_o)p(z_2)$	.....	$P(y_J\|z_2,\beta_o)p(z_2)$	$p(z_2)$
:	:	:		:	:
z_K	$P(y_1\|z_K,\beta_o)p(z_K)$	$P(y_2\|z_K,\beta_o)p(z_K)$	.....	$P(y_J\|z_K,\beta_o)p(z_K)$	$p(z_K)$
	$q(y_1)$	$q(y_2)$	.....	$q(y_J)$	1

A *simple random sample* draws independent observations from the population, each with probability law $P(y|z,\beta_o)\cdot p(z)$. The kernel of the log likelihood of this sample depends only on the conditional probability $P(y|z,\beta)$, not on the marginal density p(z); thus, maximum likelihood estimation of the structural parameters β_o does not require that the marginal distribution p(z) be parameterized or estimated.[8] In this sample, z is *ancillary* to β_o, and the observation that it can be conditioned out without loss of information on β_o can be elevated to a general principle of statistical inference (Cox and Hinckley, 1974).

We next introduce a notation for stratified or biased samples. Suppose the data are collected from one or more *strata*, indexed s = 1,..., S. Each stratum is characterized by a sampling protocol that determines the segment of the population that qualifies for interviewing. Define R(z,y,s) to be the *qualification probability* that a population member with characteristics (z,y) will qualify for the subpopulation from which the stratum s subsample will be drawn. Examples of sampling protocols and their characterizations in terms of qualification probabilities follow:

1. Simple random subsample, with $R(z,y,s) \equiv 1$.

2. Exogenous stratified sampling, with $R(z,y,s) = 1$ if $z \in A_s$ for a subset A_s of the universe Z of exogenous vectors, $R(z,y,s) = 0$ otherwise. The set A_s might define a location, such as a census tract, or a socioeconomic characteristic such as race. The protocol for identifying the qualified subpopulation under locational stratification is typically to enumerate the response units at a location, and then sample randomly from this enumeration. In the

[8] The log likelihood of an observation is $\log P(y|z,\beta) + \log p(z)$, and the kernel of this log likelihood is the part that depends on the parameter vector β.

contingency table, this corresponds to sampling randomly from one or more rows. The protocol for identifying the qualified subpopulation using a socioeconomic criterion is typically a screening interview. Exogenous stratified sampling can be generalized to differential rates by permitting R(z,y,s) to be any function from (z,s) into the unit interval; a protocol for such sampling might be, for example, a screening interview that qualifies a proportion of the respondents that is a function of respondent age.

3. Endogenous stratified sampling, with $R(z,y,s) = 1$ if $y \in B_s$, with B_s a subset of the universe of endogenous vectors Y, and $R(z,y,s) = 0$ otherwise. The set B_s might identify a single alternative or set of alternatives among discrete responses, such as the subpopulation whose appliance and energy consumption choices include an air conditioner. Alternately, B_s might identify a range of a continuous response, such as an income category. A classical choice-based sample for discrete response is the case where each response corresponds to a different stratum. In Figure 1, endogenous sampling corresponds to sampling randomly from one or more columns. Endogenous samples with strata corresponding to single columns are called pure choice-based samples. Endogenous stratified sampling can be generalized to qualification involving both exogenous and endogenous variables, with B_s defined in general as a subset of $Z \times Y$. For example, in a study of mode choice, a stratum might qualify bus riders (endogenous) over age 18 (exogenous). It can also be generalized to differential sampling rates, with a proportion R(z,y,s) between zero and one qualifying in a screening interview.

4. Sample selection/attrition, with R(z,y,s) giving the proportion of the population with variables (z,y) whose availability qualifies them for stratum s. For example, R(z,y,s) may give the proportion of subjects with variables (z,y) that can be contacted and will agree to be interviewed, or the proportion of subjects meeting an endogenous selection condition, say employment, that qualifies them for observation of wage (in z) and hours worked (in y).

The joint probability that a member of the population will have variables (z,y) and will qualify for stratum s is $R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)$. Then for stratum s, the proportion of the population qualifying into the stratum, or *qualification factor*[9], *is*

$$(2) \qquad r(s) = \sum_z \sum_y R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z),$$

and the conditional distribution of (z,y) given qualification is

$$(3) \qquad G(z,y|s) = R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)/r(s).$$

A sample from stratum s is governed by the probability law $G(z,y|s)$. Note that $G(z,y|s)$ depends on the unknown parameter vector β and on the distribution p(z) of the explanatory variables. In simple

[9] The inverse of the qualification factor is called the *raising factor*.

cases of stratification, such as Examples 1-3 above, R(z,y,s) is fully specified by the sampling protocol. The qualification factor r(s) may be known, for example when stratification is based on census tract with known sizes; estimated from the survey, for example when qualification is determined by a screening interview; or estimated from an auxiliary sample. In case of attrition or selection, R(z,y,s) may be an unknown function, or may contain unknown parameters.

Suppose a random sample of size n_s is drawn from stratum s, and let $N = \sum_s n_s$ denote total sample size. Let $n(z,y|s)$ denote the number of observations in the stratum s subsample that fall in cell (z,y).[10] Then, the log likelihood for the stratified sample is

$$(4) \qquad L = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot \text{Log } G(z,y|s).$$

This likelihood does not include screening or auxiliary data on the qualification factors, which will be informative if these factors are unknown.

2. EXOGENOUS STRATIFIED SAMPLING

When the qualification probability R(z,y,s) is independent of y, the qualification factor r(s) $= \sum_z R(z,s)p(z)$ is independent of β_o, and the log likelihood function (4) separates into the sum of a kernel

$$(5) \qquad L_1 = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot \text{Log } P(y|z,\beta)$$

and terms independent of β. Hence, the kernel is independent of the structure of exogenous stratification. This implies that estimators designed for random samples will have the same properties in exogenously stratified samples. The information matrix for the likelihood function under exogenous stratification,

$$(6) \qquad J = \sum_{s=1}^{S} \mu_s \sum_{z} \frac{R(z,s)p(z)}{r(s)} \sum_{y} P(y|z,\beta_o)\cdot[\nabla_\beta \text{Log } P(y|z,\beta_o)]\cdot[\nabla_\beta \text{Log } P(y|z,\beta_o)]',$$

depends on the sample design. Then, exogenous stratification can be used to increase the information available in a sample of given size; this is precisely the objective of classical experimental design.

[10] Note that $n(z,y|s)/n_s$ is the empirical probability measure for a random sample of size n_s from the population with law $G(z,y|s)$. In the case of discrete variables with a finite number of configurations, the $n(z,y|s)$ are simply cell counts. Nothing is changed for continuous variables, except that technically one must consider stochastic limits of empirical processes.

3. ENDOGENOUS STRATIFICATION

Suppose the qualification probability R(z,y,s) depends on y. Then the qualification factor (2) depends on β_o, and the log likelihood function (4) has a kernel depending in general not only on β, but also on the unknown marginal distribution p(z). Further, any unknowns in the qualification probability also enter the kernel. There are four possible strategies for estimation under these conditions:

1. Brute force -- Assume p(z) and, if necessary, R(z,y,s), are in parametric families, and estimate their parameters jointly with β. For example, in multivariate discrete data analysis, an analysis of variance representation absorbs the effects of stratification, and allows one to back out the structural parameters. This approach is straightforward and needs no further discussion for small problems, but is burdensome or infeasible when the Z variables have many dimensions or categories, or are continuous.
2. Weighted Exogenous Sample Maximum Likelihood -- This is a pseudo-maximum likelihood approach which starts from the likelihood function appropriate to a random sample, and reweights the data (if possible) to achieve consistency. A familiar form of this approach is the classical survey research technique of reweighting a sample so that it appears to be random.
3. Conditional Maximum Likelihood -- This approach pools the observations across strata, and then forms the conditional likelihood of y given z in this pool. This has the effect of conditioning out the unknown density p(z).
4. Full Information Maximum Likelihood -- This approach estimates p(z) nonparametrically as a function of the remaining parameters, and substitutes to concentrate the likelihood as a function of the finite parameter vector.

4. WEIGHTED EXOGENOUS SAMPLE MAXIMUM LIKELIHOOD (WESML)

Recall that the kernel of the log likelihood for exogenous sampling is given by (5). Suppose now endogenous sampling with true log likelihood (4), and consider a pseudo- maximum likelihood criterion based on (5),

$$(7) \qquad W(\beta) = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot w(z,y,s)\cdot \text{Log } P(y|z,\beta),$$

where w(z,y,s) is a weight introduced to achieve consistency. Assume that $n_s/N \to \mu_s$ as $N \to \infty$. Then, using the notation "$\to_{as}$" to denote almost sure convergence,

$$(8) \qquad n(z,y|s)/N \equiv [n(z,y|s)/n_s]\cdot[n_s/N] \to_{as} G(z,y|s)\mu_s,$$

implying from (3) that

$$(9)\qquad W(\beta)/N \to_{as} \sum_{s=1}^{S} \mu_s \sum_{z} \sum_{y} G(z,y|s)\cdot w(z,y,s)\cdot \text{Log } P(y|z,\beta)$$

$$= \sum_{z} p(z)\cdot \sum_{y} \{ \sum_{s=1}^{S} R(z,y,s)w(z,y,s)\mu_s/r(s)\}\cdot P(y|z,\beta_o)\cdot \text{Log } P(y|z,\beta).$$

A sufficient condition for consistency of the pseudo-maximum likelihood estimator is that the bracketed term,

$$(10)\qquad \sum_{s=1}^{S} R(z,y,s)w(z,y,s)n_s/N\cdot r(s)$$

be independent of y. Suppose r(s) is consistently estimated by f(s), from government statistics, survey frame data such as the average refusal rate, or an auxiliary sample. Consider the weights

$$(11)\qquad w(z,y) = \left[\sum_{s=1}^{S} R(z,y,s)n_s/Nf(s)\right]^{-1};$$

these are well-defined if the bracketed expressions are positive, and R(z,y,s) contains no unknown parameters. These weights do not depend on the stratum from which the observation is drawn, but do depend generally on the endogenous variable y.

When the qualification probabilities R(z,y,s) are strictly positive for all (z,y) and all strata, and contain no unknowns, another set of possible weights is

$$(12)\qquad w(z,y,s) = 1/R(z,y,s).$$

These can be interpreted as reweighting observations in inverse proportion to the probability with which they qualify from the population, and are precisely the weighting most commonly used in classical survey research. When the weights (11) and (12) are both feasible, the weights (11) are more efficient.

A classical application of WESML estimation is to a sample in which the strata coincide with the possible configurations of y, so that $R(z,y,s) = \mathbf{1}(y = s)$. In this case, $w(z,y) = N\cdot f(y)/n_y$, the ratio of the population to the sample frequency. Another application is to *enriched* samples, where a random subsample (s = 1) is enriched with an endogenous subsamples from one or more configurations of y; e.g., s = y = 2. Then, $w(z,1) = N/n_1$ and $w(z,2) = N\cdot f(2)/[n_1\cdot f(2) + n_2]$.

When the r(s) are known, and f(s) ≡ r(s), the WESML estimator has an asymptotic covariance matrix $J_w^{-1}H_wJ_w^{-1}$, where

$$(13)\qquad J_w = -\sum_{s=1}^{S} (\mu_s/r(s)) \sum_{z} \sum_{y} w(z,y,s)R(z,y,s)P(y|z,\beta_o)p(z)\nabla_{\beta\beta}l\ ,$$

(14) $$H_w = \sum_{s=1}^{S} \mu_s^2 \sum_z \sum_y w(z,y,s)^2[R(z,y,s)P(y|z,\beta_o)p(z)/r(s)]\cdot[\nabla_\beta l]\cdot[\nabla_\beta l]' - \sum_{s=1}^{S} q_s\cdot q_s'$$

where $l = \log P(y|x,\beta)$ and

$$q_s = \sum_z \sum_y \mu_s\, w(z,y,s)\cdot[R(z,y,s)\cdot P(y|z,\beta_o)\cdot p(z)/r(s)]\nabla_\beta l,$$

and l and its derivatives are evaluated at β_o. These covariance terms come from a Taylor's expansion of the first-order conditions for maximization of $W(\beta)$, and can be estimated consistently by replacing terms with their sample analogs.

5. CONDITIONAL MAXIMUM LIKELIHOOD (CML)

Pool the observations from the different strata. Then, the data generation process for the pool is

$$Pr(z,y) = \sum_{s=1}^{S} G(z,y|s)n_s/N,$$

and the conditional probability of y given z from this pool is

$$Pr(y|z) = \frac{\sum_{s=1}^{S} G(z,y|s)n_s/N}{\sum_y \sum_{s=1}^{S|} G(z,y|s)n_s/N} .$$

Substituting (3) yields a formula independent of p(z),

(15) $$Pr(y|z) = \frac{\sum_{s=1}^{S} R(z,y,s)\cdot P(y|z,\beta_o)\cdot n_s/N\cdot r(s)}{\sum_y \sum_{s=1}^{S} R(z,y,s)\cdot P(y|z,\beta_o)\cdot n_s/N\cdot r(s)} .$$

The CML estimator maximizes the conditional likelihood of the pooled sample in β and any unknowns in R(z,y,s). When r(s) is known, or one wishes to condition on estimates f(s) of r(s) from auxiliary samples, (15) is used directly. More generally, given auxiliary sample information on the r(s), these can be treated as parameters and estimated from the product of the likelihood (15) and the likelihood of the auxiliary sample.

For discrete response in which qualification does not depend on z, the formula (15) simplifies to $\Pr(y|z) = \dfrac{P(y|z,\beta_o)\,\alpha_y}{\sum_y P(y|z,\beta_o)\,\alpha_y}$, where $\alpha_y = \sum_{s=1}^{S} R(z,y,s)\cdot n_s/N\cdot r(s)$ can be treated as an alternative-specific constant. For multinomial logit choice models, $\Pr(y|z)$ then reduces to a multinomial logit formula with added alternative-specific constants. It is possible to estimate this model by the CML method using standard random sample computer programs for this model, obtaining consistent estimates for slope parameters, and for the sum of $\log \alpha_y$ and alternative-specific parameters in the original model. It remains necessary to use formulas for endogenous sampling to estimate the asymptotic covariance matrix consistently.

For the previous example of an enriched sample, one has $\Pr(1|z) = P(1|z,\beta_o)\cdot n_1/N\cdot D$ and $\Pr(2|z) = P(2|z,\beta_o)\cdot[n_1/N + n_2/N\cdot r(2)]/D$, where $D = n_1/N + P(2|z,\beta_o)\cdot n_2/N$. An example in a different context shows the breadth of application of (15). Suppose y is a continuous variable, and the sample consists of a single stratum in which high income families are over-sampled by screening, so that the qualification probability is $R(z,y,1) = \gamma < 1$ for $y \le y_o$ and $R(z,y,1) = 1$ for $y > y_o$. Then $\Pr(y|z) = \gamma\cdot P(y|z,\beta_o)/D$ for $y \le y_o$ and $\Pr(y|z) = P(y|z,\beta_o)/D$ for $y > y_o$, where $D = \gamma + (1-\gamma)\cdot P(y>y_o|z,\beta_o)$.

When the r(s) are known, the asymptotic covariance matrix of the CML estimator is $J_c^{-1}H_cJ_c^{-1}$, where

$$(16)\qquad J_c = -\sum_{s=1}^{S} (\mu_s/r(s)) \sum_z \sum_y R(z,y,s)P(y|z,\beta_o)p(z)\nabla_{\beta\beta}c\,,$$

$$(17)\qquad H_w = \sum_{s=1}^{S} \mu_s^2 \sum_z \sum_y [R(z,y,s)P(y|z,\beta_o)p(z)/r(s)][\nabla_\beta c]\cdot[\nabla_\beta c] - \sum_{s=1}^{S} q_s\, q_s'$$

where $c = \log \Pr(y|z,\beta)$ and $q_s = \sum_z \sum_y \mu_s\,[R(z,y,s)P(y|z,\beta_o)p(z)/r(s)]\nabla_\beta c$, and c and its derivatives evaluated at β_o. Note that the structure of this covariance matrix is the same as that for WESML.

6. FULL INFORMATION CONCENTRATED MAXIMUM LIKELIHOOD (FICLE)

Formally, the likelihood (4) can be treated as a function of the unknown parameter vector β, any unknown parameters in the qualification probabilities, and the unknown multivariate density p(z), with this whole density treated as an unknown parameter, possibly infinite dimensional. This is a *semiparametric* estimation problem, in which a finite parameter vector is to be estimated in the presence of a possibly infinite-dimensional vector of nuisance parameters. In some applications, this can be done by direct formal maximization of the likelihood in p(z), given the remaining parameters, yielding a concentrated likelihood function of the finite parameter vector.
Let

$$(18)\quad L = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot \text{Log } G(z,y|s)$$

$$+ \sum_{s=1}^{S} \lambda_s[r(s) - \sum_{z} \sum_{y} R(z,y,s)P(z,y,s)p(z)] + \lambda_o[1 - \sum_{z} p(z)]$$

be a Lagrangian for the formal maximization problem. Solving the first-order-condition for p(z) yields

$$(19)\quad p(z) = \left(\sum_{s=1}^{S} \sum_{y} n(z,y|s) \right) \Big/ \left(\sum_{s=1}^{S} \sum_{y} \lambda_s R(z,y,s)\cdot P(y|z,\beta_o) + \lambda_o \right) .$$

Substituting (19) into (18), simplifying, and dropping terms independent of the unknowns, yields

$$(20)\quad L_1 = \sum_{s=1}^{S} \sum_{z} \sum_{y} n(z,y|s)\cdot \text{Log} \frac{R(z,y,s)\cdot P_{\cdot}(y|z,\beta)/r(s)}{N + \sum_{s=1}^{S|} \lambda_s[\sum_{y} R(z,y,s)\cdot P_{\cdot}(y|z,\beta) - r(s)]}$$

$$+ \sum_{z} \left[\sum_{s=1}^{S} \sum_{y} n(z,y|s) \right] \cdot \frac{\sum_{s=1}^{S} \lambda_s[r(s) - \sum_{y} R(z,y,s)\cdot P_{\cdot}(y|z,\beta)]}{N + \sum_{s=1}^{S|} \lambda_s[\sum_{y} R(z,y,s)\cdot P_{\cdot}(y|z,\beta) - r(s)]}$$

A joint critical point of this concentrated function in β and the λ_s gives the FICLE estimator. Cosslett (1981) has shown that estimators in this class are fully efficient. Since this is a semiparametric problem, Cosslett's argument required calculation by variational methods of the least information contained in the parametric part of the problem; this method in its general form provides what are now called the Wellner efficiency bounds. The asymptotic covariance matrix of the FICLE estimators has the same general structure as the previous estimators, but the specifics are complicated by the presence of the finite vector of nuisance parameters λ_s. For straightforward response-based endogenous samples, with y used to define non-overlapping strata, the FICLE criteria and the CML criteria can be manipulated into almost the same form, with $n_s/Nf(s)$ and λ_s/N appearing in analogous positions and converging to the same limit.

7. EXTENSIONS AND CONCLUSIONS

Both the WESML and CML estimators are computationally practical in a variety of endogenous sampling situations, and have been widely used. In general, neither estimator dominates the other. Monte Carlo experience is that the WESML estimator is more efficient when the weights for different alternatives are nearly the same, and that CML is more efficient when the weights differ substantially across alternatives. The FICLE estimator has not been widely used.

When the population qualification factors r(s) are unknown, and consistently estimated by f(s) obtained from auxiliary data, then the estimators described above are consistent. However, in computing the asymptotic covariance matrices of the estimators, it is necessary to take account of presence of estimated quantities in estimation criterion. This will in general contribute additional terms to the asymptotic covariance matrix; see Newey and McFadden (1995). A more efficient procedure is to estimate the r(s) jointly using the sample and auxiliary data. Hsieh, Manski, and McFadden (1985) develop the procedures for doing this.

Extensions of the theory of endogenous sampling can be made to more complex applications, and to more complex sources of auxiliary information, such as duration data (with length-biased sampling) and endogenously recruited panel data,; see Lancaster and Imbens (1990) and McFadden (1996).

8. SELECTION

There are a variety of econometric problems where dependent variables are discrete, censored at lower or upper limits, or truncated or selected so they are not always observed. It is often convenient to model the behavior of such variables as the result of a two-stage process,

$$\begin{bmatrix} \textit{Exogenous} \\ \textit{Variables} \end{bmatrix} \longrightarrow \begin{bmatrix} \textit{Latent} \\ \textit{Dependent Variables} \end{bmatrix} \longrightarrow \begin{bmatrix} \textit{Observed} \\ \textit{Dependent Variables} \end{bmatrix},$$

where there are intermediate unobserved (latent) variables that are in the first stage determined by exogenous variables through a conventional linear model, and observed dependent variables that in the second stage are determined by some non-linear mapping. The structure of the first mapping, the dimensionality of the latent variables, and the structure of the non-linear mapping can all be varied to fit particular applications. Historically, latent variable models come from psychometrics, where both the mappings from exogenous variables to latent variables, and from latent variables to observed dependent variables are linear, and the critical feature is that the dimensionality of the latent variables is much lower than the dimensionality of the observed dependent variables. A classical psychometric application is to ability testing, where the observed dependent variables are responses to test items, and the latent variables are *factors* such as verbal, quantitative, and motor abilities. In their most general form, these are called Multiple-Indicator, Multiple Cause (MIMC) models, and analysis of the mapping from latent to observed dependent variables is called *factor analysis*. An example of an economic application of MIMC models is the Friedman permanent income hypothesis, where the observed dependent variables are measured yearly incomes and there is a single latent variable, permanent income. These lecture notes will discuss the second major application of latent variable models, to situations where the mapping from latent to observed dependent variables is nonlinear, and the observed dependent variables are not necessarily continuous.

A fairly general notation for a model with m latent variables for each observation unit is $y_j^* = x_j\beta + \varepsilon_j$, where j = 1,...m. This can be written more compactly in matrix notation as $y^* = X\beta + \varepsilon$, where $y^* \in \mathbb{R}^m$ is a m×1 vector of latent variables for one observation, X is a m×k array of explanatory variables whose rows are the x_j vectors, β is a k×1 vector of parameters, and ε is a m×1 vector of disturbances with a multivariate density $f(\varepsilon|\theta)$ that contains additional parameters θ. This

notation can accommodate β parameters that differ across equations by introducing variables in each x_j in interaction with dummies for the different equations. The observed dependent variables are given by a mapping $y = h(y^*)$ that is in general nonlinear and many-to-one. Some examples illustrate the possibilities, and indicate the scope of possible applications:

(1) $y^* \in \mathbb{R}^1$ and $y = h(y^*) = \begin{cases} +1 & if\ y^* \geq 0 \\ -1 & if\ y^* < 0 \end{cases}$ generates a binomial response model. An application might be to firms' decisions to go bankrupt or stay in business, where y^* is latent *expected* profit; see also application (5) below.

(2) $y^* \in \mathbb{R}^1$ and $y = h(y^*) = \begin{cases} y^* & if\ y^* \geq 0 \\ 0 & if\ y^* < 0 \end{cases}$ generates a *censored data* (*Tobit*) model. An application might be to expenditure on clothing in a one-week observation period, where zeros are common.

(3) $y^* \in \mathbb{R}^1$ and $y = h(y^*) = \begin{cases} y^* & if\ y^* \geq c \\ NA & if\ y^* < c \end{cases}$, where NA means no observation is available and c is a constant, generates a *truncated data* model. An application might be to competitive (among buyers) auction prices for units of a good, where a transaction is observed only if a bid exceeds a reservation price c. In case $y^* < c$, one may in one variant of this model observe x, and in another variant observe nothing about x.

(4) $y^* \in \mathbb{R}^1$ and $y = h(y^*)$ is given by $y = i$ if $\lambda_i \leq y^* < \lambda_{i+1}$ for $i = 0,...,J$, with $\lambda_0 = -\infty$ and $\lambda_{J+1} = +\infty$, where λ_1 to λ_J are parameters. This mapping generates an ordered response or *count* model. An application might be to household choice of number of children, or to wealth or income within brackets established by the questionnaire.

(5) $y^* \in \mathbb{R}^2$ and $y = h(y^*) = \begin{cases} (+1, y_2^*) & if\ y_1^* \geq 0 \\ (-1, NA) & if\ y_1^* < 0 \end{cases}$ has the following interpretation: if $y_1^* \geq 0$, then $y_1 = +1$ is an indicator for this, and $y_2 = y_2^*$ is observed. If $y_1^* < 0$, then $y_1 = -1$ is an indicator for this, and y_2 is not observed. Variants may have x_2 observed or not when y_2 is unobserved. An application is to bankruptcy decisions of the firm, where y_1^* is expected profit and y_2^* is realized profit. This is termed a *bivariate selection* model.

(6) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{1,...,m\}$, where $y = i$ if $y_i^* \geq y_j^*$ for $j \neq i$. This generates a *multinomial* response model in which the observed response corresponds to the maximum of the latent variables. An application might be to choice of occupation.

(7) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{-1,+1\}^m$, with $y_j = +1$ if $y_j^* \geq 0$, and $y_j = -1$ otherwise. This generates a *multivariate binomial* response model. An application might be to panel data on employment status.

(8) $y^* \in \mathbb{R}^m$ and $y = h(y^*)$ is a mapping from $\mathbb{R}^m$ into $\{0,1,2,...\}^m$, with $y_j = k_j$ for an integer k_j if $\lambda_{ij} \leq y_{i,j+1} < \lambda_{i,j+1}$. This is a multivariate ordered response or *count* model. An application is to numbers of units purchased of each of m goods.

Let $\mathbf{A}(y)$ denote the set of y^* that map into observation y; this can be written as $\mathbf{A}(y) = h^{-1}(y)$, where h^{-1} denotes the inverse of the (possibly) many-to-one mapping h. Then, the probability of an observation can be written

$$g(y|X,\beta,\theta) = \int_{\mathbf{A}(y)} f(y^*-X\beta|\theta)dy^*.$$

The integral should be interpreted as extending over the dimensions where the condition $y^* \in h^{-1}(y)$ gives a range of values. In the Tobit example (2) above, $y = 0$ implies $h^{-1}(0) = (-\infty,0]$, and the integral is over this interval. However, $y > 0$ implies $h^{-1}(y) = y$, and $g(y|X,\beta,\theta) = f(y-X\beta|\theta)$ without integration. In the bivariate selection model (5), the observation $(+1,y_2)$ requires integration in one dimension, $g((+1,y_2)|X,\beta,\theta) = \int_0^{+\infty} f(y_1^*-x_1\beta,y_2-x_2\beta|\theta)dy_1^*$, while the observation (-1,NA) requires integration in both dimensions, $g((-1,NA)|X,\beta,\theta) = \int_{-\infty}^{0} \int_{-\infty}^{+\infty} f(y_1^*-x_1\beta,y_2^*-x_2\beta|\theta)dy_1^*dy_2^*$.

Consider the log likelihood of an observation, $l(\beta,\theta) = \log g(y|X,\beta,\theta)$. The *score* with respect to the parameters $\gamma = (\beta,\theta)$ is

$$\nabla_\gamma l(\beta,\theta) = \frac{\int_{\mathbf{A}(y)} \left\{\nabla_\gamma \log f(y^*-X\beta|\theta)\right\} f(y^*-X\beta|\theta)dy^*}{\int_{\mathbf{A}(y)} \mathbf{A}(y) f(y^*-X\beta|\theta)dy^*}$$

$$= \mathbf{E}\left\{\nabla_\gamma \log f(y^*-X\beta|\theta) | y^* \in h^{-1}(y)\right\};$$

that is to say, *the score of the observation* y *can be expressed as the conditional expectation of the score of the latent variable model, conditioned on the event that the latent vector yields* y. If these integrals can be evaluated analytically or numerically, then it is usually feasible to do maximum likelihood estimation of the parameters. Even when the integrals are intractable, it may be possible to approximate them by simulation methods.

The basic latent variable model setup above can be extended in several ways. For time-series or panel data, X may contain variables determined by lagged latent variables. If disturbances are serially correlated, one confronts all the problems of identification, stationarity, and consistent estimation that occur in conventional linear systems, plus additional problems of dealing with initial conditions. The leading author who has worked on these problems is Heckman. The latent variable model can also be extended to have a more full-blown simultaneous-equations form,

with complex paths linking observed and latent variables, with a *multiple-indicator, multiple-cause* structure. Leading authors on MIMC models are Goldberger and Joreskog.

9. THE BIVARIATE SELECTION PROBLEM

An important economic application of latent variable models is to the problem of *selection*: Who or what we can observe about economic agents is influenced by their behavior, so that our data are not representative of the whole population. Our analysis needs to correct for the effects of selection if we are to make consistent inferences about the population. A classic example of selection occurs in the study of wages and hours worked of married women. These variables are observed only for women who are working, but the same economic factors that determine these variables also influence the decision to work. For example, an unobserved disturbance that gives Mrs. Smith a higher-than-average potential wage and Mrs. Jones a lower than-average potential wage is more likely to induce Mrs. Smith into the labor force than Mrs. Jones. Then, a regression of wage on family characteristics using data for workers will typically overestimate the potential wage of non-workers. The econometric analysis of this problem provides a good tutorial for a broad spectrum of selection problems that arise because of economic behavior or because of survey design (e.g., deliberate stratification).

Consider a bivariate latent variable model with normal disturbances,

$$\begin{aligned} y^* &= x\beta + \varepsilon \,, \\ w^* &= z\alpha + \sigma\nu \,, \end{aligned} \tag{21}$$

where x and z are vectors of exogenous variables, not necessarily all distinct, α and β are parameter vectors, again not necessarily all distinct, and σ is a positive parameter. The interpretation of y^* is latent desired hours of work, and of w^* is latent log potential wage. The disturbances ε and ν have a standard bivariate normal distribution

$$\begin{bmatrix} \varepsilon \\ \nu \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \,, \tag{22}$$

with zero means, unit variances, and correlation ρ.

There is a nonlinear *observation rule* determined by the application that maps the latent variables into observations. A typical rule might be "Observe y = 1 and w = w^* if $y^* > 0$; observe y = -1 and do <u>not</u> observe w when $y^* \leq 0$". This could correspond, for example, to an application where the event of working (y = 1) or not working (y = 0) is observed, but actual hours worked are not, and the wage is observed only if the individual works ($y^* > 0$). It is sometimes convenient to code the discrete response as s = (y+1)/2; then s = 1 for workers, s = 0 for non-workers.

The event of working is given by a *probit* model. The probability of working is $P(y=1|x) = P(\varepsilon > -x\beta) = \Phi(x\beta)$, and of not working is $P(y=-1|x) = P(\varepsilon \leq -x\beta) = \Phi(-x\beta)$, where Φ is the standard univariate cumulative normal. This can be written compactly as

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

In the bivariate normal, the conditional density of one component given the other is univariate normal,

$$\varepsilon \mid \nu \sim N(\rho\nu, 1-\rho^2) = \frac{1}{\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{\varepsilon - \rho\nu}{\sqrt{1-\rho^2}}\right)$$

and

$$\nu \mid \varepsilon \sim N(\rho\varepsilon, 1-\rho^2) = \frac{1}{\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{\grave{} - \rho\varepsilon}{\sqrt{1-\rho^2}}\right) .$$

The joint density can be written as the product of the marginal density of one component times the conditional density of the other,

$$(\varepsilon,\nu) \sim \varphi(\nu) \cdot \frac{1}{\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{\varepsilon - \rho\nu}{\sqrt{1-\rho^2}}\right) = \varphi(\varepsilon) \cdot \frac{1}{\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{\grave{} - \rho\varepsilon}{\sqrt{1-\rho^2}}\right) .$$

The density of (y^*, w^*) can then be written

$$(23) \qquad f(y^*, w^*) = \frac{1}{\sigma} \varphi\left(\frac{w^* - z\alpha}{\sigma}\right) \cdot \frac{1}{\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{y^* - x\beta - \rho(w^* - z\alpha)/\sigma}{\sqrt{1-\rho^2}}\right)$$

$$= \varphi(y^* - x\beta) \cdot \frac{1}{\sigma\sqrt{1-\rho^2}} \cdot \varphi\left(\frac{w^* - z\alpha - \rho\sigma(y^* - x\beta)}{\sigma\sqrt{1-\rho^2}}\right) .$$

Now consider the log likelihood of an observation, $l(\alpha,\beta,\sigma,\rho)$. In the case of a non-worker ($y = -1$ and $w = NA$), the density (23) is integrated over $y^* < 0$ and all w^*. Using the second form in (23), this gives probability $\Phi(-x\beta)$. In the case of a worker, the density (23) is integrated over $y^* \geq 0$. Using the first form in (23)

$$(24) \qquad e^{l(\alpha,\beta,\sigma,\rho)} = \begin{cases} \Phi(-x\beta) & if\ y = -1 \\ \frac{1}{\sigma}\varphi\left(\frac{w - z\alpha}{\sigma}\right) \cdot \Phi\left(\frac{x\beta + \rho\left(\frac{w - z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) & if\ y = 1 \end{cases} .$$

The log likelihood can be rewritten as the sum of the marginal log likelihood of the discrete variable y and the conditional log likelihood of w given that it is observed, $l(\alpha,\beta,\sigma,\rho) = l^1(\alpha,\beta) + l^2(\alpha,\beta,\sigma,\rho)$, with the marginal component,

$$(25) \qquad l^1(\beta) = \log \Phi(yx\beta) ,$$

and the conditional component (that appears only when $y = 1$),

(26) $$l^2(\alpha,\beta,\sigma,\rho) = -\log\sigma + \log\varphi\left(\frac{w-z\alpha}{\sigma}\right) + \log\Phi\left(\frac{x\beta+\rho\left(\frac{w-z\alpha}{\sigma}\right)}{\sqrt{1-\rho^2}}\right) - \log\Phi(x\beta).$$

One could estimate this model by maximizing the sample sum of the full likelihood function l, by maximizing the sample sum of either the marginal or the conditional component, or by maximizing these components in sequence. Note that asymptotically efficient estimation requires maximizing the full likelihood, and that not all the parameters are identified in each component; e.g., only β is identified from the marginal component. Nevertheless, there may be computational advantages to working with the marginal or conditional likelihood, at least in the first step of estimation. Maximization of l^1 is a conventional binomial probit problem, which can be done easily with many canned programs. Maximization of l^2 could be done either jointly in all the parameters α, β, ρ, σ; or alternately in α, ρ, σ, with the estimate of β from a first-step binomial probit substituted in and treated as fixed. The first case, maximization of l^2 in all the parameters, provides estimates whose variances are estimated by the inverse of the information matrix for l^2. The maximization of l^2 with an estimate of β substituted in requires use of the formula for the variance of a GMM estimator containing an embedded estimator; see the lecture notes on this topic. Neither of these procedures is fully efficient, and the two methods cannot be ranked in terms of efficiency.

When $\rho = 0$, the case of "exogenous" selection in which there is no correlation between the random variables determining selection into the observed population and the level of the observation, note that l^2 reduces to the log likelihood for a regression with normal disturbances, implying that the maximum likelihood estimates for α and σ will be the OLS estimates. However, when $\rho \neq 0$, selection matters and regressing of w on z will <u>not</u> give consistent estimates of α and σ.

An alternative to maximum likelihood estimation is a GMM procedure based on the moments of w. Using the property that the conditional expectation of ν given $y = 1$ equals the conditional expectation of ν given ε, integrated over the conditional density of ε given $y = 1$, plus the property of the normal that $d\varphi(\varepsilon)/d\varepsilon = -\varepsilon\cdot\varphi(\varepsilon)$, one has

(27) $$\mathbf{E}\{w|z,y=1\} = z\alpha + \sigma\mathbf{E}\{\nu|y=1\} = z\alpha + \sigma\int_{-x\beta}^{+\infty}\mathbf{E}\{\nu|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= z\alpha + \sigma\rho\int_{-x\beta}^{+\infty}\varepsilon\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= z\alpha + \sigma\rho\varphi(x\beta)/\Phi(x\beta) \equiv z\alpha + \lambda M(x\beta),$$

where $\lambda = \sigma\rho$ and $M(c) = \varphi(c)/\Phi(c)$ is called the <u>inverse Mill's ratio</u>. (As a computational note, it is much better when calculating M to use a direct approximation to this function, rather than taking the ratio of computational approximations to φ and Φ.) Further, using the relationship

$$\mathbf{E}(\nu^2|\varepsilon) = \mathrm{Var}(\nu|\varepsilon) + \{\mathbf{E}(\nu|\varepsilon)\}^2 = 1 - \rho^2 + \rho^2\varepsilon^2,$$

and the integration-by-parts formula

$$\int_{-c}^{+\infty} \varepsilon^2\varphi(\varepsilon)d\varepsilon = -\int_{-c}^{+\infty} \varepsilon\varphi'(\varepsilon)d\varepsilon = -c\varphi(c) + \int_{-c}^{+\infty} \varphi(\varepsilon)d\varepsilon = -c\varphi(c) + \Phi(c),$$

one obtains

$$(28)\qquad \mathbf{E}\{(w-z\alpha)^2|z,y=1\} = \sigma^2\mathbf{E}\{v^2|y=1\} = \sigma^2 \int_{-x\beta}^{+\infty} \mathbf{E}\{v^2|\varepsilon\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta)$$

$$= \sigma^2 \int_{-x\beta}^{+\infty} \{1 - \rho^2 + \rho^2\varepsilon^2\}\varphi(\varepsilon)d\varepsilon/\Phi(x\beta) = \sigma^2\{1 - \rho^2 + \rho^2 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta)\}$$

$$= \sigma^2\{1 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta)\} = \sigma^2\{1 - \rho^2 x\beta\cdot M(x\beta)\}.$$

Then,

$$(29)\quad \mathbf{E}\left\{[w - z\alpha - \boldsymbol{E}\{w-z\alpha|z,y=1\}]^2|z,y=1\right\} = \mathbf{E}\{(w-z\alpha)^2|z,y=1\} - [\mathbf{E}\{w-z\alpha|z,y=1\}]^2$$

$$= \sigma^2\{1 - \rho^2 x\beta\varphi(x\beta)/\Phi(x\beta) - \rho^2\varphi(x\beta)^2/\Phi(x\beta)^2\}$$

$$= \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)\}.$$

It is possible to go on and compute higher moments, using the recursion formula:

$$\mu(c,k,\lambda) \equiv \mathbf{E1}(\varepsilon>c)\cdot(\varepsilon-\lambda)^k = \int_{\varepsilon=c}^{\infty} (\varepsilon-\lambda)^k\varphi(\varepsilon)d\varepsilon$$

$$= -(c-\lambda)^{k-1}\varphi(c) - \lambda\cdot\mu(c,k-1,\lambda) + (k-1)\cdot\mu(c,k-2,\lambda).$$

A GMM estimator for this problem can be obtained by applying NLLS, for the observations with y = 1, to the equation

$$(30)\qquad w = z\alpha + \sigma\rho M(x\beta) + \zeta,$$

where ζ is a disturbance that satisfies $\mathbf{E}\{\zeta|y=1\} = 0$. This ignores the heteroskedasticity of ζ, but it is nevertheless consistent. This regression estimates only the product $\lambda \equiv \sigma\rho$, but consistent estimates of σ and ρ could be obtained in a second step: The formula for the variance of ζ,

$$(31)\qquad \mathbf{V}\{\zeta|x,z,y=1\} = \sigma^2\{1 - \rho^2 M(x\beta)[x\beta + M(x\beta)]\},$$

suggests obtaining an estimate of σ^2 by regressing the square of the estimated residual, ζ_e^2, on one and the variable $M(x\beta_e)[x\beta_e + M(x\beta_e)]$, where β_e is the estimated parameter vector. Then, the estimated coefficients *a* and *b* in the regression

$$(32)\qquad \zeta|^2 = a + b\{M(x\beta_e)[x\beta_e + M(x\beta_e)]\} + \xi$$

provide consistent estimates of σ^2 and $\sigma^2\rho^2$, respectively.

The GMM estimator above is asymptotically inefficient because it fails to correct for heteroskedasticity, but more fundamentally because there are common parameters between the regression and the variance of the disturbances, and because the disturbance ζ is not normally distributed, so there is information in moments beyond the first two. The first of these inefficiencies could be eliminated by an estimated GLS-type transformation: From the first-step NLLS regression and the estimator of σ described above, calculate the weight

$$\tau^2 = 1 - \rho|^2 M(x\beta_e)[x\beta_e + M(x\beta_e)],$$

and then rerun a weighted NLLS regression,

$$w/\tau = (z/\tau_e)\alpha + \sigma\rho(M(x\beta_e)/\tau_e) + (\zeta/\tau_e). \tag{33}$$

The variance of this regression is now σ^2, so that all the parameters of the original problem are estimated by the regression parameters plus the estimated variance of the regression.

The NLLS estimator above involves about the same amount of calculation as full maximum likelihood estimation, so that the latter method is usually preferable because it is asymptotically efficient, and the standard errors obtained from the information matrix are easier to calculate than the two-step GLS standard errors. However, there is an alternative two-step estimation procedure, due to Heckman, that requires only standard computer software, and is widely used:

[1] Estimate the binomial probit model,

$$P(y|x,\beta) = \Phi(yx\beta) \ , \tag{34}$$

by maximum likelihood.

[2] Estimate the linear regression model,

$$w = z\alpha + \lambda M(x\beta_e) + \zeta, \tag{35}$$

where $\lambda = \sigma\rho$ and the inverse Mill's ratio M is evaluated at the parameters estimated from the first stage.

To estimate σ and ρ, and increase efficiency, one can do two additional steps,

[3] Estimate σ^2 using the procedure described in (12), with estimates λ_e from the second step and β_e from the first step; and

[4] Estimate the weighted linear regression model

$$w/\tau = (z/\tau)\alpha + \lambda M(x\beta_e)/\tau + (\zeta/\tau), \tag{36}$$

where

$$\tau^2 = \{1 - \rho_e{}^2 M(x\beta_e)[x\beta_e + M(x\beta_e)]\},$$

and the parameters in this weight come from the first and second steps, plus

$$\rho_e^{\,2} = \lambda_e^{\,2}/\sigma_e^{\,2}$$

with $\lambda_e^{\,2}$ from step two and $\sigma_e^{\,2}$ from step three.

The standard errors of the first-step estimates β_e are obtained from the binomial probit maximum likelihood. However, the second-step estimates α_e and λ_e have standard errors that are not given correctly by the regression (35), both because the errors are heteroskedastic and because a right-hand-side variable contains embedded parameters from an earlier step; see the lecture notes on GMM estimation with embedded estimates for the formulas for the correct standard errors.

One limitation of the bivariate model is most easily seen by examining the regression (35). Consistent estimation of the parameters α in this model requires that the term $M(x\beta|)$ be estimated consistently. This in turn requires the assumption of normality, leading to the first-step probit model, to be exactly right. Were it not for this restriction, estimation of α in (35) would be consistent under the much more relaxed requirements for consistency of OLS estimators. To investigate this issue further, consider the bivariate selection model (21) with the following more general distributional assumptions: (i) ε has a density $f(\varepsilon)$ and associated CDF $F(\varepsilon)$; and (ii) ν has $\mathbf{E}(\nu|\varepsilon) = \rho\varepsilon$ and a second moment $\mathbf{E}(\nu^2|\varepsilon) = 1 - \rho^2$ that is independent of ε. Define the truncated moments

$$J(x\beta) = \mathbf{E}(\varepsilon|\varepsilon>-x\beta) = \int_{-x\beta}^{\infty} \varepsilon f(\varepsilon)d\varepsilon/[1 - F(-x\beta)]$$

and

$$K(x\beta) = \mathbf{E}(1 - \varepsilon^2|\varepsilon>-x\beta) = \int_{-x\beta}^{\infty} [1 - \varepsilon^2]f(\varepsilon)d\varepsilon/[1 - F(-x\beta)]\,.$$

Then, given the assumptions (i) and (ii),

$$\mathbf{E}(w|z,y=1) = z\alpha + \sigma\rho\mathbf{E}(\varepsilon|\varepsilon>-x\beta) = z\alpha + \sigma\rho J(x\beta),$$

$$\mathbf{E}((w - \mathbf{E}(w|z,y=1))^2|z,y=1) = \sigma^2\{1 - \rho^2[K(x\beta) + J(x\beta)^2]\}.$$

Thus, even if the disturbances in the latent variable model were not normal, it would nevertheless be possible to write down a regression with an added term to correct for self-selection that could be applied to observations where y = 1:

$$(37) \qquad w = z\alpha + \sigma\mathbf{E}\{\nu|x\beta+\varepsilon>0\} + \zeta = z\alpha + \sigma\rho J(x\beta) + \zeta,$$

where ζ is a disturbance that has mean zero and the heteroskedastic variance

$$\mathbf{E}(\zeta^2|z,y=1)) = \sigma^2\{1 - \rho^2[K(x\beta) + J(x\beta)^2]\}.$$

Now suppose one runs the regression (30) with an inverse Mill's ratio term to correct for self-selection, when in fact the disturbances are not normal and (36) is the correct specification. What bias results? The answer is that the closer $M(x\beta)$ is to $J(x\beta)$, the less the bias. Specifically, when (36) is the correct model, regressing w on z and $M(x\beta)$ amounts to estimating the misspecified model

$$w = z\alpha + \lambda M(x\beta) + \{\zeta + \lambda[J(x\beta) - M(x\beta)]\} .$$

The bias in NLLS is given by

$$\begin{bmatrix} \hat{\alpha} - \alpha \\ \lambda_e - \lambda \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{Ez}'z & \boldsymbol{Ez}'M \\ \boldsymbol{EMz} & \boldsymbol{EM}^2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Ez}(J-M) \\ \boldsymbol{EM}(J-M) \end{bmatrix} ;$$

this bias is small if $\lambda = \sigma\rho$ is small or the covariance of J - M with z and M is small.

Calculation for some standard distributions shows that when disturbances deviate from normal, M may not be a good approximation to J, implying that bias due to misspecification can be substantial. For example, consider as alternatives to the normal density for ε the logistic density,

$$f(\varepsilon) = e^{-a\varepsilon}/(1+e^{-a\varepsilon})^2, \; a = \frac{\sqrt{3}}{\pi} ,$$

and the bilateral exponential density,

$$f(\varepsilon) = (1/2\sqrt{2})\cdot e^{-|\varepsilon|/\sqrt{2}} .$$

For these densities, the function J can be calculated analytically. For the logistic density, one obtains $J(\varepsilon) = -x + (1/a)\cdot\log(1+e^{a\varepsilon})\cdot(1+e^{-a\varepsilon})$, and for the bilateral exponential density, one obtains $J(\varepsilon) = e^{-c|\varepsilon|}\cdot(1 + c|\varepsilon|)/2cF(\varepsilon)$, where $F(\varepsilon) = \mathbf{1}(\varepsilon<0)\cdot e^{c\varepsilon} + \mathbf{1}(\varepsilon\geq 0)\cdot(1 - e^{-c\varepsilon})$ and $c^{-1} = \sqrt{2}$. The J(·) functions have the same qualitative shape for the normal, bilateral exponential, and logistic densities, but they are substantially shifted, so that there is at least significant bias to the estimated intercept in the regression if J is misspecified.

A natural question in semiparametric estimation is whether there is a robust method for estimating α that does not require that the distributions of ε and ν be fully parametric. It should be clear intuitively that approximating the unknown true J(·) function by a series of functions of ε, such as a low order polynomial in ε, should be sufficient to approximately span the space containing J(·), and that this in turn would be sufficient to eliminate for practical purposes any bias in estimation of α. The question would remain at to how many terms to use in an approximation.

REFERENCES

Cosslett, S. (1981) "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*; 49(5), 1289-1316.

Cox, D. and Hinckley, D (1974) *Theoretical Statistics*, London: Chapman and Hall.

Goldfeld, S.; Quandt, R. (1973) "A Markov Model for Switching Regressions," *Journal-of-Econometrics*; 1(1), 3-15.

Goldfeld, S.; Quandt, R. (1975) "Estimation in a Disequilibrium Model and the Value of Information," *Journal-of-Econometrics*; 3(4), 325-48.

Hausman, J.; Wise, D. (1977) " Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*; 45(4),919-38.

Heckman, J. (1974) "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*; 42(4), 679-94.
Hsieh, D.; Manski, C.; McFadden, D. (1985) " Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal-of-the-American-Statistical-Association*; 80(391),651-62.
Lancaster, T.; Imbens, G. (1990) "Choice-Based Sampling of Dynamic Populations," in Hartog, J.; Ridder, G.; Theeuwes, J., eds. *Panel data and labor market studies*. Contributions to Economic Analysis, vol. 192, Amsterdam: North-Holland, 21-43.
Lee, L. F.; Porter, R. (1984) "Switching Regression Models with Imperfect Sample Separation Information-With an Application on Cartel Stability," *Econometrica*; 52(2), 391-418.
Maddala, G. S.; Nelson, F. (1974) "Maximum Likelihood Methods for Models of Markets in Disequilibrium," *Econometrica*; 42(6), 1013-30.
Manski, C.; Lerman, S. (1977) "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*; 45(8), 1977-88.$_e$
Manski, C. and D. McFadden (1981) "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in C. Manski and D. McFadden (eds) Structural *Analysis of Discrete Data*, Cambridge: MIT Press, 2-49.
McFadden, D. (1996) "On the Analysis of 'Intercept & Follow' Surveys," University of California Berkeley working paper.
Newey, W. and D. McFadden (1995) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden, eds. *Handbook of Econometrics*, Vol. 4, Amsterdam: North Holland, 2113-2247.

CHAPTER 3. GENERALIZED METHOD OF MOMENTS

1. INTRODUCTION

This chapter outlines the large-sample theory of Generalized Method of Moments (GMM) estimation and hypothesis testing. The properties of consistency and asymptotic normality (CAN) of GMM estimates hold under regularity conditions much like those under which maximum likelihood estimates are CAN, and these properties are established in essentially the same way. Further, the trinity of Wald, Lagrange Multiplier, and Likelihood Ratio test statistics from maximum likelihood estimation extend virtually unchanged to this more general setting. Our treatment provides a unified framework that specializes to both classical maximum likelihood methods and traditional linear models estimated on the basis of orthogonality restrictions.

Suppose data z are generated by a process that is parameterized by a k×1 vector θ. Let $l(z,\theta)$ denote the log likelihood of z, and let θ_o denote the true value of θ in the population. Suppose there is an m×1 vector of functions of z and θ, denoted $g(z,\theta)$, that have zero expectation in the population if and only if θ equals θ_o:

$$\mathbf{E}g(z,\theta) \equiv \int g(z,\theta)\cdot\exp(l(z,\theta_o))dz = 0 \text{ iff } \theta = \theta_o.$$

The $\mathbf{E}\, g(z,\theta)$ are *generalized moments*, and the analogy principle suggests that an estimator of θ_o can be obtained by solving for θ that makes the sample analogs of the population moments small. Identification normally requires that $m \geq k$. If the inequality is strict, and the moments are not degenerate, then there are *over-identifying* moments that can be used to improve estimation efficiency and/or test the internal consistency of the model.

In this setup, there are several alternative interpretations of z. It may be the case that z is a complete description of the data and $l(z,\theta)$ is the "full information" likelihood. Alternately, some components of observations may be margined out, and $l(z,\theta)$ may be a marginal "limited information" likelihood. Examples are the likelihood for one equation in a simultaneous equations system, or the likelihood for continuous observations that are classified into discrete categories. Also, there may be "exogenous" variables (covariates), and the full or limited information likelihood above may be written conditioning on the values of these covariates. From the standpoint of statistical analysis, variables that are conditioned out behave like constants. Then, it does not matter for the discussion of hypothesis testing that follows which interpretation above applies, except that when regularity conditions are stated it should be understood that they hold almost surely with respect to the distribution of covariates.

Several special cases of this general setup occur frequently in applications: First, if $l(z,\theta)$ is a full or limited information likelihood function, and $g(z,\theta) = \nabla_\theta l(z,\theta)$ is the score vector, then we obtain maximum likelihood estimation. Second, if $z = (y,x,w)$ and $g(z,\theta) = w'(y-x\theta)$ asserts orthogonality in the population between *instruments* w and regression *disturbances* $\varepsilon = y - x\theta_o$, then GMM specializes to 2SLS, or in the case that w = x, to OLS. These linear regression setups generalize immediately to nonlinear regression orthogonality conditions based on the form $g(z,\theta) = w'(y-h(x,\theta))$, where h is a function that is known up to the parameter θ. The last problem can be

interpreted as coming from a non-linear regression model where by assumption a vector of m exogenous variables w are orthogonal to the regression disturbances $y - h(x,\theta_o)$. This is an important application of GMM, and as an exercise the reader should translate all of the more abstract statements about GMM estimators into statements for this model.

Suppose an i.i.d. sample $z_1,...,z_n$ from the data generation process. A GMM estimator of θ_o is the vector T_n that minimizes the generalized distance of the sample moments from zero, where this generalized distance is defined by the quadratic form

$$Q_n(\theta) = (1/2)g_n(\theta)'W_n g_n(\theta), \quad \text{where} \quad g_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^{n} g(z_t,\theta),$$

and W_n is a m×m positive definite symmetric matrix that defines a "distance metric". When m = k, the matrix W_n does not enter the first-order-conditions for T_n (Verify), and could by default be the m×m identity matrix. When m > k, not all the components of $g_n(T_n)$ can be made zero simultaneously, and the matrix W_n determines how deviations from zero are weighted and influences the estimator. Define the m×m covariance matrix of the moments, $\Omega \equiv \mathbf{E}\, g(z,\theta_o)g(z,\theta_o)'$. Efficient weighting of a given set of m moments requires that W_n converge to Ω^{-1} as $n \to \infty$. Exercise 1 below asks you to verify this statement. A good candidate for W_n is $\Omega_n(\tau_n)^{-1}$, where

$$\Omega_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} g(z_t,\theta)g(z_t,\theta)',$$

and τ_n is a consistent preliminary estimate of θ_o. Define the m×k Jacobean matrix $G \equiv \mathbf{E}\, \nabla_\theta g(z,\theta_o)$, and let

$$G_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \nabla_\theta g(z_t,\theta).$$

Then the array $G_n(\tau_n)$ evaluated at a consistent preliminary estimate τ_n of θ_o will approach G as $n \to \infty$. Hereafter, Ω_n and G_n will be used as shorthand for $\Omega_n(\tau_n)$ and $G_n(\tau_n)$, respectively.

We will denote convergence in probability by $\to_p$, almost sure convergence by $\to_{as}$, and convergence in distribution by $\to_d$. The following regularity conditions guarantee that GMM estimators have good asymptotic properties; see Newey and McFadden (1994):

(i) The domain Θ of θ is compact, and θ_o is in its interior.
(ii) The log likelihood function $l(z,\theta)$ is almost surely in z continuously differentiable with respect to θ in a neighborhood of θ_o.
(iii) The function g is measurable in z for each θ, and almost surely is continuous and continuously differentiable in θ, with the derivative Lipschitz; i.e., there is a function $\alpha(z)$ with finite expectation such that for $\theta,\theta' \in \Theta$, $|\nabla_\theta g(z,\theta) - \nabla_\theta g(z,\theta')| \le \alpha(z)|\theta - \theta'|$.
(iv) $\mathbf{E}g(z,\theta) = 0$ if and only if $\theta = \theta_o$.
(v) Ω is a positive definite m×m matrix, and $\Omega_n \to_p \Omega$.
(vi) G is a m×k matrix of rank k, and $G_n \to_p G$.
(vii) There exists a function $\alpha(z)$, with finite expectation, that dominates $g(z,\theta)g(z,\theta)'$ and $\nabla_\theta g(z,\theta)$; i.e., $+\infty > \mathbf{E}\alpha(z)$, $|g(z,\theta)g(z,\theta)'| \le \alpha(z)$, and $|\nabla_\theta g(z,\theta)| \le \alpha(z)$.

Under these regularity conditions, Newey and McFadden (1994, Theorems 2.6 and 3.4) show that the *unconstrained* GMM estimator

$$T_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$$

is consistent and asymptotically normal (CAN), with

$$n^{1/2}(T_n - \theta_o) \rightarrow_d N(0, B^{-1});$$

where $B \equiv G'\Omega^{-1}G$.

It is useful to summarize the steps that lead to the CAN result. First consider consistency. For each fixed θ, a law of large numbers implies that $g_n(\theta) \rightarrow_p \mathbf{E}g(\theta)$. Similarly, $G_n(\theta) \rightarrow_p \mathbf{E}\nabla_\theta g(z,\theta)$ and $\Omega_n(\theta) \rightarrow_p \mathbf{E}\, g(z,\theta)g(z,\theta)'$. Using the compactness of Θ and the smoothness and dominance assumptions, these probability limits can be shown to hold uniformly in θ. This implies that if $\tau_n \rightarrow_p \theta_o$, then $g_n(\tau_n) \rightarrow_p \mathbf{E}g(\theta_o) = 0$, $G_n(\tau_n) \rightarrow_p G$, $\Omega_n(\tau_n) \rightarrow_p \Omega$, and $Q_n(\theta) \rightarrow_p (1/2)(\mathbf{E}g(\theta))'\Omega^{-1}(\mathbf{E}g(\theta))$. By construction, $Q_n(T_n) \leq Q_n(\theta_o) \rightarrow_p 0$. Outside a specified small neighborhood of θ_o, the probability limit of Q_n is uniformly bounded away from zero. Therefore, T_n is a.s. eventually within the specified neighborhood. This establishes consistency.

Next consider asymptotic normality. A central limit theorem implies

$$-\Omega^{-1/2} n^{1/2} g_n(\theta_o) \equiv U_n \rightarrow_d U \sim N(0,I). \tag{1}$$

The mean value theorem applied to the sample moments about θ_o gives

$$n^{1/2} g_n(\theta) = n^{1/2} g_n(\theta_o) + G_n n^{1/2}(\theta - \theta_o), \tag{2}$$

with G_n evaluated at points between θ and θ_o. Substituting this expression in the GMM first-order condition $0 = n^{1/2}\nabla_\theta Q_n(T_n) \equiv G_n'\Omega_n^{-1} n^{1/2} g_n(T_n)$ and using the consistency of T_n to replace G_n and Ω_n by their respective asymptotic approximations G and Ω, yields

$$0 = -G'\Omega^{-1/2}U_n + B\, n^{1/2}(T_n - \theta_o) + o_p\,,$$

where o_p denotes terms that are asymptotically negligible, implying

$$n^{1/2}(T_n - \theta_o) = B^{-1}G'\Omega^{-1/2}U_n + o_p \rightarrow_d B^{-1}G'\Omega^{-1/2}U \sim N(0,B^{-1})\,. \tag{3}$$

The asymptotic covariance matrix B^{-1} can be estimated using $G_n(\tau_n)$ and $\Omega_n(\tau_n)$, where τ_n is any $n^{1/2}$-consistent (preliminary) estimator of θ_o (i.e., $n^{1/2}(\tau_n - \theta_o)$ is stochastically bounded.) A practical procedure for estimation is to first estimate θ_o using the GMM criterion with an arbitrary Ω_n, such as an m×m identity matrix. This produces an initial $n^{1/2}$-consistent estimator τ_n. Then use the formulas above to estimate the asymptotically efficient $W_n = \Omega_n(\tau_n)^{-1}$, and use the GMM criterion with this distance metric to obtain the final estimator T_n.

Differentiating the identity $0 \equiv \int g(z,\theta)e^{l(z,\theta)}dz$, one has

$$0 \equiv \int \nabla_\theta g(z,\theta) \exp(l(z,\theta))dz + \int g(z,\theta)\nabla_\theta l(z,\theta)' \exp(l(z,\theta))dz,$$

implying at θ_o that

$$\Gamma \equiv -\mathbf{E}g(z,\theta_o)\nabla_\theta l(z,\theta_o)' \equiv \mathbf{E}\nabla_\theta g(z,\theta_o) \equiv G.$$

It will sometimes be convenient to estimate G by

$$\Gamma_n = -\frac{1}{n}\sum_{t=1}^{n} g(z_t,\tau_n)\nabla_\theta l(z_t,\tau_n)'.$$

In the maximum likelihood case $g = \nabla_\theta l$, one has $\Omega = \Gamma = G$, and the asymptotic covariance matrix of the unconstrained estimator simplifies to Ω^{-1}.

Exercise 1. Use a Taylor's expansion of the first-order-conditions for minimization of $Q_n(\theta)$ to show that when W_n converges to a matrix W other than Ω^{-1}, the resulting GMM estimator T_n is asymptotically normal with covariance matrix $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$. Show that a quadratic form in this matrix is minimized when $W = \Omega^{-1}$. (Hint: Consider a regression with m observations and k parameters, $y = G\beta + \nu$, that has $\mathbf{E}\nu\nu' = \Omega$. Then OLS applied to the transformed data $\Omega^{-1/2}y = \Omega^{-1/2}G\beta + \Omega^{-1/2}\nu$ is BLUE, and OLS applied to any other transformation $W^{1/2}y = W^{1/2}G\beta + W^{1/2}\nu$ yields estimates of β that have a larger covariance matrix.)

2. THE NULL HYPOTHESIS AND THE CONSTRAINED GMM ESTIMATOR

Suppose there is an r-dimensional null hypothesis on the data generation process,

$$H_o: a(\theta_o) = 0,$$

where $a(\cdot)$ is a r×1 vector of continuously differentiable functions. Assume that the r×k matrix $A \equiv \nabla_\theta a(\theta_o)$ has rank r. We will consider alternatives to the null of the form

$$H_1: a(\theta_o) \neq 0,$$

or *asymptotically local* alternatives of the form

$$H_{1n}: a(\theta_o) = \delta n^{-1/2} \neq 0.$$

These alternatives are of interest because in large samples alternative hypotheses of interest are often sufficiently "local" so that the asymptotic approximation will give good estimates of the power of tests. The null hypothesis may be linear or nonlinear. A particularly simple case is $H_o: \theta = \theta^o$, or $a(\theta) \equiv \theta - \theta^o$, so the parameter vector θ is completely specified under the null. Other examples are $a(\theta_o) = \theta_{1o}$, a linear hypothesis, and $a(\theta_o) = (\theta_{10}/\theta_{20} - \theta_{30}/\theta_{40})$, a non-linear hypothesis. In general there will be k-r parameters to be estimated when one imposes the null. One can define a *constrained* GMM estimator by optimizing the GMM criterion subject to the null hypothesis:

$$T_{an} = \text{argmin}_{\theta\in\Theta} Q_n(\theta) \quad \text{subject to} \quad a(\theta) = 0.$$

Newey and McFadden (1994, Theorem 9.1) establish that T_{an} is consistent under the regularity conditions above when either the null hypothesis or an asymptotically local alternative to the null holds.

Define a Lagrangian for T_{an}: $L_n(\theta,\gamma) = Q_n(\theta) + a(\theta)'\gamma$. In this expression, γ is the r×1 vector of undetermined Lagrangian multipliers; these will be non-zero when the constraints are binding. The first-order conditions for solution of this problem are

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} n^{1/2}\nabla_\theta Q_n(T_{an}) + \nabla_\theta a(T_{an})' n^{1/2}\gamma_{an} \\ n^{1/2} a(T_{an}) \end{bmatrix}.$$

The Lagrangian multipliers γ_{an} are random variables with an asymptotic distribution: The consistency of T_{an} implies $\nabla_\theta Q_n(T_{an}) \to_p G'\Omega^{-1}\mathbf{E}g(z,\theta_o) = 0$. Further, $\nabla_\theta a(T_{an}) \to_p A$, implying $A'\gamma_{an} = -\nabla_\theta Q_n(T_{an}) + o_p \to_p 0$, and since A is of full rank, $\gamma_{an} \to_p 0$. The following paragraph outlines the argument for asymptotic normality, and relates the asymptotic distributions of T_n, T_{an}, and γ_{an}. The asymptotic normality argument parallels that already given in (1)-(3) for the unconstrained estimator. Using the mean value theorem and then approximating G_n by G and Ω_n by Ω, one has

$$n^{1/2}g_n(T_{an}) = n^{1/2}g_n(\theta_o) + G_n\, n^{1/2}(T_{an}-\theta_o) = -G'\Omega^{-1/2}U_n + G\, n^{1/2}(T_{an}-\theta_o) + o_p,$$

and

$$n^{1/2}a(T_{an}) = n^{1/2}a(\theta_o) + A\, n^{1/2}(T_{an}-\theta_o) + o_p \equiv \delta + A\, n^{1/2}(T_{an}-\theta_o) + o_p.$$

Substituting these in the first-order conditions yields

$$\text{(4)} \qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G'\Omega^{-1/2}U_n \\ -\delta \end{bmatrix} - \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} n^{1/2}(T_{an} - \theta_o) \\ n^{1/2}\gamma_{an} \end{bmatrix} + o_p.$$

From the formula for partitioned inverses,

$$\text{(5)} \qquad \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1/2}MB^{-1/2} & B^{-1}A'(AB^{-1}A')^{-1} \\ (AB^{-1}A')^{-1}AB^{-1} & (AB^{-1}A')^{-1} \end{bmatrix},$$

where $M = I - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$ is a k×k idempotent matrix of rank k-r. Applying this to (4) yields

$$\text{(6)} \qquad \begin{bmatrix} n^{1/2}(T_{an} - \theta_o) \\ n^{1/2}\gamma_{an} \end{bmatrix} = \begin{bmatrix} -B^{-1}A'(AB^{-1}A')^{-1} \\ -(AB^{-1}A')^{-1} \end{bmatrix} \delta + \begin{bmatrix} B^{-1/2}MB^{-1/2} \\ (AB^{-1}A')^{-1}AB^{-1} \end{bmatrix} G'\Omega^{-1/2}U_n + o_p.$$

Then, the asymptotic distribution of $n^{1/2}(T_{an}-\theta_o)$ under a local alternative, or the null when $\delta = 0$, is $N(-B^{-1}A'(AB^{-1}A')^{-1}\delta, B^{-1/2}MB^{-1/2})$.

Writing out $M = I - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$ yields

$$(7) \qquad n^{1/2}(T_{an}-\theta_o) = B^{-1}G'\Omega^{-1/2}U_n - B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n - B^{-1}A'(AB^{-1}A')^{-1}\delta + o_p.$$

The first term on the right-hand-side of (7) and the right-hand-side of (3) are identical, to order o_p. Then, they can be combined to conclude that

$$(8) \qquad n^{1/2}(T_n-T_{an}) = B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + B^{-1}A'(AB^{-1}A')^{-1}\delta + o_p,$$

so that $n^{1/2}(T_n-T_{an})$ is asymptotically normal with mean $B^{-1}A'(AB^{-1}A')^{-1}\delta$ and covariance matrix $B^{-1/2}(I-M)B^{-1/2} \equiv B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$. Note that the asymptotic covariance matrices satisfy $acov(T_n-T_{an}) = acov(T_n) - acov(T_{an})$, or *the variance of the difference equals the difference of the variances*. This proposition is familiar in a maximum likelihood context where the variance in the deviation between an efficient estimator and any other estimator equals the difference of the variances. We see here that it also applies to *relatively* efficient GMM estimators that use available moments and constraints optimally.

The results above and some of their implications are summarized in Table 1. Each statistic is distributed as a linear transformation of a common random vector U_n that is asymptotically standard normal. Recall that $B = G'\Omega^{-1}G$ is a positive definite k×k matrix, and let $B^{-1} \equiv acov(T_n)$. Recall that $M = I - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$ is a k×k idempotent matrix of rank k-r.

Table 1

Statistic	Formula	Asymptotic Covariance Matrix
$n^{1/2}(T_n-\theta_o)$	$B^{-1}G'\Omega^{-1/2}U_n + o_p$	B^{-1}
$n^{1/2}(T_{an}-\theta_o)$	$-B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1/2}MB^{-1/2}G'\Omega^{-1/2}U_n + o_p$	$B^{-1/2}MB^{-1/2}$
$n^{1/2}(T_n-T_{an})$	$B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + o_p$	$B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$
$n^{1/2}\gamma_{an}$	$(AB^{-1}A')^{-1}\delta + (AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + o_p$	$(AB^{-1}A')^{-1}$
$n^{1/2}a(T_n)$	$\delta + AB^{-1}G'\Omega^{-1/2}U_n + o_p$	$AB^{-1}A'$
$n^{1/2}\nabla_\theta Q_n(T_{an})$	$A'(AB^{-1}A')^{-1}\delta + A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U_n + o_p$	$A'(AB^{-1}A')^{-1}A$

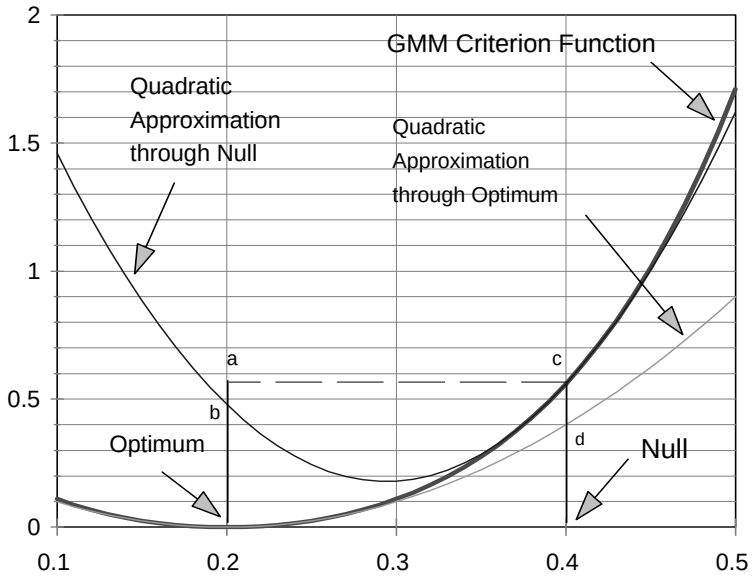
3. THE TEST STATISTICS

The test statistics for the null hypothesis fall into three major classes, sometimes called the *trinity*. *Wald statistics* are based on deviations of the unconstrained estimates from values consistent with the null. *Lagrange Multiplier* (LM) or *Score statistics* are based on deviations of the constrained estimates from values solving the unconstrained problem. *Distance metric statistics* are based on differences in the GMM criterion between the unconstrained and constrained estimators. In the case of maximum likelihood estimation, the distance metric statistic is asymptotically

equivalent to the *likelihood ratio statistic*. There are several variants for Wald statistics in the case of the general non-linear hypothesis; these reduce to the same expression in the simple case where the parameter vector is completely determined under the null. The same is true for the LM statistic. There are often significant computational advantages to using one member or variant of the trinity rather than another. On the other hand, they are all *asymptotically equivalent*. Thus, at least to first-order asymptotic approximation, there is no statistical reason to choose between them. This pattern of first-order asymptotic equivalence for GMM estimates is exactly the same as for maximum likelihood estimates.

Figure 1 illustrates the relationship between distance metric (DM), Wald (W), and Score (LM) tests. In the case of maximum likelihood estimation, this figure is inverted, the criterion is log likelihood rather than the distance metric, and the DM test is replaced by the likelihood ratio test.

FIGURE 1.
GMM TESTS



The "Optimum" and "Null" points on the θ axis give the unconstrained (T_n) and constrained (T_{an}) estimators, respectively. The GMM criterion function is plotted, along with quadratic approximations to this function through the respective arguments T_n and T_{an}. The Wald statistic (W) can be interpreted as twice the difference in the height at T_n and T_{an} of the quadratic approximation through the optimum; the height d in the figure. The Lagrange Multiplier (LM) statistic can be interpreted as twice the difference in the height at T_n and T_{an} of the quadratic approximation through the null; the difference a - b in the figure The Distance Metric (DM) statistic is twice the difference in the height at T_n and T_{an} of the GMM criterion, the height c in the figure. Note that if the criterion function were exactly quadratic, then the three statistics would be identical.

The test statistics we consider for the general non-linear hypothesis $a(\theta_o) = 0$ are given in Table 2. In this table, recall that $\text{acov}(T_n) = B$ and $\text{acov}(T_{an}) = B^{-1/2}MB^{-1/2}$. In Section 7, we consider the important special cases, including maximum likelihood and nonlinear least squares. In particular, when the hypothesis is that a subset of the parameters are constants, there are some simplifications of the statistics, and some versions are indistinguishable.

Table 2. Test Statistics

Wald Statistics	
W_{1n}	$na(T_n)'[AB^{-1}A']^{-1}a(T_n)$
W_{2n}	$n(T_n-T_{an})'\{\text{acov}(T_n) - \text{acov}(T_{An})\}^{-}(T_n - T_{an})$ $= n(T_n-T_{an})'A'(AB^{-1}A')^{-1}A(T_n-T_{an})$
W_{3n}	$n(T_n-T_{an})'B(T_n-T_{an})$
Lagrange Multiplier Statistics	
LM_{1n}	$n\gamma_{an}'AB^{-1}A'\gamma_{an}$
LM_{2n}	$n\nabla_\theta Q_n(T_{an})'\{A'(AB^{-1}A')^{-1}A'\}^{-}\nabla_\theta Q_n(T_{an})$ $= n\nabla_\theta Q_n(T_{an})'B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}\nabla_\theta Q_n(T_{an})$
LM_{3n}	$n\nabla_\theta Q_n(T_{an})'B^{-1}\nabla_\theta Q_n(T_{an})$
Distance Metric Statistic	
DM_n	$2n[Q_n(T_{an}) - Q_n(T_n)]$

Newey and McFadden (1994, Theorem 9.2) establish that under the regularity conditions (i) to (vii), the statistics are all asymptotically equivalent under the null hypothesis or under a local alternative, converging in distribution to a chi-square with r degrees of freedom under the null, and converging in distribution to a non-central chi-square with r degrees of freedom and a non-centrality parameter $\delta'(AB^{-1}A')^{-1}\delta$ under local alternatives to the null. These results are obtained by combining the expressions in Table 1. Suppose q is an expression from the table with asymptotic covariance matrix R and an asymptotic mean λ under local alternatives to the null with the property that λ lies in the subspace spanned by R. The Appendix to this chapter shows that the matrix R can be written in the form $R = S^{1/2}TS^{1/2}$, where S is symmetric and positive definite and T is idempotent with rank equal to the rank of R, that the Moore-Penrose generalized inverse of R is $R^{-} = S^{-1/2}TS^{-1/2}$, and that the condition imposed on the mean implies that $T\,S^{-1/2}\lambda = S^{-1/2}\lambda$. The transformation $S^{-1/2}q$ is then asymptotically normal with mean $S^{-1/2}\lambda$ and covariance matrix T, and consequently the statistic $q'S^{-1}q$ is asymptotically distributed noncentral chi-square with r degrees of freedom, and noncentrality parameter $\lambda'S^{-1}\lambda$ under local alternatives to the null. The transformation $TS^{-1/2}q$ has mean $T\,S^{-1/2}\lambda = S^{-1/2}$, and the asymptotic covariance of $(I - T)S^{-1/2}q$ is zero, so that $S^{-1/2}q$ and $TS^{-1/2}q$ are asymptotically equivalent.

To illustrate the argument, consider W_{1n}. Under the local alternative $a(\theta_o) = \delta n^{-1/2}$, row five of Table 1 gives $q = \delta + AB^{-1}G'\Omega^{-1/2}U$ normal with mean δ and a nonsingular r×r covariance matrix $R = AB^{-1}A'$. Then the noncentrality parameter is $\delta'R^{-1}\delta \equiv \delta'(AB^{-1}A')^{-1}\delta$. Similarly, the statistics W_{2n} and W_{3n} are obtained by noting that $q = -B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}U$ is normal with covariance matrix $R = B^{-1}A'(AB^{-1}A')^{-1}AB^{-1} = B^{-1/2}[B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}]B^{-1/2}$, where the matrix in brackets is idempotent of rank r. Then, both $q'R^{-}q$ and $q'S^{-1}q$ are noncentral chi-square

with degrees of freedom r and noncentrality parameter $\delta'(AB^{-1}A')^{-1}\delta$. The first of these expressions is asymptotically equivalent to W_{2n}, and the second to W_{3n}. Similar arguments establish the properties of the LM statistics.

To demonstrate the asymptotic equivalence of DM_n to the earlier statistics, make a Taylor's expansion of the sample moments for T_{an} about T_n, $n^{1/2}g_n(T_{an}) = n^{1/2}g_n(T_n) + G_n\, n^{1/2}(T_{an} - T_n) + o_p$, and substitute this in the expression for DM_n to obtain

$$\begin{aligned} DM_n &= 2n\{Q_n(T_{an}) - Q_n(T_n)\} \\ &= 2\, n^{1/2}(T_{an} - T_n)'G_n'\Omega_n^{-1}\, n^{1/2}g_n(T_n) + n^{1/2}(T_{an} - T_n)'G_n'\Omega_n^{-1}G_n\, n^{1/2}(T_{an} - T_n) + o_p \\ &= n(T_{an}-T_n)'B(T_{an}-T_n) + o_p \equiv W_{3n} + o_p, \end{aligned}$$

with the last equality holding since $G_n'\Omega_n^{-1}\, n^{1/2}g_n(T_n) = 0$.

The Wald statistic W_{1n} asks how close are the unconstrained estimators to satisfying the constraints; i.e., how close to zero is $a(T_n)$? This variety of the test is particularly useful when the unconstrained estimator is available and the matrix A is easy to compute. For example, when the null is that a subvector of parameters equal constants, then A is a selection matrix that picks out the corresponding rows and columns of B^{-1}, and this test reduces to a quadratic form with the deviations of the estimators from their hypothesized values in the wings, and the inverse of their asymptotic covariance matrix in the center. In the special case H_o: $\theta = \theta^o$, one has A = I.

The Wald test W_{2n} is useful if both the unconstrained and constrained estimators are available. Its first version requires only the readily available asymptotic covariance matrices of the two estimators, but for $r < k$ requires calculation of a generalized inverse. Algorithms for this are available, but are often not as numerically stable as classical inversion algorithms because near zero and exact zero characteristic roots are treated very differently. The second version involves only ordinary inverses, and is potentially quite useful for computation in applications.

The Wald statistic W_{3n} treats the constrained estimators *as if they were constants with a zero asymptotic covariance matrix*. This statistic is particularly simple to compute when the unconstrained and constrained estimators are available, as no matrix differences or generalized inverses are involved, and the matrix A need not be computed. The statistic W_{2n} is in general larger than W_{3n} in finite samples, since the center of the second quadratic form is $\text{acov}(T_n)^{-1}$ and the center of the first quadratic form is $\{\text{acov}(T_n) - \text{acov}(T_{an})\}^-$, while the tails are the same. Nevertheless, the two statistics are asymptotically equivalent.

The approach of Lagrange multiplier or score tests is to calculate the constrained estimator T_{an}, and then to base a statistic on the discrepancy from zero at this argument of a condition that would be zero if the constraint were not binding. The statistic LM_{1n} asks how close the Lagrangian multipliers γ_{an}, measuring the degree to which the hypothesized constraints are binding, are to zero. This statistic is easy to compute if the constrained estimation problem is actually solved by Lagrangian methods, and the multipliers are obtained as part of the calculation. The statistic LM_{2n} asks how close to zero is the gradient of the distance criterion, evaluated at the constrained estimator. This statistic is useful when the constrained estimator is available and it is easy to compute the gradient of the distance criterion, say using the algorithm to seek minimum distance estimates. The second version of the statistic avoids computation of a generalized inverse.

The statistic LM_{3n} bears the same relationship to LM_{2n} that W_{3n} bears to W_{2n}. This flavor of the test statistic is particularly convenient to calculate, as it can be obtained by two auxiliary regressions starting from the constrained estimator T_{an}:

a. *Regress* $\nabla_\theta l(z_t,T_{an})'$ *on* $g(z_t,T_{an})$*, and retrieve fitted values* $\nabla_\theta l^*(z_t,T_{an})'$.

b. *Regress* 1 *on* $\nabla_\theta l^*(z_t,T_{an})$*, and retrieve fitted values* $\hat{y}_t$. *Then* $LM_{3n} = \frac{1}{n}\sum_{t=1}^{n} \hat{y}_t^{\,2}$.

For MLE, $g = \nabla_\theta l$ and the first regression is redundant, so that this procedure reduces to OLS.

Another form of the auxiliary regression for computing LM_{3n} arises in the case of non-linear instrumental variable regression. Consider the model $y_t = h(x_t,\theta_o) + \varepsilon_t$ with $\mathbf{E}(\varepsilon_t | w_t) = 0$ and $\mathbf{E}(\varepsilon_t^{\,2} | w_t) = \sigma^2$, where w_t is a vector of instruments. Define $z_t = (y_t,x_t,w_t)$ and $g(z_t,\theta) = w_t[y_t-h(x_t,\theta)]$. Then $Eg(z,\theta_o) = O$ and $\mathbf{E}g(z,\theta_o)g(z,\theta_o)' = \sigma^2\mathbf{E}ww'$. The GMM criterion $Q_n(\theta)$ for this model is

$$\left(\frac{1}{n}\sum_{t=1}^{n} w_t(y_t - h(x_t,\theta))'\right)\left(\frac{1}{n}\sum_{t=1}^{n} w_t w_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} w_t(y_t-h(x_t,\theta))/2\sigma^2;\right.$$

the scalar σ^2 does not affect the optimization of this function. Consider the hypothesis $a(\theta_o) = O$, and let T_{an} be the GMM estimator obtained subject to this hypothesis. One can compute LM_{3n} by the following method:

a. Regress $\nabla_\theta h(x_t,T_{an})$ on w_t, and retrieve the fitted values $\nabla_\theta \hat{h}_t$.
b. Regress the residual $u_t = y_t - h(x_t,T_{an})$ on $\nabla_\theta \hat{h}_t$, and retrieve the fitted values $\hat{u}_t$.

Then $LM_{3n} = n \sum_{t=1}^{n} \hat{u}_t^{\,2} / \sum_{t=1}^{n} u_t^{\,2} \equiv nR^2$, with R^2 the *uncentered* multiple correlation coefficient.

Note that this is not in general the same as the standard R^2 produced by OLS programs, since the denominator of that definition is the sum of squared deviations of the dependent variable about its mean. When the dependent variable has mean zero, the centered and uncentered definitions coincide.

The approach of the distance metric test is based on the discrepancy between the value of the distance metric, evaluated at the constrained estimate, to the minimum attained by the unconstrained estimate. This estimator is particularly convenient when both the unconstrained and constrained estimators can be computed, and the estimation algorithm returns the goodness-of-fit statistics. In the case of linear or non-linear least squares, this is the familiar test statistic based on the sum of squared residuals from the constrained and unconstrained regressions.

4. TWO-STAGE GMM ESTIMATION

A common econometric problem is to do estimation when some parameters have already been estimated from a previous stage, often on the same data. One common case is where the problem contains constructed variables whose construction depended on parameters estimated in a previous round. In general, the use of consistent estimates from a previous round will not cause a problem

with consistency in later stages, but it will add noise to the problem that appears in the asymptotic covariance matrix of the later-stage estimators.

There are a few cases, such as feasible GLS with normal disturbances, where no correction of the asymptotic covariance matrix is needed. This is due in the GLS case to a block diagonality in the information matrix between regression coefficients and parameters in the covariance matrix. There is a simple rule, due to Whitney Newey, for determining whether previous stage estimation will add something to the asymptotic covariance matrix in the current stage: *There will be a contribution if and only if consistency in the first stage is necessary for consistency in the second stage.*

When a correction is required, the following generic GMM framework can be used to establish the form of this correction. Suppose one observes variables (x,y,z), where x is exogenous, and (y,z) are variables whose behavior is being modeled. Let $f(y,z|x,\alpha,\beta)$ be the joint density of the observations, conditioned on x, with parameter vectors α and β. Assume that it can be written

$$f(y,z|x,\alpha,\beta) = f^c(z|x,y,\alpha)f^m(y|x,\alpha,\beta)$$

or

$$f(y,z|x,\alpha,\beta) = f^c(z|x,y,\alpha,\beta)f^m(y|x,\alpha).$$

This is the standard decomposition of a joint density into a conditional density times a marginal density, and the only restriction we are imposing is that we can parameterize (or reparameterize) the problem so that either the conditional density or the marginal density does not depend on the parameter β. This corresponds to the usual situation in two-stage methods, where at the first stage one looks at limited information that involves a subset of the full parameter vector.

One concrete example of this setup is sequential estimation of the parameters in a two-level nested logit model, in which f^c is the likelihood of choice at the lower level, conditioned on choice of a upper level branch, and f^m is the likelihood of choice among the upper level branches. In this application, the model can be parameterized so that upper branch parameters do not appear in f^c. A second concrete example is two-step estimation of the Tobit model, in which y is an indicator for whether the response is zero or positive, z is the quantitative level of the response, f^c is the likelihood of the quantitative response conditioned on whether it is zero or not, and f^m is the likelihood of the indicator. In this example, the problem can be parameterized so that parameters that enter the quantitative response likelihood do not enter the likelihood for the indicator.

Suppose in the first stage one estimates the parameter vector α using moments

$$0 = \mathbf{E}_n h(a_n;x,y,z),$$

where $\mathbf{E}_n$ denotes empirical expectation (or sample average). A necessary condition for consistency is $\mathbf{E}h(\alpha;x,y,z) = 0$ if and only if $\alpha = \alpha_o$. Limited information maximum likelihood: $h(\alpha;x,y,z) = \nabla_\alpha l^c(z|x,y,\alpha)$, where $l^c = \log f^c$; or $h(\alpha;x,y,z) = \nabla_\alpha l^m(y|x,\alpha)$, where $l^m = \log f^m$, is an important case.

Suppose in the second stage one estimates a parameter vector β using moments

$$0 = \mathbf{E}_n g(b_n,a_n;x,y,z),$$

where a_n is inserted from the previous stage. Again, important cases are maximum likelihood: $g(\beta,\alpha;x,y,z) = \nabla_\beta l^m(y|x,\alpha,\beta)$ or $g(\beta,\alpha;x,y,z) = \nabla_\beta l^c(y|z,x,\alpha,\beta)$, with α treated as if it were known. In the first of these cases, the moments g do not depend on z. Whether or not g depends on z turns out to make a substantial difference in the final covariance formula. The case of constructed variables is handled by writing them as functions of the parameters α that enter their construction. The original parameters of the problem may be estimated, perhaps in combination with other parameters, in both the first and second stages. The classification into α and β may require reparameterization. The following rules may help: If first-stage estimates of original parameters are used solely as starting values for second-stage estimation of the same parameters, then classify these as β parameters, as these first-stage estimates are only a computational device and have no influence on the final solution of the second-stage moments. If first stage estimates of original parameters are used for other purposes, such as construction of estimated variables, and are then reestimated in the second stage, then they should appear in <u>both</u> α and β as *separate* parameters. Of course, original parameters estimated only at the first stage go into α, and original parameters estimated only at the second stage go into β.

Make a Taylor's expansion of both the first-stage and the second-stage moment conditions around the true β_o and α_o, and suppress the x,y,z arguments to simplify notation:

$$\begin{bmatrix}0\\0\end{bmatrix} = n^{1/2}\begin{bmatrix}\boldsymbol{E}_n h(\alpha_o)\\ \boldsymbol{E}_n g(\beta_o,\alpha_o)\end{bmatrix} - \begin{bmatrix}A\\B\end{bmatrix} n^{1/2}(a_n - \alpha_o) - \begin{bmatrix}0\\C\end{bmatrix} n^{1/2}(b_n - \beta_o) + o_p,$$

where $A = -\text{plim } \mathbf{E}_n\nabla_\alpha h(\alpha_o)$, $B = -\text{plim } \mathbf{E}_n\nabla_\alpha g(\beta_o,\alpha_o)$, and $C = -\text{plim } \mathbf{E}_n\nabla_\beta g(\beta_o,\alpha_o)$.

The term $n^{1/2}\begin{bmatrix}\boldsymbol{E}_n h(\alpha_o)\\ \boldsymbol{E}_n g(\beta_o,\alpha_o)\end{bmatrix}$ is asymptotically normal, by a central limit theorem, with a covariance matrix $\begin{bmatrix}\Omega_{hh} & \Omega_{hg}\\ \Omega_{gh} & \Omega_{»}\end{bmatrix}$. Solve the first block of equations and substitute them into the second block to obtain

$$0 = n^{1/2}\{\mathbf{E}_n g(\beta_o,\alpha_o) + BA^{-1}\mathbf{E}_n h(\alpha_o)\} - Cn^{1/2}(b_n - \beta_o) + o_p.$$

The term in braces on the right-hand-side of this expression has an asymptotic covariance matrix

$$\Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'.$$

Then, solving for $n^{1/2}(b_n - \beta_o)$, one obtains the result that its asymptotic covariance matrix is

$$C^{-1}\{\Omega_{gg} - BA^{-1}\Omega_{hg} - \Omega_{gh}A'^{-1}B' + BA^{-1}\Omega_{hh}A'^{-1}B'\}C'^{-1}$$

All the terms of this covariance matrix could be estimated from sample analogs, computed at the consistent estimates. The following table summarizes consistent estimators for the various covariance terms; recall that $\mathbf{E}_n$ denotes empirical expectation (sample average):

Matrix	Estimator
C	$-\mathbf{E}_n\nabla_\beta g(b_n,a_n)$
B	$-\mathbf{E}_n\nabla_\alpha g(b_n,a_n)$
A	$-\mathbf{E}_n\nabla_\alpha h(a_n)$
Ω_{hh}	$\mathbf{E}_n h(a_n)h(a_n)'$
Ω_{gh}	$\mathbf{E}_n g(b_n,a_n)h(a_n)'$
Ω_{gg}	$\mathbf{E}_n g(b_n,a_n)g(b_n,a_n)'$

The terms Ω_{gh} and Ω_{hh} add to the asymptotic covariance matrix, relative to the case of α_o known. If B = 0, there is no correction; this is the "block diagonality" case where β can be estimated consistently even if the estimator of α is not consistent. If α is estimated from an *independent* data set, then $\Omega_{gh} = 0$, but one will still need a correction due to the contribution from Ω_{hh}. Also, if g does not depend on z, then $\Omega_{gh} = \mathbf{E}_{y|x}\{g\cdot\mathbf{E}_{z|x,y}h\} = 0$. This is true, in particular, in the case that the second stage estimator is marginal maximum likelihood in which z does not appear and α is treated as given.

The identities $0 \equiv \int\int h\exp(l)dzdy$ and $0 \equiv \int\int g\exp(l)dzdy$ can be differentiated to obtain the conditions

$$A \equiv -\mathbf{E}\nabla_\alpha h = \mathbf{E}h\cdot\nabla_\alpha l\ ,\ B \equiv -\mathbf{E}\nabla_\alpha g = \mathbf{E}g\cdot\nabla_\alpha l\ ,\ C \equiv -\mathbf{E}\nabla_\beta g = \mathbf{E}g\cdot\nabla_\beta l\ .$$

If g does not depend on z, then $\mathbf{E}g\cdot\nabla_\alpha l^c = \mathbf{E}_{y|x}(g\cdot\mathbf{E}_{z|y,x}\nabla_\alpha l^c) = 0$, implying $B = \mathbf{E}g\cdot(\nabla_\alpha l^m)'$. Sample averages of these outer products estimate the corresponding matrices consistently.

Simplification occurs when the first stage is conditional maximum likelihood that does not depend on β, and the second stage is marginal maximum likelihood that treats the first stage parameter estimates as fixed. Then, $A = \mathbf{E}\nabla_\alpha l^c\cdot(\nabla_\alpha l^c)' = \Omega_{hh}$, $B = \mathbf{E}\nabla_\beta l^m\cdot(\nabla_\alpha l^m)'$, $C = \mathbf{E}\nabla_\beta l^m\cdot(\nabla_\beta l^m)' = \Omega_{gg}$, and $\Omega_{hg} = \mathbf{E}\nabla_\alpha l^c(\nabla_\beta l^m)' = 0$, so that the covariance matrix is $C^{-1} + C^{-1}BA^{-1}B'C^{-1}$.

Similarly, when the first stage is marginal maximum likelihood that does not depend on β, and the second stage is conditional maximum likelihood treating α as fixed, one has $A = \mathbf{E}\nabla_\alpha l^m\cdot(\nabla_\alpha l^m)' = \Omega_{hh}$, $B = \mathbf{E}\nabla_\beta l^c\cdot(\nabla_\alpha l^c)'$, $C = \mathbf{E}\nabla_\beta l^c\cdot(\nabla_\beta l^c)' = \Omega_{gg}$, and $\Omega_{hg} = \mathbf{E}\nabla_\alpha l^c(\nabla_\beta l^m)' = 0$, and the covariance matrix $C^{-1} + C^{-1}BA^{-1}B'C^{-1}$.

The terms in these covariance matrix expressions involve sample averages of squares and cross-products of scores (gradients) of first and second stage log likelihoods. These should all be obtainable as intermediate output from a maximum likelihood program, except for terms involving the gradient of the second-stage likelihood with respect to α. The latter would be simple to obtain in a program like TSP, which does automatic analytic differentiation, or could be obtained by numerical differentiation.

EXERCISE: Consider the problem of Heckman two-stage estimation of a Tobit model, $y = x\theta + \sigma\varphi(x\theta/\sigma)/\Phi(x\theta/\sigma) + \zeta$ for $y > 0$, where $\mathbf{E}(\zeta|y > 0\ \&\ x) = 0$, and where the inverse Mills ratio is calculated from a first-stage probit on the same data. Reparameterize $\alpha = \theta/\sigma$ and $\beta = (\theta,\sigma)$. In this case, h in the generic notation is the score of the marginal log likelihood for the probit, which

is influenced only by α, and g is the set of OLS orthogonality conditions, which depend on both α and β through the condition $y = x\theta + \sigma\varphi(x\alpha)/\Phi(x\alpha)$. Work out the corrected asymptotic covariance matrix for θ and σ.

EXERCISE: Consider the two-level nested multinomial logit model, with first stage estimation applied to the lower level of the choice tree, and used to compute summary variables ("inclusive values") that are then treated as variables in the second stage estimation.

5. ONE-STEP THEOREMS

Under standard regularity conditions, GMM estimators are *locally linear*, which means that within a suitable neighborhood of the estimator, the first-order conditions for these estimators are in large samples approximately linear, with higher-order terms being asymptotically negligible. This has an important practical implication: if one can get an initial estimator τ_n that is within the suitable neighborhood, then one can get to the full GMM estimator, or at least an asyptotically equivalent flavor of it, in one linear step. This has the computational advantage that at this stage no iterative computation is required, and the step can usually be carried out by a simple least squares regression. This also has a useful statistical advantage: the asymptotic covariance matrix of the one-step estimator will be the same as that of the GMM estimator, with its attendant efficiency properties, rather than the possibly much more complex covariance matrix of the initial estimator. For example, the initial estimator might be the result of multiple-stage estimation, as described in the previous section, with a covariance matrix of the form given in that section. However, one linear step starting from that estimator gives a result that is asymptotically equivalent to solving the full joint GMM problem. Alternately, one might start from initial GMM estimators, and in one step obtain a result that is asymptotically equivalent to full maximum likelihood estimation. Within the context of hypothesis testing with GMM estimates, it is possible to go in one linear step from any suitable initially consistent estimator to estimators that are asymptotically equivalent to either the unconstrained or constrained GMM estimators.

The first result based on these ideas is estimation of an expectation that depends on estimated parameters. Suppose one wishes to estimate $\mathbf{E}_z m(z,\theta_o)$, where m is a vector of functions of random variables z and a parameter vector θ that has true value θ_o. If τ_n is any consistent estimator of θ_o, the sample average of $m(z_t,\theta)$ converges in probability to $\mathbf{E}_z m(z,\theta)$ *uniformly* in θ, and $\mathbf{E}_z m(z,\theta)$ is continuous in θ, then

$$\frac{1}{n}\sum_{t=1}^{n} m(z_t,\tau_n) \rightarrow_p \mathbf{E}_z m(z,\theta_o).$$

This works because

$$\text{Prob}\left(\left| \frac{1}{n}\sum_{t=1}^{n} m(z_t,\tau_n) - \mathbf{E}_z m(z,\tau_n)\right| > \varepsilon\right)$$

$$\leq \frac{1}{n}\sum_{t=1}^{n} \text{Prob}(\sup_{\theta} |\, m(z_t,\theta) - \mathbf{E}_z m(z,\theta)| > \varepsilon) \to 0$$

and $\mathbf{E}_z m(z,\tau_n) \to \mathbf{E}_z m(z,\theta_o)$. Suppose one strengthens the requirement on τ_n to the condition that it be $n^{1/2}$-consistent, meaning that $n^{1/2}(\tau_n - \theta_o)$ is stochastically bounded, or for each $\varepsilon > 0$ there exists $M > 0$ such that

$$\text{Prob}(|n^{1/2}(\tau_n - \theta_o)| > M) < \varepsilon \text{ for all } n.$$

Suppose that $m(z,\theta)$ satisfies a Lipschitz condition at θ_o; i.e., there exists a function $L(z)$ with a finite expectation such that $|m(z,\theta) - m(z,\theta_o)| \leq L(z)\cdot|\theta - \theta_o|$. Then the result holds without requiring uniform convergence in probability for sample averages of $m(z,\theta)$.

The preceding result is useful for calculation of Wald or Lagrange Multiplier test statistics, which require estimation of $G(\theta_o)$, $\Omega(\theta_o)$, and/or $A(\theta_o)$. The arrays $G_n(\theta)$, $\Omega_n(\theta)$, and $A_n(\theta)$ are uniformly convergent, and the result establishes for any initial consistent estimator τ_n that $G_n(\tau_n) \to_p G(\theta_o)$, $\Omega_n(\tau_n) \to_p \Omega(\theta_o)$, and $A_n(\tau_n) \to_p A(\theta_o)$. Then, using these estimates preserves the asymptotic equivalence of the tests under the null and local alternatives. In particular, one can evaluate terms entering the definitions of these arrays at T_n, T_{an}, or any other consistent estimator of θ_o. In sample analogs that converge to these arrays by the law of large numbers, one can freely substitute sample and population terms that leave the probability limits unchanged. For example, if $z_t = (y_t,x_t)$ and τ_n is any consistent estimator of θ_o, then Ω can be estimated by (1) an analytic expression for $\mathbf{E}g(z,\theta)g(z,\theta)'$, evaluated at τ_n, (2) a sample average $\frac{1}{n}\sum_{t=1}^{n} g(z_t,\tau_n)g(z_t,\tau_n)'$, or (3) a sample average of conditional expectations $\frac{1}{n}\sum_{t=1}^{n} E_{y|x_t}\, g(y,x_t,\theta)g(y,x_t,\theta)'$ evaluated at $\theta = \tau_n$. It should be noted however that these first-order equivalences do *not* hold in finite samples, or even to higher orders of $n^{1/2}$. Thus, there may be clear choices between these when higher orders of approximation are taken into account.

The second result, called the *one-step theorem*, considers the first-order condition associated with a GMM criterion function, $0 = G_n'\Omega_n^{-1} g_n(\theta)$. Suppose one has an initial $n^{1/2}$-consistent estimator τ_n for θ_o. A Taylor's expansion of the first-order condition about τ_n yields

$$G_n'\Omega_n^{-1} g_n(\theta) = G_n'\Omega_n^{-1}g_n(\tau_n) + G_n'\Omega_n^{-1}G_n(\theta - \tau_n) + O((\theta - \tau_n)^2).$$

Then, a one-step approximation to the unconstrained GMM estimator is

$$T_{on} = \tau_n - (G_n'\Omega_n^{-1}G_n)^{-1}G_n'\Omega_n^{-1}g_n(\tau_n).$$

A Taylor's expansion around θ_o of the GMM first-order condition, evaluated at τ_n, yields

$$n^{1/2}G_n'\Omega_n^{-1}g_n(\tau_n) = n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) + G_n'\Omega_n^{-1}G_n\cdot n^{1/2}(\tau_n - \theta_o) + o_p.$$

Combine this with the condition $-G_n'\Omega_n^{-1}g_n(\tau_n) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - \tau_n)$ to conclude that

$$-n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - \theta_o) + o_p,$$

and the condition

$$-n^{1/2}G_n'\Omega_n^{-1}g_n(\theta_o) = G_n'\Omega_n^{-1}G_n n^{1/2}(T_n - \theta_o) + o_p$$

to conclude that

$$0 = G_n'\Omega_n^{-1}G_n n^{1/2}(T_{on} - T_n) + o_p,$$

so that T_{on} and T_n are asymptotically equivalent.

The one-step theorem can also be applied to the constrained GMM estimator. Suppose the null hypothesis, or a local alternative, $a(\theta_o) = \delta\cdot n^{-1/2}$, is true. Define one-step constrained estimators from the Lagrangian first-order conditions:

$$\begin{bmatrix} T_{oan} \\ \gamma_{oan} \end{bmatrix} = \begin{bmatrix} \tau_n \\ 0 \end{bmatrix} - \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_\theta Q_n(\tau_n) \\ -a(\tau_n) \end{bmatrix} .$$

Note in this definition that $\gamma = 0$ is a trivial initially consistent estimator of the Lagrangian multipliers under the null or local alternatives, and that the arrays B and A can be estimated at τ_n. The one-step theorem again applies, yielding $n^{-1/2}(T_{oan}-T_{an}) \to_p 0$ and $n^{-1/2}(\gamma_{oan}-\gamma_{an}) \to_p 0$. Then, these one-step equivalents can be substituted in any of the test statistics of the trinity without changing their asymptotic distribution.

A regression procedure for calculating the one-step expressions is often useful for computation. The adjustment from τ_n yielding the one-step unconstrained estimator is obtained by a two-stage least squares regression of the constant one on $\nabla_\theta l(z_t,\tau_n)$, with $g(z_t,\tau_n)$ as instruments; i.e.,

a. Regress each component of $\nabla_\theta l(z_t,\tau_n)$ on $g(z_t,\tau_n)$ in the sample t = 1,...,n, and retrieve fitted values $\nabla_\theta l^*(z_t,\tau_n)$;
b. Regress 1 on $\nabla_\theta l^*(z_t,\tau_n)$; and adjust τ_n by the amounts of the fitted coefficients.

Step (a) yields $\nabla_\theta l^*(z_t,\tau_n)' = g(z_t,\tau_n)\Omega_n^{-1}\Gamma_n$, and step (b) yields coefficients

$$\Delta = \left[\sum_{t=1}^{n} [\nabla_\theta l^*(z_t,\tau_n)][\nabla_\theta l^*(z_t,\tau_n)]'\right]^{-1} \sum_{t=1}^{n} \nabla_\theta l^*(z_t,\tau_n)$$
$$= (\Gamma_n'\Omega_n\Gamma_n)^{-1}\Gamma_n'\Omega_n g_n(\tau_n).$$

This is the adjustment indicated by the one-step theorem.

Computation of one-step constrained estimators is conveniently done using the formulas

$$T_{oan} = T_{on} - B^{-1}A'(AB^{-1}A')^{-1}a(T_{on}) \equiv \tau_n + \Delta - B^{-1}A'(AB^{-1}A')^{-1}[a(\tau_n) + A\Delta]$$
$$\gamma_{oan} = -(AB^{-1}A')^{-1}a(T_{on}) \equiv -(AB^{-1}A')^{-1}[a(\tau_n) + A\Delta]$$

with A and B evaluated at τ_n. To derive these formulas from the first-order conditions for the Lagrangian problem, replace $\nabla_\theta Q_n(\tau_n)$ by the expression $-(\Gamma_n'\Omega_n^{-1}\Gamma_n')(T_{on} - \tau_n)$ from the one-step definition of the unconstrained estimator, replace $a(\tau_n)$ by $a(T_{on}) + A(T_{on} - \tau_n)$, and use the formula for a partitioned inverse.

6. SPECIAL CASES

Maximum Likelihood. We have noted that maximum likelihood estimation can be treated as GMM estimation with moments equal to the score, $g = \nabla_\theta l$. The statistics in Table 2 remain the same, with the simplification that $B = \Omega (= G = \Gamma)$. The likelihood ratio statistic $2n[L_n(T_n) - L_n(T_{an})]$, where $L_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} l(z_t,\theta)$, is shown by a Taylor's expansion about T_n to be asymptotically equivalent to the Wald statistic W_{3n}, and hence to all the statistics in Table 2. Note that LR and DM occupy comparable places in the trinity for maximum likelihood and GMM estimation respectively.

Suppose one sets up an estimation problem in terms of a maximum likelihood criterion, but that one does not in fact have the true likelihood function. Suppose that in spite of this misspecification, optimization of the selected criterion yields consistent estimates. One place this commonly arises is when panel data observations are serially correlated, but one writes down the *marginal* likelihoods of the observations ignoring serial correlation. These are sometimes called *pseudo- likelihood* criteria. The resulting estimators can be interpreted as GMM estimators, so that hypotheses can be tested using the statistics in Table 2. Note however that now $G \neq \Omega$, so that $B = G'\Omega^{-1}G$ must be estimated in full, and one cannot do tests using a likelihood ratio of the pseudo-likelihood function.

Least Squares. Consider the nonlinear regression model $y = h(x,\theta) + \varepsilon$, and suppose $E(y|x) = h(x,\theta)$ and $E((y-h(x,\theta))^2|x) = \sigma^2$. The least squares criterion $Q_n(\theta) = \frac{1}{2n}\sum_{t=1}^{n} (y_t - h(z_t,\theta))^2$ is asymptotically equivalent to GMM estimation with $g(z,\theta) = (y-h(x,\theta))\nabla_\theta h(x,\theta)$ and a distance metric $\Omega_n = \frac{\sigma^2}{2n}\sum_{t=1}^{n} [\nabla_\theta h(x,\theta_o)][\nabla_\theta h(x,\theta_o)]'$. For this problem, $B = \Omega = G$. If $h(z_t,\theta) = z_t'\theta$ is linear, one has $g(z_t,\theta) = u_t(\theta)z_t$, where $u_t(\theta) = y_t - z_t'\theta$ is the regression residual, and $\Omega_n = \frac{1}{n}\sum_{t=1}^{n} z_t z_t'$.

Instrumental Variables. Consider the regression model $y_t = h(z_t,\theta_o) + \varepsilon_t$ where ε_t may be correlated with $\nabla_\theta h(z_t,\theta_o)$. Suppose there are *instruments* w such that $\mathbf{E}(\varepsilon_t|w_t) = 0$. For this problem, one has the moment conditions $g(y_t,z_t,w_t,\theta) = (y_t - h(z_t,\theta))f(w_t)$ satisfying $\mathbf{E}g(y_t,z_t,w_t,\theta_o) = 0$ for any vector of functions f(w) of the instruments, so the GMM criterion becomes

$$Q_n(\theta) = \left[\frac{1}{n}\sum_{t=1}^{n}(y_t - h(z_t,\theta))f(w_t)\right]\left[\frac{1}{n}\sum_{t=1}^{n}f(w_t)f(w_t)'\right]^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}(y_t - h(z_t,\theta))f(w_t)\right] .$$

Suppose that it were feasible to construct the conditional expectation of the gradient of the regression function conditioned on w, $q_t = \mathbf{E}(\nabla_\theta h(z_t,\theta_o)|w_t)$. This is the optimal vector of functions of the instruments, in the sense that the GMM estimator based on f(w) = q will yield estimators with an asymptotic covariance matrix that is smaller in the positive definite sense than any other distinct vector of functions of w. A feasible GMM estimator with good efficiency properties may then be obtained by first obtaining a preliminary consistent estimator τ_n employing a simple practical distance metric, second regressing $\nabla_\theta h(z_t,\tau_n)$ on a flexible family of functions of w_t, such as low-order polynomials in w, and third using fitted values from this regression as the vector of functions $f(w_t)$ in a final GMM estimation. Simplifications of this problem result when $h(z,\theta) = z'\theta$ is linear in θ; in this case, the feasible procedure above is simply 2SLS, and no iteration is needed.

Simple hypotheses. An important practical case of the general nonlinear hypothesis $a(\theta_o) = 0$ is that a subset of the parameters are zero. (A hypothesis that parameters equal constants other than zero can be reduced to this case by reparameterization.) Assume $\theta' = (\alpha',\beta')$ where β is of dimension r and α is of dimension k-r, and H_o: $\beta = 0$. The first-order conditions for solution of this problem are $0 = \nabla_\alpha Q_n(T_{an})$, $0 = \nabla_\beta Q_n(T_{an}) + \gamma_{an}$, implying $\gamma_{an} = -\nabla_\beta Q_n(T_{an})$, and $A = [0\ I_r]$ is a r×k matrix whose first k-r columns are zero. Let $C \equiv B^{-1}$ be the asymptotic covariance matrix of $n^{1/2}(T_n - \theta_o)$, and $AB^{-1}A' = C_{\beta\beta}$ the submatrix of C for β. Taylor's expansions about T_n of the first-order conditions imply $n^{1/2}(T_{1,n}-T_{1,an}) = -B_{\alpha\alpha}B_{\alpha\beta}\,n^{1/2}T_{2,n} + o_p$ and $n^{1/2}\gamma_{an} = [B_{\beta\beta}-B_{\beta\alpha}B_{\alpha\alpha}^{-1}B_{\alpha\beta}]n^{1/2}T_{2,n} + o_p = \beta|_n'C_{\beta\beta}^{-1}T_{2,n} + o_p$. Then the Wald statistics are

$$W_{1n} = nT_{2,n}'C_{\beta\beta}^{-1}T_{2,n},\ W_{2n} = n\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}'\begin{bmatrix}B_{\alpha\beta}\\ B_{\beta\beta}\end{bmatrix}C_{\beta\beta}^{-1}\begin{bmatrix}B_{\beta\alpha} & B_{\beta\beta}\end{bmatrix}\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix},$$

$$W_{3n} = n\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}'B\begin{bmatrix}T_{1,n}-T_{1,an}\\ T_{2,n}\end{bmatrix}.$$

You can check the asymptotic equivalence of these statistics by substituting the expression for $n^{1/2}(T_{1,n}-T_{1,an})$. The LM statistic, in any version, becomes $LM_n = n\nabla_\beta Q_n(T_{an})'C_{\beta\beta}\nabla_\beta Q_n(T_{an})$. Recall that B, hence C, can be evaluated at any consistent estimator of θ_o. In particular, the constrained estimator is consistent under the null or under local alternatives. The LM testing procedure for this case is then to (a) compute the constrained estimator $T_{1,an}$ subject to the condition $\beta = 0$, (b) calculate the gradient and hessian of Q_n with respect to the full parameter vector, evaluated at $T_{1,an}$ and $\beta = 0$, and (c) form the quadratic form above for LM_n from the β part of the gradient and the β submatrix of the inverse of the hessian. Note that this does not require any iteration of the GMM criterion with respect to the full parameter vector.

It is also possible to carry out the calculation of the LM_n test statistic using auxiliary regressions. This could be done using the auxiliary regression technique introduced earlier for the calculation of LM_{3n} in the case of any nonlinear hypothesis, but a variant is available for this case that reduces the size of the regressions required. The steps are as follows:

a. Regress $\nabla_\alpha l(z_t,T_{an})'$ and $\nabla_\beta l(z_t,T_{an})'$ on $g(z_t,T_{an})$, and retrieve the fitted values $\nabla_\alpha l^*(z_t,T_{an})'$ and $\nabla_\beta l^*(z_t,T_{an})'$.
b. Regress $\nabla_\beta l^*(z_t,T_{an})$ on $\nabla_\alpha l^*(z_t,T_{an})$, and retrieve the *residual* $u(z_t,T_{an})$.
c. Regress the constant 1 on the residual $u(z_t,T_{an})$, and calculate the sum of squares of the *fitted* values of 1. This quantity is LM_n.

In the case of maximum likelihood estimation, Step (a) is redundant and can be omitted.

7. TESTS FOR OVERIDENTIFYING RESTRICTIONS

Consider the GMM estimator based on moments $g(z_t,\theta)$, where g is m×1, θ is k×1, and m > k, so there are *over-identifying moments*. The criterion

$$Q_n(\theta) = (1/2)g_n(\theta)'\Omega_n^{-1}g_n(\theta),$$

evaluated at its minimizing argument T_n for any $\Omega_n \to_p \Omega$, has the property that $2nQ_n \equiv 2nQ_n(T_n) \to_d \chi^2(m-k)$ under the null hypothesis that $\mathbf{E}g(z,\theta_o) = 0$. This statistic then provides a specification test for the over-identifying moments in g. It can also be used as an indicator for convergence in numerical search for T_n.

To demonstrate this result, recall that $-\Omega^{-1/2} n^{1/2}g_n(\theta_o) = U_n \to_d U \sim N(0,I)$ and $n^{1/2}(T_n - \theta_o) = B^{-1}G'\Omega^{-1/2}U_n + o_p$. Then, a Taylor's expansion yields

$$\Omega^{-1/2} n^{1/2}g_n(T_n) = -U_n + \Omega^{-1/2}GB^{-1}G'\Omega^{-1/2}U_n + o_p = -R_nU_n + o_p,$$

where $R_n = I - \Omega^{-1/2}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1/2}$ is idempotent of rank m - k. Then

$$2nQ_n(T_n) = U_n'R_nU_n + o_p \to_d \chi^2(m-k) .$$

Suppose that instead of estimating θ using the full list of moments, one uses a linear combination $Lg(z,\theta)$, where L is r×m with $k \le r < m$. In particular, L may select a subset of the moments. Let T_{an} denote the GMM estimator obtained from these moment combinations, and assume the identification conditions are satisfied so T_{an} is $n^{1/2}$-consistent. Then the statistic $S = ng_n(T_{an})'\Omega_n^{-1/2}R_n\Omega_n^{-1/2}g_n(T_{an}) \to_d \chi^2(m-k)$ under H_o, and this statistic is asymptotically equivalent to the statistic $2nQ_n(T_n)$. This result holds for any $n^{1/2}$-consistent estimator τ_n of θ_o, not necessarily the optimal GMM estimator for the moments $Lg(z,\theta)$, or even an initially consistent estimator based on only these moments. The distance metric in the center of the quadratic form S does not depend on L, so that the formula for the statistic is invariant with respect to the choice of the initially consistent estimator. This implies in particular that the test statistics S for over-identifying restrictions, starting from different subsets of the moment conditions, are all asymptotically equivalent. However, the presence of the idempotent matrix R_n in the center of the quadratic form S is critical to its statistical properties. Only the GMM distance metric criterion using all moments, evaluated at T_n, is asymptotically equivalent to S. Substitution of another consistent estimator τ_n in place of T_n yields an asymptotically equivalent version of S, but $2nQ_n(\tau_n)$ is not asymptotically chi-square distributed.

The test for overidentifying restrictions can be recast as a LM test by artificially embedding the original model in a richer model. Partition the moments

$$g(z,\theta) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta) \end{bmatrix},$$

where g^1 is kx1 with $G_1 = \mathbf{E}\nabla_\theta g^1(z,\theta_o)$ of rank k, and g^2 is (m-k)x1 with $G_2 = E\nabla_\theta g^2(z,\theta_o)$. Embed this in the model

$$g^*(z,\theta,\psi) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta)+\psi \end{bmatrix}$$

where ψ is a (m-k) vector of additional parameters. The first-order-condition for GMM estimation of this expanded model is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G_{1n} & G_{2n} \\ 0 & I_{m-k} \end{bmatrix} \begin{bmatrix} \Omega_n & 0 \\ 0 & I_{m-k} \end{bmatrix} \begin{bmatrix} g_n(T_{an}) \\ g_n(T_{an}) - \psi_n \end{bmatrix}$$

The second block of conditions are satisfied by $\psi_n = g_n(T_{an})$, no matter what T_{an}, so T_{an} is determined by $O = G_n\Omega_n g_n(T_{an})$. This is simply the estimator obtained from the first block of moments, and coincides with the earlier definition of T_{an}. Thus, *unconstrained* estimation of the *expanded* model coincides with *restricted* estimation of the original model. Next consider GMM estimation of the expanded model subject to H_o:$\psi = O$. This constrained estimation obviously coincides with GMM estimation using all moments in the original model, and yields T_n. Thus, *constrained* estimation of the *expanded* model coincides with *unrestricted* estimation of the original model.

The Distance Metric test statistic for the constraint $\psi = 0$ in the expanded model is $DM_n = 2n[Q_n(T_n,0) - Q_n(T_n,\psi_n)] \equiv 2nQ_n(T_n)$, where Q_n denotes the criterion as a function of the expanded parameter list. One has $Q_n(T_n,0) \equiv Q_n(T_n)$ from the coincidence of the constrained expanded model estimator and the unrestricted original model estimator, and one has $Q_n(T_{an},\psi_n) = 0$ since the number of moments equals the number of parameters. Then, the test statistic $2nQ_n(T_n)$ for overidentifying restrictions is identical to a distance metric test in the expanded model, and hence asymptotically equivalent to any of the trinity of tests for H_o: $\psi = O$ in the expanded model.

We give four examples of econometric problems that can be formulated as tests for over-identifying restrictions:

Example 1. If $y = x\beta+\varepsilon$ with $E(\varepsilon|x) = 0$, $E(\varepsilon^2|x) = \sigma^2$, then the moments

$$g^1(z,\beta) = \begin{bmatrix} x(y-x\beta) \\ (y-x\beta)^2 - \sigma^2 \end{bmatrix}$$

can be used to estimate β and σ^2. If ε is normal, then these GMM estimators are MLE. Normality can be tested via the additional moments that give skewness and kurtosis,

$$g^2(x,\beta) = \begin{bmatrix} (y - x\beta)^3/\sigma^3 \\ (y - x\beta)^4/\sigma^4 - 3 \end{bmatrix} .$$

Example 2. In the linear model $y = xb+\varepsilon$ with $E(\varepsilon|x) = 0$ and $E(\varepsilon_t\varepsilon_s|x) = 0$ for $t \neq s$, but with possible heteroskedasticity of unknown form, one gets the OLS estimates b of β and $V(b) = s^2(X'X)^{-1}$ under the null hypothesis of homoskedasticity. A test for homoskedasticity can be based on the population moments $0 = \mathbf{E}\ \text{vecu}[x'x(\varepsilon^2 - \sigma^2)]$, where "vecu" means the vector formed from the upper triangle of the array. The sample value of this moment vector is

$$\text{vecu}\left[\frac{1}{n}\sum_{t=1}^{n} x_t' x_t\left((y_t - x_t\beta)^2 - s^2\right)\right] ,$$

the difference between the White robust estimator and the standard OLS estimator of $\text{vecu}[X'\Omega X]$.

Example 3. If $l(z,\theta)$ is the log likelihood of an observation, and T_n is the MLE, then an additional moment condition that should hold if the model is specified correctly is the information matrix equality

$$0 = E\,\nabla_{\theta\theta} l(z,\theta_o) + E\nabla_\theta l(z,\theta_o)\nabla_\theta l(z,\theta_o)' .$$

The sample analog is White's information matrix test, which then can be interpreted as a GMM test for over-identifying restrictions.

Example 4. In the nonlinear model $y = h(x,\theta) + \varepsilon$ with $E(\varepsilon|x) = 0$, and T_n a GMM estimator based on moments $w(x)(y-h(x,\theta))$, where w(x) is some vector of functions of x, suppose one is interested in testing the stronger assumption that ε is *independent* of x. A necessary and sufficient condition for independence is $\mathbf{E}[w(x) - \mathbf{E}w(x)]f(y - h(x,\theta_o)) = 0$ for every function f and vector of functions w for which the moments exist. A specification test can be based on a selection of such moments.

8. SPECIFICATION TESTS IN LINEAR MODELS[11]

GMM tests for over-identifying restrictions have particularly convenient forms in linear models. Three standard specification tests will be shown to have this interpretation. We will use projections and a few of their properties in the following discussion; a more detailed discussion of projections is given in the Appendix to this chapter. Let $P_X = X(X'X)^{-}X$ denote the *projection matrix* from $\mathbb{R}^n$ onto the linear subspace $\mathbf{X}$ *spanned* by a n×p array X; note that it is idempotent. (We use a Moore-Penrose generalized inverse in the definition of P_X to handle the possibility that X is less than full rank; see the Appendix.) Let $Q_X = I - P_X$ denote the projection matrix onto the linear subspace

[11] Paul Ruud contributed substantially to this section.

orthogonal to **X**. If **X** is a subspace generated by an array X and **W** is a subspace generated by an array W = [X Z] that contains **X**, then $P_X P_W = P_W P_X = P_X$ and $Q_X P_W = P_W - P_X$.

Omitted Variables Test: Consider the regression model $y = X\beta + \varepsilon$, where y is n×1, X is n×k, $\mathbf{E}(\varepsilon|X) = 0$, and $\mathbf{E}(\varepsilon\varepsilon'|X) = \sigma^2 I$. Suppose one has the hypothesis H_o: $\beta_1 = 0$, where β_1 is a p×1 subvector of β, and let X* denote the n×(k-p) array of variables whose coefficients are not constrained under the null hypothesis. Define $u = y - Xb$ to be the residual associated with an estimator b of β. The GMM criterion is then $2nQ = u'X(X'X)^{-1}X'u/\sigma^2$. The *projection matrix* $P_X \equiv X(X'X)^{-1}X'$ that appears in the center of this criterion can obviously be decomposed as $P_X \equiv P_{X^*} + (P_X - P_{X^*})$. Under H_o, $u = y - X_2 b_2$ and X′u can be interpreted as k = p + q over-identifying moments for the q parameters β_2. Then, the GMM test statistic for over-identifying restrictions is the minimum value $2nQ_n^*$ in b_2 of $u'P_X u/\sigma^2$. But $P_X u = P_{X^*} u + (P_X - P_{X^*})y$ and $\min_{b_2} u' P_{X^*} u = 0$ (at the OLS estimator under H_o that makes u orthogonal to X_2). Then $2nQ_n = y'(P_X - P_{X^*})y/\sigma^2$. The unknown variance σ^2 in this formula can be replaced by any consistent estimator s^2, in particular, the estimated variance of the disturbance from either the restricted or the unrestricted regression, without altering the asymptotic distribution, which is $\chi^2(q)$ under the null hypothesis.

The statistic $2nQ_n$ has three alternative interpretations. First,

$$2nQ_n = y'P_X y/\sigma^2 - y'P_{X^*} y/\sigma^2 = \frac{SSR_{X_2} - SSR_X}{\sigma^2},$$

which is the difference of the sum of squared residuals from the restricted regression under H_o and from the unrestricted regression, normalized by σ^2. This is a large-sample version of the usual finite-sample F-test for H_o. Second, note that the fitted value of the dependent variable from the restricted regression is $\hat{y}_o = P_{X^*} y$, and from the unrestricted regression is $\hat{y}_u = P_X y$, so that

$$2nQ_n = (\hat{y}_o'\hat{y}_o - \hat{y}_u'\hat{y}_u)/\sigma^2 = (\hat{y}_o - \hat{y}_u)'(\hat{y}_o - \hat{y}_u)/\sigma^2 = \|\hat{y}_o - \hat{y}_u\|^2/\sigma^2.$$

Then, the statistic is calculated from the distance between the fitted values of the dependent variable with and without H_o imposed. Note that it can be computed from fitted values without any covariance matrix calculation. Third, let b_o denote the GMM estimator restricted by H_o and b_u denote the unrestricted GMM estimator. Then, b_o consists of the OLS estimator for β_2 and the hypothesized value 0 for β_1, while b_u is the OLS estimator for the full parameter vector. Note that $\hat{y}_o = Xb_o$ and $\hat{y}_u = Xb_u$, so that $\hat{y}_o - \hat{y}_u = X(b_o - b_u)$. Then

$$2nQ_n = (b_o - b_u)'(X'X/\sigma^2)(b_o - b_u) = (b_o - b_u)'V(b_u)^{-1}(b_o - b_u).$$

This is the Wald statistic W_{3n}. From the equivalent form W_{2n} of the Wald statistic, this can also be written as a quadratic form $2nQ_n = b_{1,u}'V(b_{1,u})^{-1}b_{1,u}$, where $b_{1,u}$ is the subvector of unrestricted estimates for the parameters that are zero under the null hypothesis.

Two other important cases of specification tests in linear models are discussed in the following chapters. *Endogeneity tests* are discussed in the chapter on instrumental variables, and *tests for over-identifying restrictions* are discussed in the chapter on simultaneous equations.

APPENDIX

Projections: Consider a Euclidean space $\mathbb{R}^n$ of dimension n, and suppose X is a n×p array with columns that are vectors in this space. Let **X** denote the linear subspace of $\mathbb{R}^n$ that is *spanned* or *generated* by X.; and i.e., the space formed by all linear combinations of the vectors in X. Every linear subspace can be identified with an array such as X. The dimension of the subspace is the rank of X. (The array X need not be of full rank, although if it is not, then a subarray of linearly independent columns also generates **X**.) A given X determines a unique subspace, so that X characterizes the subspace. However, any set of vectors contained in the subspace that form an array with the rank of the subspace, in particular any array XA with rank equal to the dimension of X, also generates **X**. Then, X is not a unique characterization of the subspace it generates.

The *projection* of a vector y in $\mathbb{R}^n$ into the subspace **X** is defined as the point v in **X** that is the minimum Euclidean distance from y. Since each vector v in **X** can be represented as a linear combination Xα of an array X that generates **X**, the projection is characterized by the value of α that minimizes $(y-X\alpha)'(y-X\alpha)$. The solution to this problem is the OLS estimator $\hat{\alpha} = (X'X)^{-}X'y$ and $v = X\hat{\alpha} = X(X'X)^{-}X'y$. In these formulas, we use $(X'X)^{-}$ rather than $(X'X)^{-1}$; the former denotes the Moore-Penrose *generalized* inverse, and is defined even if X is not of full rank (see below). The array $P_X = X(X'X)^{-}X'$ is termed the *projection matrix* for the subspace **X**; it is the linear transformation in $\mathbb{R}^n$ that maps any vector in the space into its projection v in **X**. The matrix P_X is *idempotent* (i.e., $P_X P_X = P_X$ and $P_X = P_X'$), and every idempotent matrix can be interpreted as a projection matrix. These observations have two important implications: First, the projection matrix is uniquely determined by X, so that starting from a different array that generates **X**, say an array S = XA, implies $P_X = P_S$. (One could use the notation $P_{\mathbf{X}}$ rather than P_X to emphasize that the projection matrix depends only on the subspace, and not on any particular set of vectors that generate **X**.) Second, if a vector y is contained in **X**, then the projection into **X** leaves it unchanged, $P_X y = y$.

Define $Q_X = I - P_X = I - X(X'X)^{-1}X'$; it is the projection to the subspace orthogonal to that spanned by X. Every vector y in $\mathbb{R}^n$ is uniquely decomposed into the sum of its projection $P_X y$ onto **X** and its projection $Q_X y$ onto the subspace orthogonal to **X**. Note that $P_X Q_X = 0$, a property that holds in general for two projections onto orthogonal subspaces.

If **X** is a subspace generated by an array X and **W** is a subspace generated by an array W = [X Z] that contains X, then $\mathbf{X} \subseteq \mathbf{W}$. This implies that $P_X P_W = P_W P_X = P_X$; i.e., a projection onto a subspace is left invariant by a further projection onto a larger subspace, and a two-stage projection onto a large subspace followed by a projection onto a smaller one is the same as projecting directly onto the smaller one. The subspace of **W** that is orthogonal to **X** is generated by $Q_X W$; i.e., it is the set of linear combinations of the residuals, orthogonal to X, obtained by regressing W on X. Note that any y in $\mathbb{R}^n$ has a unique decomposition $P_X y + Q_X P_W y + Q_W y$ into the sum of projections onto three mutually orthogonal subspaces, **X**, the subspace of **W** orthogonal to **X**, and the subspace orthogonal to **W**. The projection $Q_X P_W$ can be rewritten $Q_X P_W = P_W - P_X = P_W Q_X = Q_X P_W Q_X$, or

since $Q_X W = Q_X[X\ Z] = [0\ Q_X Z]$, $Q_X P_W = P_{Q_X W} = P_{Q_X Z} = Q_X Z(Z'Q_X Z)^{-}Z'Q_X$. This establishes that P_W and Q_X commute. This condition is necessary and sufficient for the product of two projections to be a projection; equivalently, it implies that $Q_X P_W$ is idempotent since $(Q_X P_W)(Q_X P_W) = Q_X(P_W Q_X)P_W = Q_X(Q_X P_W)P_W = Q_X P_W$.

Generalized Inverses: Some test statistics are conveniently defined using generalized inverses. This section gives a constructive definition of a generalized inverse, and lists some of its properties. A k×m matrix A^- is a *Moore-Penrose generalized inverse* of a m×k matrix A if it has three properties:

(i) $AA^-A = A$,
(ii) $A^-AA^- = A^-$
(iii) AA^- and A^-A are symmetric

There are other generalized inverse definitions that have some, but not all, of these properties; in particular A^+ will denote any matrix that satisfies (i), or $AA^+A = A$.

First, a method for constructing the generalized inverse is described, and then some of the implications of the definition are developed. The construction is called the *singular value decomposition* (SVD) of a matrix, and is of independent interest as a tool for finding the eigenvalues and eigenvectors of a symmetric matrix, and for calculation of inverses of moment matrices of data with high multicollinearity; see Press *et al* (1986) for computational algorithms and programs.

Lemma 1. Every real m×k matrix A *of rank r can be decomposed into a product* $A = UDV'$ *where* D *is a r×r diagonal matrix with positive non-increasing elements down the diagonal, and* U *and* V *are column-orthonormal matrices of respective dimension m×r and k×r; i.e.,* $U'U = I_r = V'V$.

Proof: The m×m matrix AA' is symmetric and positive semidefinite. Then, there exists a m×m orthonormal matrix W, partitioned $W = [W_1\ W_2]$ with W_1 of dimension m×r, such that $W_1'(AA')W_1 = G$ is diagonal with positive, non-increasing diagonal elements, and $W_2'(AA')W_2 = 0$, implying $A'W_2 = 0$. Define D from G by replacing the diagonal elements of G by their positive square roots. Note that $W'W = I = WW' \equiv W_1 W_1' + W_2 W_2'$. Define $U = W_1$ and $V' = D^{-1}U'A$. Then, $U'U = I_r$ and $V'V = D^{-1}U'AA'UD^{-1} = D^{-1}GD^{-1} = I_r$. Further, $A = (I_m - W_2 W_2')A = UU'A = UDV'$. This establishes the decomposition. □

Note that if A is symmetric, then U is the array of eigenvectors of A corresponding to the non-zero roots, so that $A'U = UD_1$, with D_1 the r×r diagonal matrix with the non-zero eigenvalues in descending magnitude down the diagonal. In this case, $V = A'UD^{-1} = UD_1D^{-1}$. Since the elements of D_1 and D are identical except possibly for sign, the columns of U and V are either equal (for positive roots) or reversed in sign (for negative roots).

Lemma 2. The *Moore-Penrose generalized inverse of a* m×k *matrix* A *(which has a SVD A = UDV') is the matrix* $A^- = VD^{-1}U$, *where* V is k×r, D is r×r, *and* U *is* r×m. *Let* A^+ *denote any matrix, including* A^-, *that satisfies* $AA^+A = A$. *These matrices satisfy:*

(1) $A^+ = A^{-1}$ *if* A *is square and non-singular*.

(2) *The system of equations* $Ax = y$ *has a solution if and only if* $y = AA^+y$, *and the linear subspace of all solutions is the set of vectors* $x = A^+y + [I - A^+A]z$ *for all* $z \in \mathbb{R}^k$.
(3) AA^+ *and* A^+A *are idempotent*.
(4) *If* A *is idempotent, then* $A = A^-$.
(5) *If* $A = BCD$ *with* B *and* D *nonsingular, then* $A^- = D^{-1} C^- B^{-1}$, *and any matrix* $A^+ = D^{-1}C^+B^{-1}$ *satisfies* $AA^+A = A$.
(6) $(A')^- = (A^-)'$
(7) $(A'A)^- = A^-(A^-)'$
(8) $(A^-)^- = A = AA'(A^-)' = (A^-)'A'A$.

(9) *If* $A = \sum_i A_i$ *with* $A_i'A_j = 0$ *and* $A_iA_j' = 0$ *for* $i \neq j$, *then* $A^- = \sum_i A_i^-$.

Lemma 3. If A *is square, symmetric, and positive semidefinite of rank r, then*
(1) There exist Q *positive definite and* R *idempotent of rank r such that* $A = QRQ$ *and* $A^- = Q^{-1}RQ^{-1}$.
(2) *There exists a* $k \times r$ *column-orthonormal matrix* U *such that* $U'AU = D$ *is non-singular diagonal and* $A^- = U(U'AU)^{-1}U'$.
(3) A *has a symmetric square root* $B = A^{1/2}$, *and* $A^- = B^-B^-$.

Proof: Let $W = [U\ W_2]$ be an orthogonal matrix diagonalizing A. Then, $U'AU = D$, a diagonal matrix of positive eigenvalues, and $AW_2 = 0$. Define $Q = W \begin{bmatrix} D^{1/2} & 0 \\ 0 & I_{m-r} \end{bmatrix} W'$, $R = WW'$, and $B = UD^{1/2}U'$. □

Lemma 4. Suppose $y \sim N(\lambda, A)$, *with* A *of rank r, and let* $A = S^{1/2}TS^{1/2}$ *be a decomposition of* A *in terms of a positive definite matrix* S *and an idempotent matrix* T *of rank r. Suppose* λ *is contained in the space spanned by* A*; i.e.,* $TS^{-1/2}\lambda = S^{-1/2}\lambda$., *Then* $y'S^{-1}y$ *and* $y'A^-y$ *are identical, and are distributed noncentral chi-square with r degrees of freedom and noncentrality parameter* $\lambda'A^-\lambda$.

Proof: Let $W = [U\ W_2]$ be an orthonormal matrix that diagonalizes A, as in the proof of Lemma 3, with $U'AU = D$, a positive diagonal $r \times r$ matrix, and $W'AW_2 = 0$, implying $AW_2 = 0$. Then, the nonsingular transformation $z = \begin{bmatrix} D^{-1/2} & 0 \\ 0 & I \end{bmatrix} W'y$ has mean $\begin{bmatrix} D^{-1/2}U'A\lambda \\ 0 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, so that $z_1 = D^{-1/2}U'y$ is distributed $N(D^{-1/2}U'A\lambda, I_r)$, $z_2 = W_2y = 0$, implying $W'y = [D^{1/2}z_1\ 0]$. It is standard that $z'z$ has a noncentral chi-square distribution with r degrees of freedom and noncentrality parameter $\lambda'AUD^{-1}U'A\lambda = \lambda'A\lambda$. The condition $A = AA^+A$ implies

$U'AU = U'AWW'A^{+}WW'AU$, or $D = [D|0]W'A^{+}W[D\ 0]' = D(U'A^{+}U)D$. Hence, $U'A^{+}U = D^{-1}$. Then

$$y'A^{+}y = y'WW'A^{+}WW'y = [z_1'D^{1/2}\ 0](W'A^{+}W)[D^{1/2}\ z_1'0]'$$
$$= z_1'D^{1/2}(U'A^{+}U)D^{1/2}z_1 = z_1'z_1. \ \square$$

CHAPTER 4. INSTRUMENTAL VARIABLES

1. INTRODUCTION

Consider the linear model $y = X\beta + \varepsilon$, where y is n×1, X is n×k, β is k×1, and ε is n×1. Suppose that *contamination* of X, where some of the X variables are correlated with ε, is suspected. This can occur, for example, if ε contains omitted variables that are correlated with the included variables, if X contains measurement errors, or if X contains endogenous variables that are determined jointly with y.

OLS Revisited: Premultiply the regression equation by X' to get

$$X'y = X'X\beta + X'\varepsilon. \tag{1}$$

One can interpret the OLS estimate b_{OLS} as the solution obtained from (1) by first approximating $X'\varepsilon$ by zero, and then solving the resulting k equations in k unknowns,

$$X'y = X'Xb_{OLS}, \tag{2}$$

for the unknown coefficients. Subtracting (1) from (2), one obtains the condition

$$X'X(b_{OLS} - \beta) = X'\varepsilon, \tag{3}$$

and the error in estimating β is linear in the error caused by approximating $X'\varepsilon$ by zero. If $X'X/n \to_p A$ positive definite and $X'\varepsilon/n \to_p 0$, (3) implies the result that $b_{OLS} \to_p \beta$. What makes OLS consistent when $X'\varepsilon/n \to_p 0$ is that approximating $X'\varepsilon$ by zero is reasonably accurate in large samples. On the other hand, if one has instead $X'\varepsilon/n \to_p C \neq 0$, then b_{OLS} is not consistent for β, and instead $b_{OLS} \to_p \beta + A^{-1}C$.

Instrumental Variables: Suppose there is a n×j array of variables W, called *instruments*, that have two properties: (i) These variables are uncorrelated with ε; we say in this case that these instruments are *clean*. (ii) The matrix of correlations between the variables in X and the variables in W is of maximum possible rank (= k); we say in this case that these instruments are *fully correlated*. Call the instruments *proper* if they satisfy (i) and (ii). The W array should include any variables from X that are themselves clean. To be fully correlated, W must include at least as many variables as are in X, so that $j \geq k$. Another way of stating this necessary condition is that *the number of instruments in* W *that are excluded from* X *must be at least as large as the number of contaminated variables that are included in* X.

Instead of premultiplying the regression equation by X' as we did for OLS, premultiply it by $R'W'$, where R is a j×k weighting matrix that we get to choose. (For example, R might select a subset of k from the j instrumental variables, or might form k linear combinations of these variables. The only restriction is that R must have rank k.) This gives

(4) $$R'W'y = R'W'X\beta + R'W'\varepsilon.$$

The idea of an instrumental variables (IV) estimator of β is to approximate $R'W'\varepsilon$ by zero, and solve

(5) $$R'W'y = R'W'X\, b_{IV}$$

for $b_{IV} = [R'W'X]^{-1}R'W'y$. Subtract (4) from (5) to get the IV analog of the OLS relationship (3),

(6) $$R'W'X(b_{IV} - \beta) = R'W'\varepsilon.$$

If $R'W'X/n$ converges in probability to a nonsingular matrix and $R'W'\varepsilon/n \rightarrow_p 0$, then $b_{IV} \rightarrow_p \beta$. Thus, in problems where OLS breaks down due to correlation of right-hand-side variables and the disturbances, you can use IV to get consistent estimates, provided you can find proper instruments.

The idea behind (5) is that W and ε are orthogonal in the population, a generalized moment condition. Then, (5) can be interpreted as the solution of a generalized method of moments problem, based on the sample moments $W'(y - X\beta)$. The properties of the IV estimator could be deduced as a special case of the general theory of GMM estimators. However, because the linear IV model is such an important application in economics, we will give IV estimators an elementary self-contained treatment, and only at the end make connections back to the general GMM theory.

2. OPTIMAL IV ESTIMATORS

If there are exactly as many instruments as there are explanatory variables, j = k, then the IV estimator is uniquely determined, $b_{IV} = (W'X)^{-1}W'y$, and R is irrelevant. However, if j > k, each R determines a different IV estimator. What is the best way to choose R? An analogy to the generalized least squares problem provides an answer: Premultiplying the regression equation by W' yields a system of j > k equations in k unknown β's, $W'y = W'X\beta + W'\varepsilon$. Since there are more equations than unknowns, we cannot simply approximate all the $W'\varepsilon$ terms by zero simultaneously, but will have to accommodate at least j-k non-zero residuals. But this is just like a regression problem, with j observations, k explanatory variables, and disturbances $\nu = W'\varepsilon$. Suppose the disturbances ε have a covariance matrix $\sigma^2\Omega$, and hence the disturbances $\nu = W'\varepsilon$ have a non-scalar covariance matrix $\sigma^2W'\Omega W$. If this were a conventional regression satisfying $\mathbf{E}(\nu|W'X) = 0$, then we would know that the generalized least squares (GLS) estimator of β would be BLUE; this estimator is

(7) $$b_{GLSIV} = [X'W(W'\Omega W)^{-1}W'X]^{-1}X'W(W'\Omega W)^{-1}W'y.$$

This corresponds to using the weighting matrix $R = (W'\Omega W)^{-1}W'X$. In truth, the conditional expectation of ν given $W'X$ is not necessarily zero, but clean instruments will have the property that $(W'X)'\varepsilon/n \rightarrow_p 0$ because W and ε are uncorrelated in the population. This is enough to make the analogy work, so that (7) gives the IV estimator that has the smallest asymptotic variance among those that could be formed from the instruments W and a weighting matrix R.

If one makes the usual assumption that the disturbances ε have a scalar covariance matrix, $\Omega = I$, then the best IV estimator reduces to

(8) $$b_{2SLS} = [X'W(W'W)^{-1}W'X]^{-1}X'W(W'W)^{-1}W'y.$$

This corresponds to using the weighting matrix $R = (W'W)^{-1}W'X$. But this formula provides another interpretation of (8). If you regress each variable in X on the instruments, the resulting OLS coefficients are $(W'W)^{-1}W'X$, the same as R. Then, the best linear combination of instruments WR equals the fitted value $X^* = W(W'W)^{-1}W'X$ of the explanatory variables from a OLS regression of X on W. Further, you have $X'W(W'W)^{-1}W'X = X'X^* = X^{*\prime}X^*$ and $X'W(W'W)^{-1}W'y = X^{*\prime}y$, so that the IV estimator (8) can also be written

(9) $$b_{2SLS} = (X^{*\prime}X)^{-1}X^{*\prime}y = (X^{*\prime}X^*)^{-1}X^{*\prime}y.$$

This provides a two-stage least squares (2SLS) interpretation of the IV estimator: First, a OLS regression of the explanatory variables X on the instruments W is used to obtain fitted values X^*, and second a OLS regression of y on X^* is used to obtain the IV estimator b_{2SLS}. Note that in the first stage, any variable in X that is also in W will achieve a perfect fit, so that this variable is carried over without modification in the second stage.

The 2SLS estimator (8) or (9) will no longer be best when the scalar covariance matrix assumption $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$ fails, but under fairly general conditions it will remain consistent. The best IV estimator (7) when $\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$ can be reinterpreted as a conventional 2SLS estimator applied to the transformed regression $Ly = LX\beta + \eta$ using the instruments $(L')^{-1}W$, where L is a Cholesky array that satisfies $L\Omega L' = I$. When Ω depends on unknown parameters, it is often possible to use a feasible generalized 2SLS procedure (FG2SLS): First estimate β using (8) and retrieve the residuals $u = y - Xb_{2SLS}$. Next use these residuals to obtain an estimate Ω^* of Ω. Then find a Cholesky transformation L satisfying $L\Omega^*L' = I$, make the transformations $y = Ly$, $X = LX$, and $W = (L')^{-1}W$, and do a 2SLS regression of y on X using W as instruments. This procedure gives a feasible form of (7), and is also called three-stage least squares (3SLS).

3. STATISTICAL PROPERTIES OF IV ESTIMATORS

IV estimators can behave badly in finite samples. In particular, they may fail to have moments. Their appeal relies on their behavior in large samples, although an important question is when samples are large enough so that the asymptotic approximation is reliable. We first discuss asymptotic properties, and then return to the issue of finite-sample properties.

We already made an argument that IV estimators are consistent, provided some limiting conditions are met. We did not show that IV estimators are unbiased, and in fact they usually are not. An exception where b_{IV} is unbiased is if the original regression equation actually satisfies Gauss-Markov assumptions. Then, no contamination is present, IV is not really needed, and if IV is used, its mean and variance can be calculated in the same way this was done for OLS, by first taking the conditional expectation with respect to ε, given X and W. In this case, OLS is BLUE, and since IV is another linear (in y) estimator, its variance will be at least as large as the OLS variance.

We show next that IV estimators are asymptotically normal under some regularity conditions, and establish their asymptotic covariance matrix. This gives a relatively complete large-sample theory for IV estimators. Let $\sigma^2\Omega$ be the covariance matrix of ε, given W, and assume that it is finite and of full rank. Make the assumptions:

[1] rank(W) = j $\geq$ k
[2a] $W'W/n \rightarrow_p H$, a positive definite matrix
[2b] $W'\Omega W/n \rightarrow_p F$, a positive definite matrix
[3] $X'W/n \rightarrow_p G$, a matrix of rank k
[4] $W'\varepsilon/n \rightarrow_p 0$
[5] $n^{-1/2}W'\varepsilon \rightarrow_d N(0,\sigma^2 F)$

Assumption [1] can always be met by dropping linearly dependent instruments, and should be thought of as true by construction. Assumption [1] implies that $W'W/n$ and $W'\Omega W/n$ are positive definite; Assumption [2] strengthens these to hold in the limit. Proper instruments have $X'W/n$ of rank k from the fully correlated condition and $\mathbf{E}(W'\varepsilon/n) = 0$ by the clean condition. Assumption [3] strengthens the fully correlated condition to hold in the limit. Assumption [4] will usually follow from the condition that the instruments are clean by applying a weak law of large numbers. For example, if the ε are independent and identically distributed with mean zero and finite variance, given W, then Assumption [2a] plus the Kolmogorov WLLN imply Assumption [4]. Assumption [5] will usually follow from Assumption [2b] by applying a central limit theorem. Continuing the i.i.d. example, the Lindeberg-Levy CLT implies Assumption [5]. There are WLLN and CLT that hold under much weaker conditions on the ε's, requiring only that their variances and correlations satisfy some bounds, and these can also be applied to derive Assumptions [4] and [5]. Thus, the statistical properties of IV can be established in the presence of many forms of heteroskedasticity and serial correlation.

Theorem: Assume that [1], [2b], [3] hold, and that an IV estimator is defined with a weighting matrix R_n that may depend on the sample n, but which converges to a matrix R of rank k. If [4] holds, then $b_{IV} \rightarrow_p \beta$. If both [4] and [5] hold, then

(10) $$n^{1/2}(b_{IV} - \beta) \rightarrow_d N(0, \sigma^2(R'G')^{-1}R'FR(GR)^{-1}).$$

Suppose $R_n = (W'W)^{-1}W'X$ and [1]-[5] hold. Then the IV estimator specializes to the 2SLS estimator b_{2SLS} given by (8) which satisfies $b_{2SLS} \rightarrow_p \beta$ and

(11) $$n^{1/2}(b_{2SLS} - \beta) \rightarrow_d N(0, \sigma^2(GH^{-1}G')^{-1}(GH^{-1}FH^{-1}G')(GH^{-1}G')^{-1}).$$

Suppose $R_n = (W'\Omega W)^{-1}W'X$ and [1]-[5] hold. Then the IV estimator specializes to the GLSIV estimator b_{GLSIV} given by (7) which satisfies $b_{GLSIV} \rightarrow_p \beta$ and

(12) $$n^{1/2}(b_{GLSIV} - \beta) \rightarrow_d N(0, \sigma^2(GF^{-1}G')^{-1}).$$

Further, the GLSIV estimator is the minimum asymptotic variance estimator; i.e., $\sigma^2(R'G')^{-1}R'FR(GR)^{-1} - \sigma^2(GF^{-1}G')^{-1}$ is positive semidefinite. If $\Omega = I$, then the 2SLS and GLSIV estimators are the same, and the 2SLS estimator has limiting distribution (12) and is asymptotically best among all IV estimators that use instruments W.

The first part of this theorem is proved by dividing (6) by n and using assumptions [2], [3], and [4], and then dividing (6) by $n^{1/2}$ and applying assumptions [2], [3], and [5]. Substituting the definitions of R for the 2SLS and GLSIV versions then gives the asymptotic properties of these estimators. Finally, a little matrix algebra shows that the GLSIV estimator has minimum asymptotic variance among all IV estimators: Start with the matrix $I - F^{-1/2}G'(GF^{-1}G')^{-1}GF^{-1/2}$ which equals its own square, so that it is idempotent, and therefore positive semidefinite. Premultiply this idempotent matrix by $(R'G')^{-1}R'F^{1/2}$, and postmultiply it by the transpose of this matrix; the result remains positive semidefinite, and equals $(R'G')^{-1}R'FR(GR)^{-1} - (GF^{-1}G')^{-1}$. This establishes the result.

In order to use the large-sample properties of b_{IV} for hypothesis testing, it is necessary to find a consistent estimator for σ^2. The following estimator works: Define IV residuals

$$u = y - Xb_{IV} = [I - X(R'W'X)^{-1}R'W']y = [I - X(R'W'X)^{-1}R'W']\varepsilon,$$

the *Sum of Squared Residuals* SSR = u′u, and $s^2 = u'u/(n-k)$. If $\varepsilon'\varepsilon/n \rightarrow_p \sigma^2$, then s^2 is consistent for σ^2. To show this, simply write out the expression for u′u/n, and take the probability limit:

(13)
$$\begin{aligned}\text{plim } u'u/n &= \text{plim } \varepsilon'\varepsilon/n - 2 \text{ plim } [\varepsilon'W/n]R([X'W/n]R)^{-1}[X'\varepsilon/n] \\ &\quad + [\varepsilon'W/n]R([X'W/n]R)^{-1}[X'X/n](R'[W'X/n])^{-1}R'[W'\varepsilon/n] \\ &= \sigma^2 - 2\cdot 0\cdot R\cdot(GR)^{-1}C + 0\cdot R\cdot(GR)^{-1}A(R'G')^{-1}R'\cdot 0 = \sigma^2.\end{aligned}$$

We could have used n-k instead of n in the denominator of this limit, as it makes no difference in large enough samples. The consistency of the estimator s^2 defined above holds for any IV estimator, and so holds in particular for the 2SLS or GLSIV estimators. Note that this consistent estimator of σ^2 substitutes the IV estimates of the coefficients into the <u>original</u> equation, and uses the original values of the X variables to form the residuals. When working with the 2SLS estimator, and calculating it by running the two OLS regression stages, you might be tempted to estimate σ^2 using a regression program printed values of SSR or the variance of the second stage regression, which is based on the residuals $\hat{u} = y - X^*b_{2SLS}$. It tuns out that this estimator is <u>not</u> consistent for σ^2: A few lines of matrix manipulation shows that $\hat{u}'\hat{u}/n \rightarrow_p \sigma^2 + \beta'[A - GF^{-1}G']\beta$. The second term is positive semidefinite, so this estimator is asymptotically biased upward.

Suppose $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$, so that 2SLS is best among IV estimators using instruments W. The sum of squared residuals SSR = u′u, where $u = y - Xb_{2SLS}$, can be used in hypothesis testing in the same way as in OLS estimation. For example, consider the hypothesis that $\beta_2 = 0$, where β_2 is a r×1 subvector of β. Let SSR_0 be the sum of squared residuals from the 2SLS regression of y on X with $\beta_2 = 0$ imposed, and SSR_1 be the sum of squared residuals from the unrestricted 2SLS regression of y on X. Then, $[(SSR_0 - SSR_1)/m]/[SSR_1/(n-k)]$ has an approximate F-distribution under the null with m and n-k degrees of freedom. There are several cautions to keep in mind when considering use of this test statistic. This is a large sample approximation, rather than an exact distribution, because it is derived from the asymptotic normality of the 2SLS estimator. Its actual size in small samples could differ substantially from its nominal (asymptotic) size. Also, the large sample distribution of the statistic assumed that the disturbances ε have a scalar covariance matrix. Otherwise, it is mandatory to do a FGLS transformation before computing the test statistic above. For example, if $y = X\beta + \varepsilon$ represents a stacked system of equations such as structural equations in a simultaneous

equations system, or if ε exhibits serial correlation, as may be the case in time-series or panel data, then one should estimate β consistently using 2SLS, retrieve the residuals $u = y - Xb_{2SLS}$ and use them to make an estimate Ω^* of $\Omega = \mathbf{E}\varepsilon\varepsilon'$, make the transformations $y = Ly$, $X = LX$, $\nu = L\varepsilon$, and $W = (L')^{-1}W$ where L is a Cholesky matrix such that $L\Omega^*L'$ is proportional to an identity matrix, and finally apply 2SLS to the regression $y = X\beta + \nu$ with W as instruments and carry out the hypothesis testing using this model. The reason for the particular transformation of W is that one has $W'\nu = W'\varepsilon$, so that the original property that the instruments were uncorrelated with the disturbances is preserved. The 3SLS procedure just described corresponds to estimating β using a feasible version of the GLSIV estimator.

What are the finite sample properties of IV estimators? Because you do not have the condition $\mathbf{E}(\varepsilon|X) = 0$ holding in applications where IV is needed, you cannot get simple expressions for the moments of $b_{IV} = [R'W'X]^{-1}R'W'y = \beta + [R'W'X]^{-1}R'W'\varepsilon$ by first taking expectations of ε conditioned on X and W. In particular, you cannot conclude that b_{IV} is unbiased, or that it has a covariance matrix corresponding to its asymptotic covariance matrix. In fact, b_{IV} can have very bad small-sample properties. To illustrate, consider the case where the number of instruments equals the number of observations, j = n. (This can actually arise in dynamic models, where often all lagged values of the exogenous variables are legitimate instruments. It can also arise when the candidate instruments are not only uncorrelated with ε, but satisfy the stronger property that $\mathbf{E}(\varepsilon|w) = 0$. In this case, all functions of w are also legitimate instruments.) In this case, W is a square matrix, and

$$b_{2SLS} = [X'W(W'W)^{-1}W'X]^{-1}X'W(W'W)^{-1}W'y$$
$$= [X'WW^{-1}W'^{-1}W'X]^{-1}X'WW^{-1}W'^{-1}W'y = [X'X]^{-1}X'y = b_{OLS}.$$

We know OLS is inconsistent when $\mathbf{E}(\varepsilon|X) = 0$ fails, so clearly the 2SLS estimator is also biased if we let the number of instruments grow linearly with sample size. This shows that for the IV asymptotic theory to be a good approximation, n must be much larger than j. One rule-of-thumb for IV is that n - j should exceed 40, and should grow linearly with n in order to have the large-sample approximations to the IV distribution work well.

Considerable technical analysis is required to characterize the finite-sample distributions of IV estimators analytically; the names associated with this problem are Nagar, Phillips, and Mariano. However, simple numerical examples provide a picture of the situation. Consider first a regression $y = x\beta + \varepsilon$ where there is a single right-hand-side variable, and a single instrument w, and assume x, w, and ε have the simple joint distribution given in the table below, where λ is the correlation of x and w, ρ is the correlation of x and ε, and $0 \leq \lambda,\rho$ and $\lambda + 2\rho < 1$:

x	w	ε	Prob
1	1	1	$(1+\lambda)/8$
-1	1	1	$(1-\lambda)/8$
1	-1	1	$(1-\lambda+2\rho)/8$
-1	-1	1	$(1+\lambda-2\rho)/8$
1	1	-1	$(1+\lambda)/8$
-1	1	-1	$(1-\lambda)/8$
1	-1	-1	$(1-\lambda-2\rho)/8$
-1	-1	-1	$(1+\lambda+2\rho)/8$

These random variables then satisfy $\mathbf{E}x = \mathbf{E}w = \mathbf{E}\varepsilon = 0$, $\mathbf{E}x\varepsilon = \rho$, $\mathbf{E}xw = \lambda$, and $\mathbf{E}w\varepsilon = 0$, and their products have the joint distribution

xw	wε	xε	Prob
1	1	1	$(1+\lambda+\rho)/4$
-1	-1	1	$(1-\lambda+\rho)/4$
-1	1	-1	$(1-\lambda-\rho)/4$
1	-1	-1	$(1+\lambda-\rho)/4$

Least squares is biased if $\rho \neq 0$, and IV is consistent if $\lambda \neq 0$. Suppose n = 2. Then the exact distribution of the relevant random variables is

$\sum xw$	$\sum w\varepsilon$	$\sum x\varepsilon$	$b_{OLS}-\beta$	$b_{IV}-\beta$	Prob
2	2	2	1	1	$(1+\lambda+\rho)^2/16$
0	0	2	1	0	$((1+\rho)^2-\lambda^2)/8$
0	2	0	0	$+\infty$	$(1-(\lambda+\rho)^2)/8$
2	0	0	0	0	$((1+\lambda)^2-\rho^2)/8$
-2	-2	2	1	1	$(1-\lambda+\rho)^2/16$
-2	0	0	0	0	$((1-\lambda)^2-\rho^2)/8$
0	-2	0	0	$-\infty$	$(1-(\lambda-\rho)^2)/8$
-2	2	-2	-1	-1	$(1-\lambda-\rho)^2/16$
0	0	-2	-1	0	$((1-\rho)^2-\lambda^2)/8$
2	-2	-2	-1	-1	$(1+\lambda-\rho)^2/16$

Note first that there is a positive probability that b_{IV} is not defined; hence, technically it has no finite moments. Collecting terms from this table, the exact CDF of $b_{OLS} - \beta$ and $b_{IV} - \beta$ satisfy

c	$\text{Prob}(b_{OLS}-\beta \leq c)$	$\text{Prob}(b_{IV}-\beta \leq c)$
$-\infty$	0	$(1-(\lambda-\rho)^2)/8$
-1	$(1-\rho)^2/4$	$(1-\lambda(1-\rho))/4$
0	$(1-\rho)(3+\rho)/4$	$(3-\lambda(1-\rho))/4$
1	1	$(\lambda+\rho)^2/2$
$+\infty$	1	1

Also, $\text{Prob}(|b_{IV}-\beta| > |b_{OLS}-\beta|) = (1-\lambda^2-\rho^2)/4$. Then for this small sample there is a substantial probability that the IV estimator will be further away from the true value than the OLS estimator. As an exercise, carry through this example for n = 3, and show that in this case b_{IV} will always exist, but there continues to be a large probability that b_{OLS} is closer to β than b_{IV}. As n increases, the probability that b_{OLS} is closer than b_{IV} shrinks toward zero, but there is always a positive probability that the IV estimator is worse than the OLS estimator, and for n odd a positive probability that the IV estimator is infinite, so it never has any finite moments.

The second example is the one-variable model $y = x\beta + \varepsilon$ with one instrument w where (x,w,ε) are jointly normal with zero means, unit variances, $\mathbf{E}wx = \lambda$, $\mathbf{E}x\varepsilon = \rho$, and $\mathbf{E}w\varepsilon = 0$. A difficult

technical analysis can be used to derive the exact distribution of the IV estimator in terms of a non-central Wishart distribution. However, for purposes of getting an idea of how IV performs, it is much simpler to do a small computer simulation. For the values $\rho = .2$ and $\lambda = .8$, the table below gives the results of estimating a true value $\beta = 1$ in 1000 samples of sizes n = 5, 10, 20, or 40. Because the denominator in the IV estimator is small with some probability, the IV estimator tends to produce large deviations that lead to a large mean square error (MSE). In this example, the probability that the IV estimator is closer to β than the OLS estimator exceeds 0.5 only for samples of size 20 or greater, and the IV estimator has a smaller MSE only for samples of size 40 or larger. The smaller ρ or λ, the larger the sample size needed to make IV better than OLS in terms of MSE.

Sample Size	Mean Bias in b_{OLS} (1000 samples)	Mean Bias in b_{IV} (1000 samples)	MSE of b_{OLS} (1000 samples)	MSE of b_{IV} (1000 samples)	Frequency of b_{IV} as good as b_{OLS} (1000 samples)
5	0.18	-0.15	0.25	63.5	39.6%
10	0.19	-0.04	0.15	0.70	45.7%
20	0.20	-0.02	0.09	0.10	54.6%
40	0.20	-0.00	0.07	0.04	69.2%

In practice, in problems where sample size minus the number of instruments exceeds 40, the asymptotic approximation to the distribution of the IV estimator is reasonably good, and one can use it to compare the OLS and IV estimates. To illustrate, continue the example of a regression in one variable, $y = x\beta + \varepsilon$. Suppose as before that x and ε have a correlation coefficient $\rho \neq 0$, so that OLS is biased, and suppose that there is a single proper instrument w that is uncorrelated with ε and has a correlation $\lambda \neq 0$ with x. Then, the OLS estimator is asymptotically normal with mean $\beta + \rho\sigma_\varepsilon/\sigma_x$ and variance $\sigma_\varepsilon^2/n\sigma_x^2$. The 2SLS estimator is asymptotically normal with mean β and variance $\sigma_\varepsilon^2/n\sigma_x^2\lambda^2$. The mean squares of the two estimators are then, approximately,

$$MSE_{OLS} = (\rho^2 + 1/n)\sigma_\varepsilon^2/\sigma_x^2$$
$$MSE_{2SLS} = \sigma_\varepsilon^2/n\sigma_x^2\lambda^2.$$

Then, 2SLS has a lower MSE than OLS when

$$1 < \rho^2\lambda^2 n/(1-\lambda^2) \approx (b_{2SLS}-b_{OLS})^2/(V(b_{2SLS})-V(b_{OLS})),$$

or approximately $n > (1 - \lambda^2)/\rho^2\lambda^2$. When $\lambda = 0.8$ and $\rho = 0.2$, this asymptotic approximation suggests that a sample size of about 14 is the tip point where b_{IV} should be better than b in terms of MSE. However, the asymptotic formula underestimates the probability of very large deviations arising from a denominator in b_{IV} that is near zero, and as a consequence is too quick to reject b_{OLS}. The right-hand-side of this approximation to the ratio of the MSE is the Hausman test statistic for exogeneity, discussed below; for this one-variable case, one should reject the null hypothesis of exogeneity when the statistic exceeds one. Under the null, the statistic is approximately chi-square with one degree of freedom, so that this criterion corresponds to a type I error probability of 0.317.

4. RELATION OF IV TO OTHER ESTIMATORS

The 2SLS estimator can be interpreted as a member of the family of *Generalized Method of Moments* (GMM) estimators. You can verify by differentiating to get the first-order condition that the 2SLS estimator of the equation $Ly = LX\beta + L\varepsilon$ using the instruments $(L')^{-1}W$, where $\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$ and L is a Cholesky matrix satisfying $L\Omega L' = I$, solves

$$\text{(14)} \qquad \text{Min}_{\beta}\ (y-X\beta)'W(W'\Omega W)^{-1}W'(y-X\beta).$$

In this quadratic form objective function, $W'(y-X\beta)$ is the moment that has expectation zero in the population when β is the true parameter vector, and $(W'\Omega W)^{-1}$ is a "distance metric" in the center of the quadratic form. Define $P = (L')^{-1}W(W'\Omega W)^{-1}W'(L)^{-1}$, and note that *P* is idempotent, and thus is a projection matrix. Then, the GMM criterion chooses β to minimize the length of the vector $L(y-X\beta)$ projected onto the subspace spanned by *P*. The properties of GMM hypothesis testing procedures follow readily from the observation that $L(y-X\beta)$ has mean zero and a scalar covariance matrix. In particular, $\text{Min}_{\beta}\ (y-X\beta)'W(W'\Omega W)^{-1}W'(y-X\beta)/\sigma^2$ is asymptotically chi-squared distributed with degrees of freedom equal to the rank of *P*.

It is possible to give the 2SLS estimator a *pseudo-MLE* interpretation. Premultiply the regression equation by $W'L^{-1}$ to obtain $W'y = W'X\beta + W'\varepsilon$. Now treat $W'\varepsilon$ *as if* it were normally distributed with mean zero and j×j covariance matrix $\lambda^2W'\Omega W$, conditioned on $W'X$. Then, the log likelihood of the sample would be

$$L = -(j/2)\log 2\pi - (j/2)(\tfrac{1}{2})\log\lambda^2 - (\tfrac{1}{2})\log\det(W'\Omega W) - (1/2\lambda^2)(W'y-W'X\beta)'(W'\Omega W)^{-1}(W'y-W'X\beta).$$

The first-order condition for maximization of this pseudo-likelihood is the same as the condition defining the 2SLS estimator.

5. TESTING EXOGENEITY

Sometimes one is unsure whether some potential instruments are clean. If they are, then there is an asymptotic efficiency gain from including them as instruments. However, if they are not, estimates will be inconsistent. Because of this tradeoff, it is useful to have a specification test that permits one to judge whether suspect instruments are clean or not. To set the problem, consider a regression $y = X\beta + \varepsilon$, an array of proper instruments Z, and an array of instruments W that includes Z plus other variables that may be either clean or contaminated.

Several superficially different problems can be recast in this framework:

(1) The regression may be one in which some right-hand-side variables are known to be exogenous and others are suspect, Z is an array that contains the known exogenous variables and other clean instruments, and W contains Z and the variables in X that were excluded from Z because of the possibility that they might be dirty. In this case, 2SLS using W reduces to OLS, and the problem is to test whether the regression can be estimated consistently by OLS.

(2) The regression may contain known endogenous and known exogenous variables, Z is an array that contains the known exogenous variables and other clean instruments, and W is an array that contains Z and additional suspect instruments from outside the equation. In this case, one has a consistent 2SLS estimator using instruments Z, and a 2SLS estimator using instruments W that is more efficient under the hypothesis that W is exogenous, but inconsistent otherwise. The question is whether to use the more inclusive array of instruments.

(3) The regression may contain known endogenous, known exogenous, and suspect right-hand-side variables, Z is an array that contains the known exogenous variables plus other instruments from outside the equation, and W is an array that contains Z plus the suspect variables from the equation. The question is whether it is necessary to instrument for the suspect variables, or whether they are clean and can themselves be used as instruments.

In the regression $y = X\beta + \varepsilon$, you can play it safe and use only the Z instruments. This gives $b_Q = (X'QX)^{-1}X'Qy$, where $Q = (L')^{-1}Z(Z'\Omega Z)^{-1}Z'(L)^{-1}$. Alternately, you use W, including the suspect instruments, taking a chance with inconsistency to gain efficiency. This gives

$$b_P = (X'PX)^{-1}X'Py, \text{ where } P = (L')^{-1}W(W'\Omega W)^{-1}W'(L)^{-1}.$$

If the suspect instruments are clean and both estimators are consistent, then b_Q and b_P should be close together, as they are estimates of the same β; further, b_P is efficient relative to b_Q, implying that the covariance matrix of $(b_Q - b_P)$ equals the covariance matrix of b_Q minus the covariance matrix of b_P. However, if the suspect instruments are contaminated, b_P is inconsistent, and $(b_Q - b_P)$ has a nonzero probability limit. This suggests a test statistic of the form

$$(15) \qquad (b_Q - b_P)'[V(b_Q) - V(b_P)]^{-}(b_Q - b_P),$$

where $[\cdot]^{-}$ denotes a generalized inverse, could be used to test if W is clean. This form is the exogeneity test originally proposed by Hausman. Under the null hypothesis that W is clean, this statistic will be asymptotically chi-square with degrees of freedom equal to the rank of the covariance matrix in the center of the quadratic form.

Another formulation of an exogeneity test is more convenient to compute, and can be shown (in one manifestation) to be equivalent to the Hausman test statistic. This alternative formulation has the form of an omitted variable test, with appropriately constructed auxiliary variables. We describe the test in the case $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$ and leave as an exercise the extension to the case $\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$.

First do an OLS regression of X on Z and retrieve fitted values $X^* = QX$, where $Q = Z(Z'Z)^{-1}Z'$. (This is necessary only for variables in X that are not in Z, since otherwise this step just returns the original variable.) Second, using W as instruments, do a 2SLS regression of y on X, and retrieve the sum of squared residuals SSR_1. Third, do a 2SLS regression of y on X and a subset of m columns of X^* that are linearly independent of X, and retrieve the sum of squared residuals SSR_2. Finally, form the statistic $[(SSR_1 - SSR_2)/m]/[SSR_2/(n-k)]$. Under the null hypothesis that W is clean, this statistic has an approximate F-distribution with m and n-k degrees of freedom, and can be interpreted as a test for whether the m auxiliary variables from X^* should be omitted from the regression. When a subset of X^* of maximum possible rank is chosen, this statistic turns out to be asymptotically equivalent to the Hausman test statistic. Note that if W contains X, then the 2SLS in the second and third steps reduces to OLS.

We next show that this test is indeed an exogeneity test. Consider the 2SLS regression

$$y = X\beta + X_1^*\gamma + \eta,$$

where X_1^* is a subset of $X^* = QX$ such that $[X,X_1^*]$ is of full rank. The 2SLS estimates of the parameters in this model, using W as instruments, satisfy

$$\begin{bmatrix} b_P \\ c_P \end{bmatrix} = \begin{bmatrix} X'PX & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'Py \\ X_1'Qy \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix} + \begin{bmatrix} X'PX & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'P\varepsilon \\ X_1'Q\varepsilon \end{bmatrix}.$$

But $X'Q\varepsilon/n \to_p \text{plim}(X'Z/n)\cdot(\text{plim}(Z'Z/n))^{-1}\cdot\text{plim}(Z'\varepsilon/n) = 0$ by assumptions [1]-[4] when Z is clean. Similarly, $X'P\varepsilon/n \to_p GH^{-1}\cdot\text{plim}(W'\varepsilon/n) = 0$ when W is clean, but $X'P\varepsilon/n \to_p GH^{-1}\cdot\text{plim}(W'\varepsilon/n) \neq 0$ when W is contaminated. Define

$$\begin{bmatrix} X'PX/n & X'QX_1/n \\ X_1'QX/n & X_1'QX_1/n \end{bmatrix}^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

From the formula for a partitioned inverse,

$$A_{11} = (X'[P - QX_1(X_1'QX_1)^{-1}X_1'Q]X/n)^{-1}$$

$$A_{22} = (X_1'Q[I - X(X'PX)^{-1}X']QX_1/n)^{-1}$$

$$A_{21} = -(X_1'QX_1)^{-1}X_1'QX\cdot A_{11} = -A_{22}(X_1'QX)(X'PX)^{-1} = A_{12}'$$

Hence,

$$c_P = A_{22}\cdot\{X_1'Q\varepsilon/n - (X_1'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\}. \quad (16)$$

If W is clean and satisfies assumptions [4] and [5], then $c_P \to_p 0$ and $n^{1/2}c_P$ is asymptotically normal. On the other hand, if W is contaminated, then c_P has a non-zero probability limit. Then, a test for $\gamma = 0$ using c_P is a test of exogeneity.

The test above can be reinterpreted as a Hausman test involving differences of b_P and b_Q. Recall that $b_Q = \beta + (X'QX)^{-1}X'Q\varepsilon$ and $b_P = \beta + (X'PX)^{-1}X'P\varepsilon$. Then

$$(X'QX)(b_Q - b_P) = \{X'Q\varepsilon/n - (X'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\}. \quad (17)$$

Then in particular for a linearly independent subvector X_1 of X,

$$A_{22}(X_1'QX)(b_Q - b_P) = A_{22}\{X_1'Q\varepsilon/n - (X_1'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\} = c_P.$$

Thus, c_P is a linear transformation of $(b_Q - b_P)$. Then, testing whether c_P is near zero is equivalent to testing whether a linear transformation of $(b_Q - b_P)$ is near zero. When X_1 is of maximum rank, this equivalence establishes that the Hausman test in its original form is the same as the test for c_P.

6. EXOGENICITY TESTS ARE GMM TESTS FOR OVERIDENTIFICATION

The Hausman Exogeneity Test. Consider the regression model $y = X\beta + \varepsilon$, and suppose one wants to test the exogeneity of p variables X_1 in X. Suppose R is an array of instruments, including X_2; then $Z = P_R X_1$ are instruments for X_1. Let $W = [Z\ X]$ be all the variables that are orthogonal to

ε in the population under the null hypothesis that X and ε are uncorrelated. As in the omitted variables problem, consider the test statistic for over-identifying restrictions, $2nQ_n = \min_b u'P_W u/\sigma^2$, where $u = y - Xb$. Decompose $P_W = P_X + (P_W - P_X)$. Then $u'(P_W - P_X)u = y'(P_W - P_X)y$ and the minimizing b sets $u'P_X u = 0$, so that $2nQ_n = y'(P_W - P_X)y/\sigma^2$. Since $P_W - P_X = P_{Q_X W}$, one also has $2nQ_n = y' P_{Q_X W} y$. This statistic is the same as the test statistic for the hypothesis that the coefficients of Z are zero in a regression of y on X and Z; thus the test for over-identifying restrictions is an omitted variables test. One can also write $2nQ_n = \|\hat{y}_W - \hat{y}_X\|^2/\sigma^2$, so that a computationally convenient equivalent test is based on the difference between the fitted values of y from a regression on X and Z and a regression on X alone. Finally, we will show that the statistic can be written

$$2nQ_n = (b_{1,2SLS} - b_{1,OLS})[V(b_{1,2SLS}) - V(b_{1,OLS})]^{-1}(b_{1,2SLS} - b_{1,OLS}).$$

In this form, the statistic is the Hausman test for exogenicity in the form developed by Hausman and Taylor, and the result establishes that the Hausman test for exogeneity is equivalent to a GMM test for over-identifying restrictions.

Several steps are needed to demonstrate this equivalence. Note that $b_{2SLS} = (X'P_M X)^{-1}X'P_M y$, where $M = [Z\ X_2]$. Write

$$\begin{aligned} b_{2SLS} - b_{OLS} &= (X'P_M X)^{-1}X'P_M y - (X'X)^{-1}X'y \\ &= (X'P_M X)^{-1}[X'P_M - X'P_M X(X'X)^{-1}X']y \\ &= (X'P_M X)^{-1}X'P_M Q_X y. \end{aligned}$$

Since X_2 is in M, $P_M X_2 = X_2$, implying $X'P_M Q_X = \begin{bmatrix} X_1'P_M Q_X \\ X_2'P_M Q_X \end{bmatrix} = \begin{bmatrix} X_1'P_M Q_X \\ X_2'Q_X \end{bmatrix} = \begin{bmatrix} X_1'P_M Q_X \\ 0 \end{bmatrix}$.

Also, $X'P_M X = \begin{bmatrix} X_1'P_M X_1 & X_1'P_M X_2 \\ X_2'P_M X_1 & X_2'P_M X_2 \end{bmatrix} = \begin{bmatrix} X_1'P_M X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$. Then $\begin{bmatrix} X_1'P_M Q_X y \\ 0 \end{bmatrix} = (X'P_M X)(b_{2SLS} - b_{OLS}) \equiv \begin{bmatrix} X_1'P_M X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_{1,2SLS} - b_{1,OLS} \\ b_{2,2SLS} - b_{2,OLS} \end{bmatrix}$. From the second block of equations, one obtains the result that the second subvector is a linear combination of the first subvector. This implies that a test statistic that is a function of the full vector of differences of 2SLS and OLS estimates can be written equivalently as a function of the first subvector of differences. From the first block of equations, substituting in the solution for the second subvector of differences expressed in terms of the first, one obtains

$$[X_1'P_M X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1](b_{1,2SLS} - b_{1,OLS}) = X_1'P_M Q_X y$$

The matrix on the left-hand-side can be rewritten as $X_1'P_M\ Q_{X_2}\ P_M X_1$, so that

$$b_{1,2SLS} - b_{1,OLS} = (X_1'P_M \;\; Q_{X_2} \;\; P_MX_1)^{-1}X_1'P_MQ_Xy.$$

Next, we calculate the covariance matrix of $b_{2SLS} - b_{OLS}$, and show that it is equal to the difference of $V(b_{2SLS}) = \sigma^2(X'P_MX)^{-1}$ and $V(b_{OLS}) = \sigma^2(X'X)^{-1}$. From the formula $b_{2SLS} - b_{OLS} = (X'P_MX)^{-1}X'P_MQ_Xy$, one has $V(b_{2SLS} - b_{OLS}) = \sigma^2(X'P_MX)^{-1}X'P_MQ_XP_MX(X'P_MX)^{-1}$.
On the other hand,

$$\begin{aligned} V(b_{2SLS}) - V(b_{OLS}) &= \sigma^2(X'P_MX)^{-1}\{X'P_MX - X'P_MX(X'X)^{-1}X'P_MX\}(X'P_MX)^{-1} \\ &= \sigma^2(X'P_MX)^{-1}\{X'P_M[I - X(X'X)^{-1}X']P_MX\}(X'P_MX)^{-1} \\ &= \sigma^2(X'P_MX)^{-1}X'P_MQ_XP_MX(X'P_MX)^{-1}. \end{aligned}$$

Thus, $V(b_{2SLS} - b_{OLS}) = V(b_{2SLS}) - V(b_{OLS})$. This is a consequence of the fact that under the null hypothesis OLS is efficient among the class of linear estimators including 2SLS. Expanding the center of this expression, and using the results $P_MX_2 = X_2$ and hence $Q_XP_MX_2 = 0$, one has

$$X'P_MQ_XP_MX = \begin{bmatrix} X_1'P_MQ_XP_MX_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence, $V(b_{2SLS}) - V(b_{OLS})$ is of rank p; this also follows by noting that $b_{2,2SLS} - b_{2,OLS}$ could be written as a linear transformation of $b_{1,2SLS} - b_{1,OLS}$.

Next, use the formula for partitioned inverses to show for N = M or N = I that the northwest corner of $\begin{bmatrix} X_1'P_NX_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}$ is $(X_1'P_NQ_{X_2}P_NX_1)^{-1}$. Then,

$$V(b_{1,2SLS} - b_{1,OLS}) = \sigma^2(X_1'P_M \;\; Q_{X_2} \;\; P_MX_1)^{-1}X_1'P_MQ_XP_MX_1(X_1'P_M \;\; Q_{X_2} \;\; P_MX_1)^{-1}.$$

Using the expressions above, the quadratic form can be written

$$\begin{aligned} &(b_{1,2SLS} - b_{1,OLS})V(b_{1,2SLS} - b_{1,OLS})^{-1}(b_{1,2SLS} - b_{1,OLS}) \\ &= y'Q_XP_MX_1(X_1'P_MQ_XP_MX_1)^{-1}X_1'P_MQ_Xy/\sigma^2. \end{aligned}$$

Finally, one has, from the test for over-identifying restrictions,

$$2nQ_n = y'(P_W - P_X)y/\sigma^2 = \; y'P_{Q_XW}y \; /\sigma^2$$

$$\equiv y'Q_XP_MX_1(X_1'P_MQ_XP_MX_1)^{-1}X_1'P_MQ_Xy/\sigma^2,$$

so that the two statistics coincide.

A Generalized Exogenicity Test: Consider the regression $y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon$, and the null hypothesis that X_1 is exogenous, where X_2 is known to be exogenous, and X_3 is known to be endogenous. Suppose N is an array of instruments, including X_2, that are sufficient to identify the coefficients when the hypothesis is false. Let $W = [N \; X_1]$ be the full set of instruments available when the null hypothesis is true. Then the best instruments under the null hypothesis are $X_o = P_WX \equiv [X_1 \;\; X_2 \;\; X_3^*]$, and the best instruments under the alternative are $X_u = P_NX \equiv [X_1^* \;\; X_2 \;\; X_3^*]$. The

test statistic for over-identifying restrictions is $2nQ_n = y'(P_{X_o} - P_{X_u})y/\sigma^2$, as in the previous cases. This can be written $2nQ_n = (SSR_{X_o} - SSR_{X_u})/\sigma^2$, with the numerator the difference in sum of squared residuals from a OLS regression of y on X_u and a OLS regression of y on X_o. Also, $2nQ_n = \| \hat{y}_{X_o} - \hat{y}_{X_u} \|^2/\sigma^2$, the difference between the fitted values of y from a regression on X_u and a regression on X_o. Finally,

$$2nQ_n = (b_{2SLS_o} - b_{2SLS_u})'[V(b_{2SLS_u}) - V(b_{2SLS_o})]^{-} (b_{2SLS_o} - b_{2SLS_u}),$$

an extension of the Hausman-Taylor exogeneity test to the problem where some variables are suspect and others are known to be exogenous. One can show that the quadratic form in the center of this quadratic form has rank equal to the rank of X_1, and that the test statistic can be written equivalently as a quadratic form in the subvector of differences of the 2SLS estimates for the X_1 coefficients, with the ordinary inverse of the corresponding submatrix of differences of variances in the center of the quadratic form.

7. INSTRUMENTAL VARIABLES IN TIME-SERIES MODELS

The treatment of IV estimation up to this point applies in principle to observations made either in cross section or over time. For example, if the observations correspond to time periods and $\mathbf{E}(\varepsilon\varepsilon' | W) = \sigma^2\Omega$ with Ω either known or estimated, the 2SLS estimator (2) or the two-stage feasible generalized least squares estimator (10) with Ω estimated using residuals obtained by application of (2), can be applied to problems where the structure of Ω comes from serial correlation. However, for time series applications it is useful to examine in more detail the structure of W and the orthogonality conditions used in forming IV estimators. In particular, one should ask how conventional sources of contamination in explanatory variables such as omitted variables or measurement error and conventional sources of serial correlation such as behavioral lags in adjustment are likely to affect the serial correlation structure of disturbances and the correlation of contemporaneous disturbances with explanatory variables for various transformations of the model.

Start with the example of a linear model with measurement error in explanatory variables, and suppose that in the absence of this measurement error problem the disturbance in the equation would follow an AR1 process. Let z_t denote the ideal variables without measurement error, and $x_t = z_t + \eta_t$ denote the observed explanatory variables. Then, the model can be written

$$y_t = z_t\beta + \varepsilon_t \text{ with } \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t,$$

or

$$y_t = x_t\beta + \nu_t - \eta_t\beta + \rho\nu_{t-1} + \rho^2\nu_{t-2} + ..., \tag{18}$$

where the ν_t are i.i.d. innovations and $\rho^2 < 1$. This model can also be written

$$y_t = y_{t-1}\rho + x_t\beta - x_{t-1}\beta\rho + (\nu_t - \eta_t\beta + \eta_{t-1}\beta\rho). \tag{19}$$

The form (19) removes the serial correlation in the ideal equation disturbance, but in doing so introduces a moving average of the measurement errors. Only in the unlikely case that all components of η_t follow an AR1 process with the same ρ as the ε_t process will serial correlation be fully removed.[12] Application of OLS to either (18) or (19) will then in general result in inconsistent estimates. The issue for application of IV methods is whether proper instruments can be found. In (18), the variables in x_t that are measured with error would require instrumenting. If the z_t are serially correlated, and the η_t are not, then x_{t-1}, x_{t-2},... are potential clean instruments for x_t. However, if there is serial correlation in the measurement errors, one would need to find proper instruments from outside the model. In (19), all of the explanatory variables y_{t-1}, x_t, and x_{t-1} are contaminated, but if the z_t are correlated with a sufficiently long lag and the η_t are uncorrelated, then x_{t-2}, x_{t-3}, x_{t-4},... are potential clean instruments. It is important to not introduce x's with too high lags as instruments, because this requires truncating the sample in order to observe the instruments for each date used in the estimation, and the good statistical properties of the IV method begins to break down as the number of instruments ceases to be small relative to the remaining sample size.

Omitted variables leads to models similar to (18) and (19). In this case, interpret the disturbance in the model $y_t = x_t\beta + \varepsilon_t$ as including the omitted variables. If these omitted variables are themselves serially correlated, then they will induce serial correlation in ε_t, perhaps adding to serial correlation in a disturbance component that arises for reasons other than omitted variables. A transformation of the model in this case may be able to remove serial correlation in the disturbance, but does not remove the contamination. The issue will be to find proper instruments. If the included x's are themselves serially correlated and the final disturbance is AR1, then the equation $y_t = y_{t-1}\rho + x_t\beta - x_{t-1}\beta\rho + \varepsilon_t - \rho\varepsilon_{t-1}$ obtained by partial differencing will have y_{t-1}, x_{t-1}, x_{t-2},... as potential clean instruments. For this to work, the AR1 specification for ε_t must be correct, and x_t must not have the same AR1 process.

The preceding examples illustrate several important points about the use of IV methods in time-series models. First, there is likely to be an interaction between the source of the contamination and the nature of the serial correlation in the model. Second, the process followed by the explanatory variables will determine what variables are clean (i.e., uncorrelated with the contemporaneous disturbance) and what variables might be available as instruments. Third, choice of instruments is not clear-cut, and may involve the question of what variables are potential clean instruments and how many potential instruments to introduce given the fairly poor small sample properties of IV. The use of lags of y_t or x_t as instruments exacerbates the sample size problem, since it decreases the operating sample size as the number of instruments rises. Further, lagged variables may fail to be proper instruments, either because assumptions of zero correlation are not robust and fail due to a more complex pattern of serial correlation than the econometrician assumes, or because these lagged variables are not correlated with the variables they are instrumenting. Together, these observations suggest that careful consideration of the nature of contamination and serial correlation is needed in time-series applications of IV, and that this method be used with caution.

[12]The situation in which all the variables in a model follow the same AR process does has some chance of arising in stationary state equilibria, because equilibrium pressures may force all variables to move nearly in lock-step along a dynamic path determined by the largest root of the system.

8. INSTRUMENTAL VARIABLES IN NONLINEAR MODELS

The method of instrumental variables in its most commonly used 2SLS form is applied to models linear in variables and in parameters, $y = X\beta + \varepsilon$. If there are proper instruments W for X and if $\mathbf{E}(\varepsilon|W) = \sigma^2\mathbf{I}$, then the 2SLS estimator (2) is consistent for β and efficient among all IV estimators using these instruments; see the theorem in Section 3. However, the orthogonality conditions invoked to justify the IV method do not necessarily extend to nonlinear transformations, because expectations are not preserved. For example, economic applications may postulate a zero correlation between variables for behavioral reasons, such as the rational expectations hypothesis that intertemporally optimized consumption is a random walk whose innovations are uncorrelated with history. This is not sufficient to guarantee that innovations in a nonlinear transformation of consumption are uncorrelated with history. To investigate what happens without linearity, consider three cases of nonlinearity:

(a) Models nonlinear in parameters only: $y = x\beta(\theta) + \varepsilon$
(b) Models nonlinear in variables only: $y = f(x)\beta + \varepsilon$
(c) Models nonlinear in both variables and parameters: $y = h(x,\theta) + \varepsilon$

A case such as (a) might arise for example when partial differencing is done to handle AR1 serial correlation. In this case, $y = x\alpha + \eta$ and $\eta = \rho\eta_{-1} + \nu$ with ν i.i.d., and transformation yields $y = \rho y_{-1} + x\alpha - x_{-1}\alpha\rho + \nu$, a model that has i.i.d. disturbances, but the parameters α and ρ appearing in nonlinear combination. Suppose in the model (a) that one first does an OLS regression of x on proper instruments w, and retrieves fitted values x^*, and second does a nonlinear least squares regression for the model $y = x^*\beta(\theta) + \varepsilon^*$. Examine the first-order conditions for the last regression, and show as an exercise that orthogonality of the instruments and the disturbances in the original regression implies consistency, just as in the fully linear case.[13] , It is the linearity of the first-order condition in the instruments and in ε that guarantees that the initial condition that the instruments be uncorrelated with ε continues to suffice.

Next consider the case $y = f(x)\beta + \varepsilon$ with nonlinear transformation of the explanatory variables but linearity in parameters. If instruments w are available that are uncorrelated with ε and fully correlated with f(x), then GMM estimation using the criterion function

$$(20) \qquad \left[\sum_{i=1}^{N} w_i(y_i - f(x_i)\beta)\right]' \cdot \left[\sum_{i=1}^{N} w_i w_i'\right]^{-1} \cdot \left[\sum_{i=1}^{N} w_i(y_i - f(x_i)\beta)\right] ,$$

will be consistent; see Chapter 3. Solution of this GMM problem can be given a 2SLS interpretation: First do an OLS regression of $f(x_i)$ on w_i, and retrieve fitted values f^*, then do an OLS regression of y_i on f^*. Then, the form and computation of the IV estimator are not affected by nonlinearity in variables. However, there are substantial issues regarding specification of the instruments. In particular, given an initial set of "raw" instruments z, should they be given nonlinear transformations to improve the efficiency of the IV estimator? An initial issue is whether postulated orthogonality of z and ε will be preserved for nonlinear transformations of z. This will depend on the economic application and the nature of z. If the application can guarantee only that z is

[13] The usual limiting regularity conditions are assumed to hold, as in Section 3, and the parameter θ is assumed to be identified in the sense the mapping from θ to β is one-to-one for β in its range.

uncorrelated with ε, this property will not in general be preserved under nonlinear transformation, and the only clean instruments w will be the untransformed z. However, if the application can guarantee that z is statistically independent of ε, then any nonlinear transformation of z will be uncorrelated with ε, and is a potential clean instrument. For the remainder of this section, assume that z and ε are statistically independent.

What transformations of z make good instruments? In some cases it is feasible to apply the nonlinear transformation f to z_i, and tempting to use $f(z_i)$ to instrument $f(x_i)$. For example, if x_i is a variable measured with error, and z_i is an independent measurement of the same variable, then provided one is persuaded that the error in z_i is statistically independent of ε_i, $f(z_i)$ seems to be a reasonable instrument for $f(x_i)$; e.g., $\log(z_i)$ seems to be a natural instrument for $\log(x_i)$. This is a practical thing to do, and will often give a more precise IV estimator than one that just uses the raw instruments. However, it will not in general yield the most efficient possible IV estimator. The reason for this is the proposition that expectations are not preserved under nonlinear transformations.

The best instruments are given by the conditional expectation of $f(x_i)$ given z_i: $w^* \equiv \omega(z_i) = \mathbf{E}(f(x_i)|z_i)$. To see this, first observe that the asymptotic covariance matrix for the IV estimator using instruments w_i that are any specified transformations of z_i is

$$\sigma^2[(\mathbf{E}w'f(x))'(\mathbf{E}w'w)^{-1}(\mathbf{E}w'f(x))]^{-1}. \text{ But } \mathbf{E}w'f(x) = \mathbf{E}_z w'\mathbf{E}_{x|z}f(x) = \mathbf{E}_z w'w^*.$$

The asymptotic covariance matrix of this IV estimator can be written

$$\sigma^2[(\mathbf{E}w'w^*)'(\mathbf{E}w'w)^{-1}(\mathbf{E}w'w^*)]^{-1}.$$

If $w = w^*$, this covariance matrix reduces to $\sigma^2(\mathbf{E}w^{*\prime}w^*)^{-1}$. It is a standard exercise to show that $w = w^*$ minimizes the asymptotic variance. Let $F = \mathbf{E}w^{*\prime}w^*$, $G = \mathbf{E}w'w^*$, and $H = \mathbf{E}w'w$. Then the quadratic form

$$[I \ -G'H^{-1}]\cdot \begin{bmatrix} F & G' \\ G & H \end{bmatrix} \cdot[I \ -G'H^{-1}]' = F - G'H^{-1}G$$

is positive semidefinite, which implies that $[G'H^{-1}G]^{-1} - F^{-1}$ is positive semidefinite. From this result, the IV estimator using the instruments w^* is called the best nonlinear 2SLS estimator (BN2SLS).

In general, the BN2SLS estimator is not practical in applications because computation of the conditional expectation $\mathbf{E}_{x|z}f(x)$ is intractable. Obviously, in any application where direct computation of $\mathbf{E}_{x|z}f(x)$ is tractable, it should be used. In the remaining cases, it is possible to approximate $\mathbf{E}_{x|w}f(x)$. A method proposed by Kelejian (1971) and Amemiya (1974) is to make an approximation in terms of low-order polynomials in the raw instruments z; i.e., regress $f(x_i)$ on z_i, squares and cross-products of components of z_i, third-order interactions, and so forth. One interpretation of this procedure is that one is making a series approximation using the leading terms in a Taylor's expansion of $\mathbf{E}_{x|w}f(x)$, or in other words the low order conditional moments of x given w. This method can be implemented in the LSQ procedure in TSP by expanding the list of specified instruments in the command to include the desired low-order polynomials in the raw instruments. Viewed more generally, the expression $\mathbf{E}_{x|z}f(x)$ can be written as

$$\mathbf{E}_{x|z}f(x) = \int_x f(x)\cdot g(x|z)\cdot dx \equiv \psi(z), \tag{21}$$

where g(x,z) is the joint density of x and z, and g(x|z) is the conditional density of x given z. If g(x|z) is known (or can be estimated consistently as a parametric function), but analytic computation of the integral is intractable, it may be possible to use simulation methods, drawing a "pseudo-sample" x_{ij} from g(x|z_i) for j = 1,...,J and estimating $\mathbf{E}_{x|z}f(x)$ as the mean of $f(x_{ij})$ in this pseudo-sample. If the pseudo-sample size J grows at a sufficient rate with sample size (typically, faster than $N^{1/2}$), then IV using this approximation will have the same asymptotic covariance matrix as BN2SLS. If the conditional density is itself not known or tractable, it may be possible to estimate it nonparametrically, say using a kernel estimator; see Chapter 7. Alternately, viewing $\psi(z)$ as a nonparametric function of z, the problem can be approached as a nonparametric regression $f(x_i) = \psi(z_i) + \zeta_i$, and ψ estimated by a variety of nonparametric procedures; again see Chapter 7. In particular, one approach to nonparametric regression is series approximation, where $\psi(z_i)$ is approximated by a linear combination of initial terms in a series approximation. In particular, the Kelejian-Amemiya method falls within this class, and nonparametric estimation theory provides a guide to choice of the truncation level as a function of sample size. The bottom line is that by simulation or nonparametric procedures, one may be able to "adaptively" achieve the asymptotic covariance matrix of the BN2SLS estimator without having to solve an intractable problem of determining $\mathbf{E}_{x|z}f(x)$ analytically. Existing software may not be sufficiently "adaptive" to automatically achieve the BN2SLS asymptotic efficiency level, so that it is up to the user to specify instruments in a form that achieves this adaptation. In practice, the issue of adaptiveness has no real bite in determining a good set of instruments in a given finite data set, and the properties of the asymptotic approximation may not tell you much about the actual finite-sample distribution of your estimators. Bootstrap methods, discussed in Chapter 7, may be one useful way to give a better approximation to finite-sample distributions and guide choice among estimators using different sets of instruments.

Finally, consider models that are nonlinear in both variables and parameters, $y = h(x,\theta) + \varepsilon$. First observe that if there are proper raw instruments z, then minimizing the GMM criterion

$$(22) \qquad \left[\sum_{i=1}^{N} z_i(y_i - h(x_i,\theta))\right]' \cdot \left[\sum_{i=1}^{N} z_i z_i'\right]^{-1} \cdot \left[\sum_{i=1}^{N} z_i(y_i - h(x_i,\theta))\right]$$

in θ will produce a consistent initial estimator θ_N for θ. There is an iterative procedure that can be used to calculate θ_N. From starting values $\theta^{(0)}$, suppose one has reached $\theta^{(r)}$. Linearize the model about $\theta^{(r)}$, obtaining

$$(23) \qquad y_i - h(x_i,\theta^{(r)}) = f^{(r)}(x_i)\cdot(\theta - \theta^{(r)}) + \upsilon_i,$$

where $f^{(r)}(x_i) = \nabla_\theta h(x_i,\theta^{(r)})$ and υ_i is a disturbance that includes the remainder from the linear approximation. Apply conventional 2SLS to this model, with the instruments z_i. The estimated coefficients provide the adjustments that produce the next iterate $\theta^{(r+1)}$. For a suitably chosen starting point, the iterates $\theta^{(r)}$ will converge to a limit at θ_N. It may be necessary to consider alternative starting values to obtain convergence to the minimand of the GMM criterion.

Start from the consistent initial estimator θ_N, and the linearized model (23) evaluated at θ_N, with $f_N(x) = \nabla_\theta h(x_i,\theta_N)$. Treating θ_N as a vector of constants, (23) now has the same form as the model that is nonlinear in variables but linear in parameters that was discussed above. As in the previous case, estimate this model using 2SLS and an approximation to the best instruments $\mathbf{E}_{x|z}f_N(x)$; this will approximate the BN2SLS estimator. This procedure, with the best instruments approximated by user-specified combinations of the raw instrumental variables, is used by the LSQ command in TSP.

It is possible to iterate the procedure described in this paragraph, but the first application of the procedure is already asymptotically equivalent to the BN2SLS estimator (provided the approximation to the best instruments is adaptive), and there is no further gain in (first-order) asymptotic efficiency from iteration.

CHAPTER 5. SYSTEMS OF REGRESSION EQUATIONS

1. MULTIPLE EQUATIONS

Consider the regression model setup

$$y_{nt} = x_{nt}\beta_n + u_{nt},$$

where n = 1,...,N, t = 1,...,T, x_{nt} is 1×k, and β_n is k×1. This is a version of the standard regression model where the observations are indexed by the two indices n and t rather than by a single index. Applications where this setup occurs are

- n indexes equations, with different dependent variables, and t indexes observation units. Example: $y_{1t},...,y_{nt}$ are the input demands of firm t. In this example, there are likely to be parameters in common across equations.
- n indexes observation units, t indexes time, and the data come from a time-series of cross-sections. Example: y_{nt} is the income of household n in the Census Public Use Sample in year t.
- n indexes observation units, t indexes time, and the data come from a longitudinal <u>panel</u> of time series observations on each observation unit. Examples: y_{nt} is hours supplied by the head of household n in year t in the Panel Study of Income Dynamics; <u>or</u> y_{nt} is the excess return on stock market asset n on day t in the CRISP financial database.

These problems may contain the usual litany of econometric problems: (1) a non-scalar covariance matrix due to heteroskedasticity across observation units, serial correlation over time, or covariance across equations within an observation unit; and (2) the potential for correlation of explanatory variables and disturbances when x includes lagged dependent variables. They also provide an opportunity for a richer analysis of covariance patterns, since observations across units can be used to identify covariance patterns over time, and observations across time can be used to identify heteroskedasticities across units.

2. STACKING THE DATA

For analysis (and computation), it is useful to organize the observations in vectors in which all the observations for n = 1 are stacked on top of all the observations for n = 2, etc. Use the notation:

$$y_n = \begin{bmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nT} \end{bmatrix}, X_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nT} \end{bmatrix}, u_n = \begin{bmatrix} u_{n1} \\ u_{n2} \\ \vdots \\ u_{nT} \end{bmatrix},$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix} \quad X = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & X_N \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & 0 \\ \vdots & \dots & \vdots \\ x_{1T} & \dots & 0 \\ \vdots & \dots & \vdots \\ 0 & \dots & x_{N1} \\ \vdots & \dots & \vdots \\ 0 & \dots & x_{NT} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Then, the system can be written

$$y_n = x_n\beta_n + u_n \text{ , } n = 1,...,N$$

or in stacked form,

(1) $$y = X\beta + u \text{ .}$$

The vector y_n is of dimension T×1, the array X_n is of dimension T×k, the vector y is of dimension NT×1, the array X is of dimension NT×Nk. We wrote down the system assuming the number of parameters k was the same in each equation, but this is not necessary. One could have X_n of dimension T×k_n and X of dimension NT×(k_1+..+k_n). If there are parameters in common across different equations, then the corresponding explanatory variables will be stacked in the same column rather than placed in different columns, and the overall number of columns in X reduced accordingly.

Suppose the observations are independent and identically distributed for different t, but the covariances $\mathbf{E}(u_{nt}u_{mt}) = \sigma_{nm}$ are not necessarily zero. Let $\Sigma = (\sigma_{nm})$ be the N×N array of covariances of the observations for each t. The covariance matrix of the stacked disturbance vector u is then

$$\mathbf{E}(uu') = \begin{bmatrix} \boldsymbol{\sigma_{11}I_T} & \boldsymbol{\sigma_{12}I_T} & \dots & \boldsymbol{\sigma_{1N}I_T} \\ \boldsymbol{\sigma_{21}I_T} & \boldsymbol{\sigma_{22}I_T} & \dots & \boldsymbol{\sigma_{2N}I_T} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\sigma_{N1}I_T} & \boldsymbol{\sigma_{N2}I_T} & \dots & \boldsymbol{\sigma_{NN}I_T} \end{bmatrix},$$

where $\mathbf{I}_T$ denotes a T×T identity matrix.

Define the Kronecker Product A⊗B of a n×m matrix A and a p×q matrix B:

$$A \otimes B = \begin{bmatrix} \boldsymbol{a_{11}B} & \boldsymbol{a_{12}B} & \dots & \boldsymbol{a_{1m}B} \\ \boldsymbol{a_{21}B} & \boldsymbol{a_{22}B} & \dots & \boldsymbol{a_{2m}B} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{a_{n1}B} & \boldsymbol{a_{n2}B} & \dots & \boldsymbol{a_{nm}B} \end{bmatrix}.$$

Then, A⊗B is (np)×(mq). Kronecker products have the following properties:

$(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$ when the matrices are commensurate
$(A\otimes B)^{-1} = (A^{-1})\otimes(B^{-1})$ when A and B are square and nonsingular
$(A\otimes B)' = (A')\otimes(B')$
$\text{trace}(A\otimes B) = (\text{trace}(A))\cdot(\text{trace}(B))$ when A and B are square
$\det(A\otimes B) = (\det(A))^p(\det(B))^n$ when A is n×n and B is p×p

Applying the Kronecker product notation to the covariance matrix of u, $\mathbf{E}(uu') = \Sigma\otimes\mathbf{I}_T$.

3. ESTIMATION

The problem of estimating the stacked model $y = X\beta + u$ when the covariance matrix of the disturbances is $\Sigma\otimes\mathbf{I}_T$ and Σ is known is a straightforward GLS problem, provided there are no additional complications of correlation of explanatory variables and disturbances. Using the rule for inverses of Kronecker products, the GLS estimator is

$$b = (X'(\Sigma^{-1}\otimes\mathbf{I}_T)X)^{-1}X'(\Sigma^{-1}\otimes\mathbf{I}_T)y\ .$$

Computationally, the most practical way to do this regression is to calculate a triangular Cholesky matrix L such that $L'L = \Sigma^{-1}$. Then, the transformed model

$$(L\otimes\mathbf{I}_T)y = (L\otimes\mathbf{I}_T)X\beta + (L\otimes\mathbf{I}_T)u \tag{2}$$

satisfies Gauss-Markov conditions (Verify), and the BLUE estimator of β is OLS applied to this equation. The data transformations can be carried out separately for each t, and recursively for n = 1,...,N.

When Σ is unknown, one can do FGLS estimation: First apply OLS to (1) and retrieve fitted residuals $\hat{u}$. Then, estimate the elements σ_{nm} of Σ from the average (over T) of the squares and cross-products of the fitted residuals,

$$s_{nm} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{nt}\hat{u}_{mt}\ .$$

Finally, apply OLS to (2), with L a Cholesky factor of the estimated Σ^{-1}.

The problem of estimating β in (1) when there are <u>no</u> cross-equation restrictions on the β_n is called the <u>seemingly unrelated regressions problem</u>. Summarizing, the β_n can be estimated consistently equation-by-equation using OLS; in most cases, this is inefficient compared to GLS; and FGLS is asymptotically fully efficient. There is one case in which there is no efficiency gain from use of GLS rather than OLS: Suppose no cross-equation restrictions on parameters and <u>common</u> explanatory variables across equations; i.e., $X_1 = X_2 = \ldots = X_N$. Then, $X = \mathbf{I}_N\otimes X_1$, and the GLS estimator is

$$b = ((\mathbf{I}_N\otimes X_1')(\Sigma^{-1}\otimes\mathbf{I}_T)(\mathbf{I}_N\otimes X_1))^{-1}(\mathbf{I}_N\otimes X_1')(\Sigma^{-1}\otimes\mathbf{I}_T)y\ .$$

As an exercise, use the Kronecker product rules to show that this formula reduces to the OLS estimator $b_n = (X_1'X_1)^{-1}X_1'y_n$ for each n. Intuitively, the reason OLS is efficient in this case is that the OLS residuals in, say, the first equation are automatically orthogonal to the (common) exogenous variables in each of the other equations, so that there is no additional information on the first equation parameters to be distilled from the cross-equation orthogonality conditions. Put another

way, GLS can be interpreted as OLS applied to linear combinations of the original equations, with the linear combinations obtained from the Cholesky factorization of the covariance matrix of the disturbances. But these linear combinations of the common exogenous variables leaves one with the same exogenous variables, and the orthogonality conditions satisfied by the GLS estimates are the same as the orthogonality conditions satisfied by OLS on the first equation in the original system.

4. AN EXAMPLE

Suppose a firm t utilizes N = 3 inputs, and has a Diewert unit cost function,

$$C_t = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} \sqrt{p_{it} p_{jt}} \quad ,$$

where p_{it} is input i price, and the α's are nonnegative parameters with $\alpha_{ij} = \alpha_{ji}$. By Shephard's lemma, the unit input demand functions are given by the derivatives of the unit cost function with respect to the input prices:

$$z_{nt} = \sum_{j=1}^{N} \alpha_{nj} \sqrt{p_{jt}/p_{nt}} \quad .$$

Written in stacked form, these equations become

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \begin{bmatrix} 1_T & (\sqrt{p_2/p_1})_T & (\sqrt{p_3/p_1})_T & 0_T & 0_T & 0_T \\ 0_T & (\sqrt{p_1/p_2})_T & 0_T & 1_T & (\sqrt{p_3/p_2})_T & 0_T \\ 0_T & 0_T & (\sqrt{p_1/p_3})_T & 0_T & (\sqrt{p_2/p_3})_T & 1_T \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{13} \\ \alpha_{22} \\ \alpha_{23} \\ \alpha_{33} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} ,$$

where 1_T denotes a T×1 vector of 1's and $\left(\sqrt{p_1/p_2}\right)_T$ denotes a T×1 vector with components $\sqrt{p_{1t}/p_{2t}}$. Note that the parameter restrictions across equations lead to variables appearing stacked in the same column. The disturbances can be interpreted as coming from random variations across firms around the respective "average" parameters $\alpha_{11}, \alpha_{22}, \alpha_{33}$. The interesting econometric feature of this setup is that even if there is considerable multicollinearity in prices so that OLS equation by equation is imprecise, this multicollinearity is broken when the data are stacked. Then, there is likely to be a substantial efficiency gain from estimating the equations in stacked form with the cross-equation restrictions imposed, even at the first OLS stage before the additional efficiency gain from the second-stage FGLS is achieved.

5. PANEL DATA

The application of systems of regressions equations to panel data, where n indexes observation units that are followed over time periods t, is very important in economics. A typical model for panel data is

$$y_{nt} = x_{nt}\beta + \alpha_n + u_{nt} \quad \text{for } n = 1,...,N \text{ and } t = 1,...,T .$$

In this model, the β parameters are not subscripted by n or t; this implies they are the same for every unit and every time period. (This is not as restrictive as it might appear, because variation in parameters over time or with some characteristics of the units can be reintroduced by including in the x's interactions with time dummies or with unit dummies.) The α_n are termed individual effects. They may be treated as intercept terms that vary across units. The model with this interpretation is called a fixed effects (FE) model. Alternately, the α_n may be interpreted as components of the disturbance that vary randomly across units. The model with the second interpretation is called a random effects (RE) model. Often, the assumption is made that once the individual effects are isolated, the remaining disturbances u_{nt} are independent and identically distributed across n as well as t. Alternately, the u_{nt} could be serially correlated; this requires another layer of calculation for GLS.

The questions that arise in analysis of the panel data model are (a) under what conditions the model parameters can be estimated consistently, in either the fixed effects or the random effects interpretation; (b) what is the form of consistent or efficient estimators; and (c) whether the random effects or the fixed effects model is "better" in applications. I first analyze the fixed effects case, then the random effects case, and after this return to these questions to see what can be said.

6. FIXED EFFECTS

The fixed effects model can be rewritten by stacking the T observations on unit n,

$$y_n = x_n\beta + 1_T\alpha_n + u_n , \tag{3}$$

where 1_T is a T×1 vector of ones. Equation (3) is a special case of a general system of regression equations, and can be approached in the same way. Stacking the unit data, first unit followed by second unit, etc., gives the stacked model

$$y = X\beta + D\alpha + u , \tag{4}$$

where $D = [d_1\ d_2\ ...\ d_N]$ is a NT×N array whose columns are dummy variables such that d_m is one for observations from unit m, and zero otherwise, and α is a N×1 vector with components α_n. (Exercise: Verify that this setup follows from the general stacking pattern shown in Section 2.)

In (4), note first that any column of X that does not change over t, within the observations for a unit, is linearly dependent on the columns of D. Then, when there are fixed effects, there is no possibility of identifying the separate effects of X variables that are time-invariant. Suppose we remove any such columns from X, so that only time-varying variables are left. For good measure, we can also remove from X the within-unit means of the X variables, so that X now denotes deviations from within-unit means. The model (3) can be rewritten as a relationship in unit means plus relationships in deviations from within unit means:

(5) $\bar{y}_n = \alpha_n + \bar{u}_n$

(6) $Y_n = X_n\beta + \tilde{u}_n,$

where $\bar{y}_n$ and $\bar{u}_n$ are unit means, Y_n is a vector of deviations of the unit n observations from the unit mean, and X_n is an array of deviations that has zero unit means by construction. Stack these models further, with the unit one data followed by the unit two data, etc., to obtain

(7) $$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \beta + \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_N \end{bmatrix} .$$

The deviations in (7) eliminate the fixed effects. Then, (7) can be estimated by OLS, which is consistent for β as $N \to +\infty$ or $T \to +\infty$ or both. (Note that (7) has one redundant observation for each observation unit, since the within group deviations must sum to zero. One can eliminate any one of the observations in each unit, or alternately leave it in the regression and remember that the number of observations is really N(T-1) rather than NT.) The regression (7) is called the <u>within</u> regression. One can estimate the fixed effect for each unit n using the formula $\hat{\alpha}_n = \bar{y}_n$; this is called the <u>between</u> regression. The fixed effects are estimated consistently only if $T \to +\infty$.

The particularly simple formula above for the fixed effects estimates came from normalizing the x's to have zero within-unit means. In the general case where the x's can have non-zero unit means, the fixed effect estimators become $\hat{\alpha}_n = \bar{y}_n - \bar{x}_n b$, where *b* is the vector of estimates from (7).

Exercise 1: Using the projection notation $Q_D = I - D(D'D)^{-1}D'$, note that the OLS estimator of β in (4) is $b = (X'Q_DX)^{-1}X'Q_Dy$. Show that this is the same as the within estimator of β.

7. RANDOM EFFECTS

Suppose the α's in (3) are treated as components of the disturbance, so that (3) can be rewritten as $y = X\beta + \nu$, where $\nu_{nt} = \alpha_n + u_{nt}$. Then, an OLS regression of y on X yields a consistent estimator of β as $NT \to +\infty$, provided the x's and the disturbances are uncorrelated. The covariance matrix of the stacked disturbances is now $\mathbf{E}(\nu\nu') = \mathbf{I}_N \otimes \Omega$, where Ω is the T×T matrix of covariances of the disturbances $\alpha_n + u_{nt}$ for given n, with the form

(8) $$\Omega = \begin{bmatrix} \sigma_\alpha^2+\sigma_u^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2+\sigma_u^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2+\sigma_u^2 \end{bmatrix} \equiv \sigma_\alpha^2 1_N 1_N' + \sigma_u^2 \mathbf{I}_N .$$

Efficiency of estimation can be improved by GLS. Verify as an exercise that $L = \frac{1}{\sigma_u}(\mathbf{I}_N - \lambda 1_N 1_N')$,

with $\lambda = \frac{1}{T}(1 - \sigma_u / \sqrt{\sigma_u^2 + T\sigma_\alpha^2})$, satisfies $L\Omega L' = \mathbf{I}_N$. Then, GLS is the same as OLS applied to the transformed data $(\mathbf{I}_N \otimes L)y = (\mathbf{I}_N \otimes L)X\beta + (\mathbf{I}_N \otimes L)\nu$. In practice, Ω is unknown and FGLS must be used. Intuition for how to estimate σ_α^2 and σ_u^2 can be obtained from an analogy to population moments. Let ν_n^* denote the unit mean of ν_{nt}. We know that $\mathbf{E}\nu_{nt}^2 = \sigma_u^2 + \sigma_\alpha^2$ and that $\mathbf{E}\nu_n^{*2} = \sigma_u^2/T + \sigma_\alpha^2$. Solve these two equations for σ_u^2 and σ_α^2:

(9) $$\sigma_u^2 = \frac{T}{T-1}(\mathbf{E}\nu_{nt}^2 - \mathbf{E}\nu_n^{*2}) \text{ and } \sigma_\alpha^2 = (T\,\mathbf{E}\nu_n^{*2} - E\nu_{nt}^2)/(T-1).$$

Then, substituting sample moments of fitted OLS disturbances in place of the population moments will give consistent estimates of the variance components. The steps to do FGLS are then to first regress y on X and retrieve the fitted residuals v_{nt}, and second, estimate $\mathbf{E}v_{nt}^2$ and $\mathbf{E}v_n^{*2}$ by the respective formulas

$$\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} v_{nt}^2 \quad \text{and} \quad \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} v_{nt}\right)^2$$

Third, substitute these expressions in (9) to estimate the variance components and substitute the results into the L matrix, carry out the data transformations unit by unit, and run OLS on the transformed stacked data to get the FGLS estimates. The variance component estimates above are the same as in Greene except for degrees of freedom adjustments. (Since only consistency of the estimates of σ_α^2 and σ_u^2 matter for the efficiency of the FGLS estimator, unbiasedness is no particular virtue. Finite sample monte carlo results on the value of degrees of freedom adjustments are not compelling. Thus, in most cases, it is probably not worth making these adjustments.) The estimator of σ_α^2 can go negative in finite samples. The usual recommendation in this case is to set the estimator to zero and assume there are no individual effects. Show as a (difficult) exercise that if the α's and u's are normal and uncorrelated with each other, then the estimators above are the maximum likelihood estimators for the variances.

Suppose that instead of starting from the original stacked data, we had started from the *within* regression model

(10) $$Y = X\beta + \nu^*,$$

which contains the stacked deviations from unit means, and constitutes N(T-1) observations if redundant observations are excluded; and the *between* regression model

(11) $$\bar{y} = \bar{x}\beta + \bar{\nu},$$

which contains the N stacked unit means. Provided the coefficients are identified (e.g., each variable is time-varying so that no columns of X are identically zero), one could estimate β consistently by applying OLS to either (10) or (11) separately. Greene shows that the OLS estimator can be

interpreted as a weighted combination of the within and between OLS estimators, and that the GLS estimator can be interpreted as a different weighted combination that gives less weight to the between model. For comparison, the fixed effects estimator of β was given by the within regression only.

8. FIXED EFFECTS VERSUS RANDOM EFFECTS

In the (unusual) case that you need estimates of the individual effects, you have no choice but to estimate the fixed effects model; even then, you need $T \to +\infty$ to estimate the α's consistently. The fixed effects model has the advantage that the estimates of β are consistent even if X is correlated with the individual effects, provided of course that X and the individual effects are uncorrelated with u. Its major drawbacks are that it uses up quite a few degrees of freedom, and makes it impossible to identify the effects of time-invariant explanatory variables. The random effects model economizes on degrees of freedom, and permits consistent estimation of the effects of all explanatory variables, including ones that are time-invariant, provided that all these explanatory variables are uncorrelated with the disturbances. (This is an advantage only if you have a convincing story to support the identifying assumption that there is zero correlation of these variables and the α's.)

As $T \to +\infty$, the FE and RE estimators merge, and the FE estimator can be interpreted as estimation of the RE model by conditioning on the realized values of the α's. From this, one can see how to test the RE model specification by examining the correlation of α and X. One way to do this is to regress the fitted α on X, and carry out a conventional F test that the coefficients in this regression are all zero. Unless T is very large, or the assumption that α is uncorrelated with X particularly implausible, it is usually better to work with the RE model.

9. SPECIFICATION TESTING

Standard regression model hypothesis testing of linear hypotheses on model coefficients, using Wald, LR, or SSR test statistics, carries over to the case of systems of regressions. This is most transparent when the FGLS estimators are given by OLS applied to data that is transformed to give a (asymptotically) scalar covariance matrix. This setup allows one to test not only hypotheses about coefficients in one equation, but also hypotheses connecting coefficients across equations, or in the panel context, across time.

For tests on covariance parameters, such as a test for homoskedasticity across equations, or a test for serial correlation, two useful ways to get suitable test statistics are to proceed by analogy with single-indexed regression problems, and to derive LM statistics under the assumption that disturbances are normal. One example is a Durbin-Watson like test for serial correlation in panel data, using the estimated coefficient from a regression of v_{nt} on $v_{n,t-1}$ for n = 1,...N and t = 2,...,T.

Exercise 2: Consider the panel data model in which $T \to +\infty$. If the disturbances are uncorrelated with the right-hand-side variables, then both the FE and RE model estimates will be consistent and the RE estimates will be efficient. On the other hand, if there is correlation between the disturbances and the right-hand-side variables, only the FE estimates will be consistent. From these observations, suggest a simple specification test for the hypothesis that the disturbances are uncorrelated with the right-hand-side variables. Use (10) and (11) to show that this test is equivalent to a test for over-identifying restrictions.

Exercise 3: One of the ways a panel data model might come about is from a regression model $y_{nt} = x_{nt}\gamma_{nt} + u_{nt}$, where the γnt are random coefficients that vary with n (or t). When does this model reduce to the standard panel data model with random n effects? What are the generalizations of the standard RE and FE estimators when $\gamma_{nt} = \delta + \kappa_n + \lambda_t$?

10. VECTOR AUTOREGRESSION

The generic systems of equations model (1) with n indexing dependent variables and t indexing time, and with the right-hand-side variables various lags of the dependent variables, is called a *vector autoregression* (VAR) model. The model may include current and lagged exogenous variables, but is often applied to macroeconomic data where all the variables in the analysis are treated as dependent variables. To write out the lag structure, form the date-t vectors

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Nt} \end{bmatrix}, \; X_t = \begin{bmatrix} x_{1t} & 0 & \cdots & 0 \\ 0 & x_{2t} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x_{Nt} \end{bmatrix}, \; u_t = \begin{bmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{Nt} \end{bmatrix},$$

and then

$$y_t = X_t\beta + A_1 y_{t-1} + \ldots + A_J y_{t-J} + u_t, \tag{12}$$

where the A_j are N×N arrays of lag coefficients. The VAR assumption is that with inclusion of sufficient lags, the disturbances in (12) are i.i.d. innovations that are statistically independent of $X_t, y_{t-1}, y_{t-2}, \ldots$. In this case, the variables $X_t, y_{t-1}, y_{t-2}, \ldots$ are said to be *strongly predetermined* in (12). The X_t are often assumed, further, to be *strongly exogenous*; i.e., u_t is statistically independent of X_t and all leads and lags of X_t.

The dynamics of the system (12) are most easily analyzed by defining

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-J+1} \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{J-1} & A_J \\ I_J & 0_J & \cdots & 0_J & 0_J \\ \vdots & \vdots & & \vdots & \vdots \\ 0_J & 0_J & \cdots & I_J & 0_J \end{bmatrix},$$

and rewriting the system in the form

$$\mathbf{y}_t = \begin{bmatrix} X_t\beta \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathbf{A}\mathbf{y}_{t-1} + \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The system (12) with the strongly exogenous forcing variables X_t and the disturbances u_y omitted, is a *stable* difference equation if all the characteristic roots of **A** are less than one in modulus. The long-run dynamics of a stable system will be dominated by the largest (in modulus) characteristic

root of **A**, and will have the feature that the impact on y_t of a shock in the disturbance in a specified period eventually damps out. Further, the most slowly decaying component in each variable in y_t will damp out at the same rate. (There is an exception if the characteristic vector associated with the largest characteristic root lies in a subspace spanned by a subset of the variables.) In the stable case, i.i.d. innovations, combined with strongly exogenous variables that have a stationary distribution, will produce y_t with a stationary distribution. In particular, the covariance matrix of y_t will not vary with t, so that the y_t are homoskedastic. The estimation and hypothesis testing procedures discussed in Section 3 will then apply, with the predetermined and strongly exogenous variables treated the same. There will in general be contemporaneous correlation, so that (12) has the structure of a seemingly unrelated regressions problem for which GLS can be used to obtain BLUE estimates of the coefficients. If the strictly exogenous variables are the same in every equation, there are no exclusion restrictions in the lag coefficients, and no restrictions on coefficients across equations, GLS estimation reduces to OLS applied to each equation separately, as before.

If **A** has one or more roots of modulus one or greater, then the impact of past disturbances does not damp out, the system (12) is unstable, and the variance of y_t rises with t. The occurrence of modulus one (unit) roots seems to be fairly common in macroeonomic time series. Statistical inference in such systems is quite different than in stable systems. In particular, detection and testing for unit roots, and the corresponding characteristic roots that determine cointegrating relationships among the variables, require a special statistical analysis. The topic of testing for unit roots and cointegrating relationships is discussed extensively by Stock "Unit Roots, Structural Breaks, and Trends," and Watson "Vector Autoregression and Cointegration," both in R. Engle and D. McFadden, eds., Handbook of Econometrics IV, 1994.

11. SYSTEMS OF NONLINEAR EQUATIONS

The systems of equations linear in variables and parameters, with additive disturbances, that were introduced at the beginning of this chapter, can be extended easily to systems that retain the assumption of additive disturbances, but are nonlinear in variables and/or parameters:

$$y_{nt} = h_n(x_{nt},\beta_n) + u_{nt}, \tag{13}$$

where n = 1,...,N, t = 1,...,T, and β_n is $k_n \times 1$. Assume for the following discussion that the disturbances u_{nt} are independent for different t. If the x_{nt} are strongly predetermined, implying that $\mathbf{E}(u_{nt} | x_{nt}) = 0$, then each equation in (13) can be estimated by nonlinear least squares. This can be interpreted as a "limited information" or "marginal" GMM estimation procedure in which information from the equations for the remaining variables is not used. Chapter 3 discusses the statistical properties of nonlinear least squares estimators.

In general, there will be an efficiency gain from taking into account the covariance structure of the disturbances u_{nt} for different n. This can be done practically in TSP by using the LSQ command applied to all the equations in the model. This procedure then applies nonlinear least squares to each equation separately, retrieves fitted residuals, uses these residuals to estimate the covariance matrix of the disturbances at each t, and then does feasible generalized nonlinear least squares employing the estimated covariance matrix.

CHAPTER 6. SIMULTANEOUS EQUATIONS

1. INTRODUCTION

Economic systems are usually described in terms of the *behavior* of various economic agents, and the *equilibrium* that results when these behaviors are reconciled. For example, the operation of the market for Ph.D. economists might be described in terms of *demand behavior*, *supply behavior*, and *equilibrium* levels of employment and wages. The market clearing process feeds back wages into the behavioral equations for demand and supply, creating *simultaneous* or *joint* determination of the equilibrium quantities. This causes econometric problems of correlation between explanatory variables and disturbances in estimation of behavioral equations.

Example 1. In the market for Ph.D. economists, let q = number employed, w = wage rate, s = college enrollment, and m = the median income of lawyers. Assume that all these variables are in logs. The behavioral, or structural, equation for *demand* in year t is

$$q_t = \beta_{11} + \beta_{12} s_t + \beta_{13} w_t + \varepsilon_{1t} ; \tag{1}$$

this equation states that the demand for economists is determined by college enrollments and by the wage rate for economists. The behavioral equation for *supply* is

$$q_t = \beta_{21} + \beta_{22} m_t + \beta_{23} w_t + \beta_{24} q_{t-1} + \varepsilon_{2t} ; \tag{2}$$

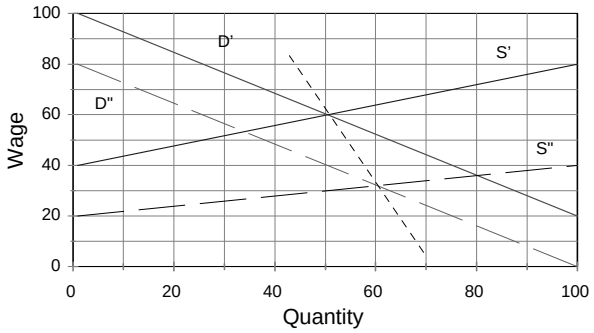
this equation states that the supply of economists is determined by the wage rate, the income of lawyers, which represents the opportunity cost for students entering graduate school, and lagged quantity supplied, which reflects the fact that the pool of available economists is a stock that adjusts slowly to market innovations. Equations (1) and (2) together define a *structural simultaneous equations system*. The disturbances ε_{1t} and ε_{2t} reflect the impact of various unmeasured factors on demand and supply. For this example, assume that they are uncorrelated over time. Assume that college enrollments s_t and lawyer salaries m_t are *exogenous*; meaning that they are determined outside this system, or functionally, that they are uncorrelated with the disturbances ε_{1t} and ε_{2t}. Then, (1) and (2) are a *complete* system for the determination of the two *endogenous* or dependent variables q_t and w_t.

Suppose you are interested in the parameters of the demand equation, and have data on the variables appearing in (1) and (2). How could you obtain good statistical estimates of the demand equation parameters? It is useful to think in terms of the "experiment" run by Nature, and the experiment that you would ideally like to carry out to form the estimates.

Figure 1 shows the demand and supply curves corresponding to (1) and (2), with w and q determined by market equilibrium. Two years are shown, with solid curves in the first year and dashed curves in the second. The equilibrium wage and quantity are of course determined by the condition that the market clear. If both the demand and supply curves shift between periods due to random disturbances, then the locus of equilibria will be a scatter of points (in this case, two) which will not in general lie along either the demand curve or the supply curve. In the case illustrated, the dotted line which passes through the two observed equilibria has a slope substantially different than

the demand curve. If the disturbances mostly shift the demand curve and leave the supply curve unchanged, then the equilibria will tend to map out the supply curve. Only if the disturbances mostly shift the supply curve and leave the demand curve unchanged will the equilibria tend to map out the demand curve. These observations have several consequences. First, an ordinary least squares fit of equation (1) will produce a line like the dotted line in the figure that is a poor estimate of the demand curve. Only when most of the shifts over time are coming in the supply curve so that the equilibria lie along the demand curve will least squares give satisfactory results. Second, exogenous variables shift the demand and supply curve in ways that can be estimated. In particular, the variable m that appears in the supply curve but not the demand curve shifts the supply curve, so that the locus of w,q pairs swept out when only m changes lies along the demand curve. Then, the ideal experiment you would like to run in order to estimate the slope of the demand curve is to vary m, holding all other things constant. Put another way, you need to find a statistical analysis that mimics the ideal experiment by isolating the partial impact of the variable m on both q and w.

Fig. 1. Demand & Supply of Economists



The structural system (1) and (2) can be solved for q_t and w_t as functions of the remaining variables

$$(3) \qquad w_t = \frac{\left\{(\beta_{11}-\beta_{21}) + \beta_{12}s_t - \beta_{22}m_t - \beta_{24}q_{t-1} + (\varepsilon_{1t}-\varepsilon_{2t})\right\}}{\beta_{23} - \beta_{13}}$$

$$(4) \qquad q_t = \frac{\left\{(\beta_{11}\beta_{23}-\beta_{21}\beta_{13}) + \beta_{23}\beta_{12}s_t - \beta_{13}\beta_{22}m_t - \beta_{13}\beta_{24}q_{t-1} + (\beta_{23}\varepsilon_{1t}-\beta_{13}\varepsilon_{2t})\right\}}{\beta_{23} - \beta_{13}}$$

Equations (3) and (4) are called the *reduced form*. For this solution to exist, we need $\beta_{23} - \beta_{13}$ non-zero. This will certainly be the case when the elasticity of supply β_{23} is positive and the elasticity of demand β_{13} is negative. Hereafter, assume that the true $\beta_{23} - \beta_{13} > 0$. Equations (3) and (4) constitute a system of regression equations, which could be rewritten in the stacked form

$$(5)\qquad \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \\ q_1 \\ q_2 \\ \vdots \\ q_T \end{bmatrix} = \begin{bmatrix} 1 & s_1 & m_1 & q_0 & 0 & 0 & 0 & 0 \\ 1 & s_2 & m_2 & q_1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_T & m_T & q_{T-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & s_1 & m_1 & q_0 \\ 0 & 0 & 0 & 0 & 1 & s_2 & m_2 & q_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & s_T & m_T & q_{T-1} \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{14} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \\ \pi_{24} \end{bmatrix} + \begin{bmatrix} \nu_{11} \\ \nu_{12} \\ \vdots \\ \nu_{1T} \\ \nu_{21} \\ \nu_{22} \\ \vdots \\ \nu_{2T} \end{bmatrix}, \text{ or } y = Z\pi + \nu ,$$

where the π's are the combinations of behavioral coefficients, and the ν's are the combinations of disturbances, that appear in (3) and (4). The system (5) can be estimated by GLS. In general, the disturbances in (5) are correlated and heteroskedastic across the two equations. However, exactly the same explanatory variables appear in each of the two equations. If the correlation pattern is the same in each equation, so that $\mathbf{E}\nu_{it}\nu_{js} = \sigma_{ij}\rho_{ts}$, or $\mathbf{E}\nu\nu' = R\otimes\Sigma$, then GLS using this covariance structure collapses to GLS applied separately to each equation. When there is no correlation across t, GLS collapses to OLS.

Suppose you are interested in estimating the parameters of the behavioral demand equation (1). For OLS applied to (1) to be consistent, it is necessary that the disturbance ε_{1t} be uncorrelated with the right-hand-side variables, which are s_t and w_t. This condition is met for s_t, provided it is indeed exogenous. However, from (3), an increase in ε_{1t} increases w_t, other things being equal, and in (1) this results in a positive correlation of the RHS variable w_t and the disturbance ε_{1t}.

Instrumental variables estimation is one alternative for the estimation of (1). In this case, one needs to introduce at least as many instrumental variables as there are RHS variables in (1), and these variables need to be uncorrelated with ε_{1t} and fully correlated with the RHS variables. The list of instruments should include the exogenous variables in (1), which are the constant, 1, and s_t. Other candidate instruments are the exogenous and predetermined variables elsewhere in the system, m_t and q_{t-1}.

Will IV work? In general, to have enough instruments, there must be at least as many predetermined variables excluded from (1) and appearing elsewhere in the system as there are endogenous variables on the RHS of (1). When this is true, (1) is said to satisfy the *order condition* for identification. In the example, there is one RHS endogenous variable, w_t, and two excluded exogenous and predetermined variables, m_t and q_{t-1}, so the order condition is satisfied. If there are enough instruments, then from the general theory of IV estimation, the most efficient IV estimator is obtained by first projecting the RHS variables on the space spanned by the instruments, and then using these projections as instruments. In other words, the best combinations of instruments are obtained by regressing each RHS variable in (1) on the instruments 1, s_t, m_t, and q_{t-1}, and then using the fitted values from these regressions as instruments. But the reduced form equation (3) is exactly this regression. Therefore, the best IV estimator is obtained by first estimating the reduced form equations (3) and (4) by OLS and retrieving fitted values, and then estimating (1) by OLS after replacing RHS endogenous variables by their fitted values from the reduced form. For this to yield instruments that are fully correlated with the RHS variables, it must be true that at least one of the

variables m_t and q_{t-1} truly enters the reduced form, which will happen if at least one of the coefficients β_{22} or β_{24} is nonzero. This is called the *rank* condition for identification.

2. STRUCTURAL AND REDUCED FORMS

In general a behavioral or structural simultaneous equations system can be written

$$y_t'B + z_t'\Gamma = \varepsilon_t', \tag{6}$$

where $y_t' = (y_{1t}, .,y_{Nt})$ is a 1×N vector of the endogenous variables, B is a N×N array of coefficients, $z_t' = (z_{n1}, .,z_{Mt})$ is a 1×M vector of predetermined variables, Γ is a M×N array of coefficients, and ε_t' is a 1×N vector of disturbances. Let Σ denote the N×N covariance matrix of ε_t. The reduced form for this system is

$$y_t' = z_t'\Pi + v_t', \tag{7}$$

where $\Pi = -\Gamma B^{-1}$ and $v_t' = \varepsilon_t'B^{-1}$, so that the covariance matrix of v_t is $\Omega = B'^{-1}\Sigma B^{-1}$. Obviously, for (6) to be a well-defined system that determines y_t, it is necessary that B be non-singular.

3. IDENTIFICATION

It should be clear that some restrictions must be imposed on the coefficient arrays B and Γ, and possibly on the covariance matrix Σ, if the remaining coefficients are to be estimated consistently. First, post-multiplying (6) by a nonsingular diagonal matrix leaves the reduced form solution (7) unchanged, so that all versions of (6) that are rescaled in this way are observationally equivalent. Then, for estimation of (6) it is necessary to have a scaling normalization for each equation. Second, counting parameters, B, Γ, and Σ contain N(N-1) + NM + N(N+1)/2 parameters, excluding the N parameters determined by the scaling normalizations and taking into account the symmetry of Σ. However, Π and Ω contain only NM + N(N+1)/2 parameters. Therefore, an additional N(N-1) restrictions on parameters are necessary to determine the remaining structural parameters from the reduced form parameters.

It is traditional in econometrics texts to work out detailed *order* and *rank* conditions for identification. These come from the structure of the B and Γ matrices and the condition that $\Pi B + \Gamma = 0$ relating the reduced form coefficients to the structural parameters. However, it is much simpler to think of identification in terms of the possibility for IV estimation: *An equation (with associated restrictions) is identified if and only if there exists a consistent IV estimator for the parameters in the equation; i.e., if there are sufficient instruments for the RHS endogenous variables that are fully correlated with these variables.* Even covariance matrix restrictions can be used in constructing instruments. For example, if you know that the disturbance in an equation you are trying to estimate is uncorrelated with the disturbance in another equation, then you can use a consistently estimated residual from the second equation as an instrument. If you are not embarrassed to let a computer do your thinking, you can even leave identification to be checked numerically: an equation is identified if and only if you can find an IV estimator for the equation that empirically has finite variances.

Exercise 1. Show that the condition above requiring N(N-1) restrictions on parameters will hold if the *order condition*, introduced in the example of the market for economists, holds for each equation. In the general case, the order condition for an equation states that the number of excluded

predetermined (including strictly exogenous) variables is at least as great as the number of included RHS endogenous variables. Add the number of excluded RHS endogenous variables to each side of this inequality, and sum over equations to get the result.

4. 2SLS

For discussions of estimators for simultaneous equations systems, it is convenient to have available the systems (6) and (7) stacked two different ways. First, one can stack (6) and (7) vertically by observation to get

(8) $$YB + Z\Gamma = \varepsilon$$

and

(9) $$Y = Z\Pi + \nu,$$

where Y, ε, and ν are T×N and Z is T×K. With this stacking, one has $\mathbf{E}\varepsilon'\varepsilon/T = \Sigma$ and $\mathbf{E}\nu'\nu/T = B^{-1}\Sigma B'^{-1}$. Note that post-multiplying (8) by a non-singular diagonal matrix leaves the reduced form unchanged; hence this modification is observationally equivalent. Then, we can choose any convenient diagonal matrix as a normalization. In particular, we can renumber the equations and rescale them so that the dependent variable y_{nt} appears with a coefficient of one in the n-th equation. This is equivalent to saying that we can write B = **I** - A, where A is a matrix with zeros down the diagonal, and that the behavioral system (8) can be written

(10) $$Y = YA - Z\Gamma + \varepsilon \equiv [Y \mid Z] \begin{bmatrix} A \\ -\Gamma \end{bmatrix} \equiv XC + \varepsilon.$$

In this setup, Y and ε are T×N, X is T×(N+K), and C is (N+K)×N. Restrictions that exclude some variables from some equations will force some of the parameters in C to be zero. Rewrite the n-th equation from (10), taking these restrictions into account, as

(11) $$y_n = Y_n A_n - Z_n \Gamma_n + \varepsilon_n \equiv X_n C_n + \varepsilon_n,$$

where this equation includes M_n endogenous variables and K_n predetermined variables on the RHS. Then, y_n is T×1, Y_n is T×M_n, and Z_n is T×K_n, and X_n is T×(M_n+K_n).

A second method of stacking which is more convenient for empirical work is to write down all the observations for the first equation, followed by all the observations for the second equation, etc. This amounts to starting from (11), and stacking the T observations for the first equation, followed by the T observations for the second equation, etc. Since the C_n differ across equations, the stacked system looks like

(12) $$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & X_N \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \equiv \mathbf{Xc} + \mathbf{e}.$$

Note that **X** in (12) is <u>not</u> the same as X in (10); **X** is NT×J, where $J = J_1 + . + J_N$ and $J_n = M_n + K_n$ is the number of RHS variables in the n-th equation. The system (12) has the appearance of a system

of regression equations. Because of RHS endogenous variables, OLS will not be consistent, so that we have to turn to IV methods. In addition, there are GLS issues due to the correlation of disturbances across equations.

Suppose you are interested in estimating a single equation from the system, say $y_1 = Y_1A_1 - Z_1\Gamma_1 + \varepsilon_1 \equiv X_1c_1 + \varepsilon_1$. The IV method states that if you can find instruments W that are uncorrelated with ε_1 and fully correlated with X_1, then the best IV estimator, $\hat{c}_1 = [X_1'W(W'W)^{-1}W'X_1]^{-1}X_1'W(W'W)^{-1}W'y_1$ is consistent. But the potential instruments for this problem are $Z = [Z_1 \mid Z_{-1}]$, where Z_{-1} denotes the predetermined variables that are in Z, but not in Z_1. The *order* condition for identification of this equation is that the number of variables in Z_{-1} be at least as large as the number of variables in Y_1, or *the number of <u>excluded</u> predetermined must be as large as the number of <u>included</u> RHS endogenous*. The *rank* condition is that X_1' W be of maximum rank. For consistency, you need to have $X_1'W/T$ converging in probability to a matrix of maximum rank.

Exercise 2. Show that the rank condition implies the order condition. Show in the example of the supply and demand for economists that the order condition can be satisfied, but the rank condition can fail, so that the order condition is necessary but not sufficient for the rank condition.

The best IV estimator can be written $\hat{c}_1 = [X_{1e}'X_{1e}]^{-1}X_{1e}'y_1$, where $X_{1e} = W(W'W)^{-1}W'X_1$ is the array of fitted values from an OLS regression of X_1 on the instruments W = Z; i.e., the reduced form regression. Then, the estimator has a *two-stage OLS* (2SLS) interpretation:

(1) Estimate the reduced form by OLS, and retrieve the fitted values of the endogenous variables.

(2) Replace endogenous variables in a behavioral equation by their fitted values from the reduced form, and apply OLS.

Recall from the general IV method that the procedure above done by conventional OLS programs will not produce consistent standard errors. Correct standard errors can be obtained by first calculating residuals from the 2SLS estimators in the original behavioral model, $u_1 = y_1 - X_1\hat{c}_{2SLS}$, estimating $\hat{\sigma}^2 = u_1'u_1/(T-K_1)$, and then estimating $V_e(\hat{c}_{2SLS}) = \hat{\sigma}^2[X_1'X_1]^{-1}$.

5. 3SLS

The 2SLS method does not exploit the correlation of the disturbances across equations. You saw in the case of systems of regression equations that using FGLS to account for such correlations improved efficiency. This will also be true here. To motivate an estimator, write out all the moment conditions available for estimation of each equation of the system:

$$(13) \qquad \begin{bmatrix} Z'y_1 \\ Z'y_2 \\ \vdots \\ Z'y_N \end{bmatrix} = \begin{bmatrix} Z'X_1 & 0 & \dots & 0 \\ 0 & Z'X_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & Z'X_N \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} + \begin{bmatrix} Z'\varepsilon_1 \\ Z'\varepsilon_2 \\ \vdots \\ Z'\varepsilon_N \end{bmatrix} \equiv [(\mathbf{I}_N \otimes Z')\mathbf{X}]\mathbf{c} + (\mathbf{I}_N \otimes Z')\varepsilon.$$

The disturbances in the NK×1 system (13) have the covariance matrix $\Sigma \otimes (Z'Z)$. Then, by analogy to GLS, the best estimator for the parameters should be

(14) $$\hat{c}_{3SLS} = \left\{\boldsymbol{X}'(\boldsymbol{I}_N\otimes Z)(\Sigma^{-1}\otimes(Z'Z)^{-1})(\boldsymbol{I}_N\otimes Z')\boldsymbol{X}\right\}^{-} \mathbf{X}'(\mathbf{I}_N\otimes Z)(\Sigma^{-1}\otimes(Z'Z)^{-1})(\mathbf{I}_N\otimes Z')y$$

$$= \left\{\boldsymbol{X}'(\Sigma^{-1}\otimes(Z(Z'Z)^{-1}Z'))\boldsymbol{X}\right\}^{-1} \mathbf{X}'(\Sigma^{-1}\otimes(Z(Z'Z)^{-1})Z'))y\ .$$

This estimator can be obtained in three OLS stages, hence its name:

(1-2) Do 2SLS on each equation of the system, and retrieve the residuals calculated at the 2SLS estimators and the original (not the fitted) RHS variables.

(3) Estimate Σ from the residuals just calculated, and then do FGLS regression of y on X using the GLS weighting matrix $\Sigma^{-1}\otimes(Z(Z'Z)^{-1})Z')$.

The large-sample approximation to the covariance matrix for (14) is, from the usual GLS theory,

(15) $$V(\hat{c}_{3SLS}) = \left\{\boldsymbol{X}'(\Sigma^{-1}\otimes(Z(Z'Z)^{-1}Z'))\boldsymbol{X}\right\}^{-1}\ .$$

The FGLS third stage for the 3SLS estimator can be done conveniently by a OLS on transformed data. Let L be a lower triangular Cholesky factor of Σ_e^{-1} and Q be a lower triangular Cholesky factor of $(Z(Z'Z)^{-1})Z'$. Then $(L\otimes Q)(L\otimes Q)' = \Sigma_e^{-1}\otimes(Z(Z'Z)^{-1})Z')$. Transform $(L\otimes Q)y = (L\otimes Q)\mathbf{Xc} + \boldsymbol{\eta}$ and apply OLS to this system to get the 3SLS estimators.

The main advantage of 3SLS over 2SLS is a gain in asymptotic efficiency. The main disadvantage is that the estimators for a single equation are potentially less robust, since they will be inconsistent if the IV assumptions that Z is predetermined fail in any equation, not just a particular one of interest.

6. TESTING FOR OVER-IDENTIFYING RESTRICTIONS

Consider an equation $y = X\beta + u$ from a system of simultaneous equations, and let W denote the array of instruments (exogenous and predetermined variables) in the system. Let $X^* = P_W X$ denote the fitted values of X obtained from OLS estimation of the reduced form; where $P_W = W(W'W)^{-}W'$ is the projection operator onto the space spanned by W. The equation is *over-identified* if the number of instruments W exceeds the number of right-hand-side variables X. From Chapter 3, the GMM test statistic for over-identification is the minimum in β of

$$2nQ_n(\beta) = u'P_W\, u/\sigma^2 = u'P_{X^*}\, u/\sigma^2 + u'(P_W - P_{X^*})u/\sigma^2,$$

where $u = y - X\beta$. One has $u'(P_W - P_{X^*})u = y'(P_W - P_{X^*})y$, and at the minimum in β, $u'P_{X^*}u = 0$, so that $2nQ_n = y'(P_W - P_{X^*})y/\sigma^2$. Under H_o, this statistic is asymptotically chi-squared distributed with degrees of freedom equal to the difference in ranks of W and X^*. This statistic can be interpreted as the difference in the sum of squared residuals from the 2SLS regression of y on X and the sum of squared residuals from the reduced form regression of y on W, normalized by σ^2. A computationally convenient equivalent form is $2nQ_n = \|\hat{y}_W - \hat{y}_{X^*}\|^2/\sigma^2$, the sum of squares of the difference between the reduced form fitted values and the 2SLS fitted values of y, normalized by σ^2. Finally, $2nQ_n = y'Q_{X^*}P_WQ_{X^*}y/\sigma^2 = nR^2/\sigma^2$, where R^2 is the multiple correlation coefficient from regressing the 2SLS residuals on all the instruments; this result follows from the equivalent formulas for the projection onto the subspace of **W** orthogonal to the subspace spanned by X^*. This test statistic does *not* have a version that can be written as a quadratic form with the wings containing a difference of coefficient estimates from the 2SLS and reduced form regressions. Note that if the equation is *just identified*, with the number of proper instruments excluded from the equation exactly

equal to the number of right-hand-side included endogenous variables, then there are no over-identifying restrictions and the test has no power. However, when the number of proper instruments exceeds the minimum for just identification, this test amounts to a test that all the exclusions of the instruments from the structural equation are valid.

7. TIME-SERIES APPLICATIONS OF SIMULTANEOUS EQUATIONS MODELS

The example of the market for economists that introduced this chapter was a time- series model that involved lagged dependent variables. In the example, we assumed away serial correlation, but in general serial correlation will be as issue to be dealt with in applications of simultaneous equations models to time series. The setup (6) for a linear simultaneous equations model can be expanded to make dependence on lagged dependent variables explicit:

$$y_t'B + y_{t-1}'\Lambda + z_t'\Gamma = \varepsilon_t'. \tag{16}$$

Recall that the variables y_{t-1} and z_t in this model are *predetermined* if they are uncorrelated with the disturbance ε_t, and *strongly predetermined* if ε_t is statistically independent of y_{t-1} and z_t. In this model, the strictly exogenous variables z_t may include lags (and, if it makes economic sense, leads). It is not restrictive to write the model as a first-order lag in y_t, as higher-order lags can be incorporated by including lagged values of the dependent variables as additional components of y_t, with identities added to the system of equations to link the variables at different lags. (This was done in Chapter 5 in discussing the stability of vector autoregressions.)

The reduced form for the system (16), also called the *final form* in time series applications, is

$$y_t' = y_{t-1}'\Theta + z_t'\Pi + \nu_t', \tag{17}$$

where $\Theta = -\Lambda B^{-1}$, $\Pi = -\Gamma B^{-1}$, and $\nu_t' = \varepsilon_t' B^{-1}$, so that the covariance matrix of ν_t is $\Omega = B'^{-1}\Sigma B^{-1}$. Identification of the model requires that B be nonsingular, and that there be exclusion and/or covariance restrictions that satisfy a rank condition. Stability of the model requires that the characteristic roots of Θ all be less than one in modulus. If one started with a stable structural model that had disturbances that were serially correlated with an autoregressive structure, then with suitable partial differencing the model could be rewritten in the form (17), the disturbances ν_t would be innovations that are independent across t, and the explanatory variables in (17) would be strongly predetermined. Further, the dynamics of the system would be dominated by the largest modulus characteristic root of Θ. In this stable case, estimation of the model can proceed in the manner already discussed: Estimate the reduced form, use fitted values of y_t (along with z_t and y_{t-1}) as instruments to obtain 2SLS estimates of each equation in (17), and finally use fitted covariances from these equations (calculated at the 2SLS estimates) to carry out 3SLS.

If the final form (17) is not stable, and in particular Λ has one or more unit roots, then the statistical properties of 2SLS or 3SLS estimates are quite different: some estimates may converge in asymptotic distribution at rate T rather than the customary $T^{1/2}$, and the asymptotic distribution may not be normal. Consequently, one must be careful in conducting statistical inference using these estimates. There is an extensive literature on analysis of systems containing unit roots; see the chapter by Jim Stock in the Handbook of Econometrics IV. When a system is known to contain a unit root, then it may be possible to transform to a stable system by appropriate differencing.

8. NONLINEAR SIMULTANEOUS EQUATIONS MODELS

In principle, dependent variables may be simultaneously determined within a system of equations that is nonlinear in variables and parameters. One might, for example, consider a system

$$F_i(y_{1t},y_{2t},...,y_{Nt};z_{it},\theta) = \varepsilon_{it},\ i = 1,...,N \tag{18}$$

for the determination of $(y_{1t},y_{2t},...,y_{Nt})$ that depends on a K×1 vector of parameters θ, vectors of exogenous variables z_{it}, and disturbances ε_{it}. Such systems might arise naturally out of economic theory. For example, consumer or firm optimization may be characterized by first-order conditions that are functions of dependent decision variables and exogenous variables describing the economic environment of choice, with the ε_{it} appearing due to errors in optimization by the economic agents, arising perhaps because ex post realizations differ from ex ante expectations, or due to approximation errors by the analyst. For many plausible economic models, linearity of the system (18) in variables and parameters would be the exception rather than the rule, with the common linear specification justifiable only as an approximation. The nonlinear system (18) is well-determined if it has a unique solution for the dependent variables, for every possible configuration of the z's and ε's, and for all θ's in a specified domain. If it is well-determined, then it has a reduced form

$$y_{it} = f_i(z_{1t},z_{2t},...,z_{Nt},\varepsilon_{1t},\varepsilon_{2t},...,\varepsilon_{Nt},\theta),\ i = 1,...,N. \tag{19}$$

This reduced form can also be written

$$y_{it} = h_i(z_{1t},z_{2t},...,z_{Nt},\theta) + u_{it},\ i = 1,...,N \tag{20}$$

where

$$h_i(z_{1t},z_{2t},...,z_{Nt},\theta) = \mathbf{E}\{f_i(z_{1t},z_{2t},...,z_{Nt},\varepsilon_{1t},\varepsilon_{2t},...,\varepsilon_{Nt},\theta)\,|\,z_t\},$$

and u_{it} is the disturbance with conditional mean zero that makes (20) hold. In this form, (20) is a system of nonlinear equations in the form considered in Chapter 5, and the treatment there can also be applied to estimate the structural parameters from this reduced form. (The specification (20) guarantees that the reduced form disturbances have conditional expectation zero; but the additional assumption that u's are statistically independent of z's, or even that they are homoskedastic, is rarely justifiable from economic theory. Then statistical analysis based on this assumption may be invalid and misleading for many application.)

Recall that in Chapter 4, estimation of a nonlinear equation with contaminated explanatory variables was discussed, a best nonlinear 2SLS (BN2SLS) estimator was defined, and practical approximations to the BN2SLS were discussed. The equations in (18) would correspond directly to this structure if in equation i, one had

$$F_i(y_{1t},y_{2t},...,y_{Nt};z_{it},\theta) = y_{it} - h(y_{1t},...,y_{i-1,t},y_{i+1,t},...,y_{Nt},z_{it},\theta), \tag{21}$$

Absent this normalization, some other normalization is needed for identification in F_i, either on the scale of the dependence of F_i on one variable, or in the scale of ε_{it}. This is no different in spirit than the normalizations needed in a linear simultaneous equations specification. Given an identifying

normalization, it is possible to proceed in essentially the same way as in Chapter 4. Make a first-order Taylor's expansion of (18) about an initial parameter vector θ_o to obtain

$$(22) \qquad F_i(y_{1t},y_{2t},...,y_{Nt};z_{it},\theta_o) \approx - \sum_{k=1}^{K} \frac{\partial F_i(y_{1t},y_{2t},...,y_{Nt};z_{it},\theta_o)}{\partial \theta_k} \cdot(\theta_k-\theta_{ok}) + \varepsilon_{it}.$$

Treat the expressions $x_{itk} = -\partial F_i(y_{1t},y_{2t},...,y_{Nt};z_{it},\theta_o)/\partial\theta_k$ as contaminated explanatory variables, and the expectations of x_{ikt} given $z_{1t},...,z_{Nt}$ as the ideal best instruments. Approximate these best instruments by regressing the x_{itk} on suitable functions of the z's, as in Chapter 4, and then estimate (22) by this approximation to best 2SLS. Starting from an initial guess for the parameters, iterate this process to convergence, using the estimated coefficients from (22) to update the parameter estimates. The left-hand-side of (22) is the dependent variable in these 2SLS regressions, with the imposed normalization guaranteeing that the system is identified. This procedure can be carried out for the entire system (22) at one time, rather than equation by equation. This will provide nonlinear 2SLS estimates of all the parameters of the system. These will not in general be best system estimates because they do not take into account the covariances of the ε's across equations. Then, a final step is to apply 3SLS to (22), using the previous 2SLS estimates to obtain the feasible GLS transformation. The procedure just described is what the LSQ command in TSP does when applied to a system of nonlinear equations without normalization, with instrumental variables specified.

When the nonlinear reduced form (20) can be obtained as an analytic or computable model, it is possible to apply nonlinear least squares methods directly, either equation by equation as N2SLS or for the system as N3SLS. This estimation procedure is described in Chapter 5. One caution is that while the disturbances u_{it} in (20) have conditional mean zero by construction, economic theory will rarely imply that they are, in addition, homoskedastic, and the large sample statistical theory needs to be reworked when heteroskedasticity of unknown form is present. Just as in linear models, consistency is generally not at issue, but standard errors will typically not be estimated consistently. At minimum, one should be cautious and use robust standard error estimates that are consistent under heteroskedasticity of unknown form.

CHAPTER 7. ROBUST METHODS IN ECONOMETRICS

1. THE PARAMETERS OF ECONOMETRICS

Econometrics deals with complex multivariate relationships and employs non-experimental or "field" data that are influenced by many factors. Occasionally econometricians have data from *designed experiments* in which treatments are randomized, and/or other factors are held constant, to assure that there can be no confounding of the measured effects of treatments. Almost as good are "*natural experiments*", also called "*quasi-experiments*", in field data where a factor of direct interest (or an instrument correlated with a factor of interest) has clearly operated in a manner that is independent of confounding effects. The scientific value of such quasi-experiments is high, and econometricians should actively seek designed or natural experiments that can illuminate economic issues. That said, there remain important problems in economic theory and policy for which experimental data are not available within the time frame in which answers are needed. It is imperative that econometricians deal with these problems using the best tools available, rather than reverting to an orthodoxy that they are too "messy" for econometric treatment.

Econometricians must make educated guesses about the structure of the data generation processes in non-experimental data. The studies that result rely on these structural assumptions can be misleading if the assumptions are not realistic. This has important implications for the conduct of econometric analysis. First, it is desirable to have large data sets in which the "signal" contained in systematic relationships is strong relative to the "statistical noise". Second, it is important to "proof" econometric models, testing the plausibility of the specification both internally and against other data and other studies, and avoiding complex or highly parametric formulations whose plausibility is difficult to check. Fourth, it is desirable to use statistical methods that are "robust" in the sense that they do not force conclusions that are inconsistent with the data, or rely too heavily on small parts of the data.

Most of classical econometric analysis, from linear regression models to maximum likelihood estimation of non-linear models, lays out the assumptions under which the procedures will produce good statistical results, and simply assumes that these postulates can be checked and will be checked by users. To some extent, the development of diagnostic and specification tests provides the capacity to make these checks, and good econometric studies use these tests. However, some basic assumptions are difficult to check, and they are too often accepted in econometric studies without serious examination. Fortunately, in many economic applications, particularly using linear models, the analysis is more robust than the assumptions, and sensibly interpreted will provide useful results even if some assumptions fail. Further, there are often relatively simple estimation alternatives that provide some protection against failures, such as use of instrumental variables or heteroskedasticity-consistent standard errors. New developments in econometrics expand the menu of procedures that provide protection against failures of classical assumptions. This chapter introduces three areas in which "robust" methods are available: the use of nonparametric and semiparametric methods, the use of simulation methods and "indirect inference", and the use of bootstrap methods.

Econometrics first developed from classical parametric statistics, with attention focused on linear systems. This was the only practical alternative in an era when computation was difficult and data limited. Linear parametric models remain the most useful tool of the applied econometrician.

However, the assumption of known parametric functional forms and distributions interposes an untidy veil between econometric analysis and the propositions of economic theory, which are mostly abstract without specific dimensional or functional restrictions. Buoyed by good data and computers, contemporary econometricians have begun to attack problems which are not *a priori* parametric. One major line of attack is to use general nonparametric estimation methods to avoid distributional assumptions. The second, closer to classical methods, is to use flexible forms to approximate unknown functions, and specification tests to search for parsimonious representations. The added dimension in a modern rendition of the second approach is explicit recognition of the statistical consequences of adding terms and parameters as sample sizes grow.

Many problems of econometric inference can be cast into some version of the following setup: There is a random vector $(Y,X) \in \mathbb{R}^k \times \mathbb{R}^m$ such that X has a (unknown) density $g(x)$ and almost surely Y has a (unknown) conditional density $f(y|x)$. There is a known transformation $t(y,x)$ from $\mathbb{R}^k \times \mathbb{R}^m$ into the real line $\mathbb{R}$, and the conditional expectation of this transformation, $\theta(x) = \mathbf{E}(t(Y,x)|X=x)$, is the target of the econometric investigation. Examples of transformations of interest are (1) $t(y,x) \equiv y$, in which case $\theta(x) = \mathbf{E}(Y|X=x)$ is the conditional expectation of Y given x, or the *regression function* of Y on x; (2) $t(y,x) = yy'$, in which case $\theta(x) = \mathbf{E}(YY'|X=x)$ is the array of second conditional moments, and this function combined with the first example, $\mathbf{E}(YY'|X=x) - \{\mathbf{E}(Y|X=x)\}\{\mathbf{E}(Y|X=x)\}'$ is the conditional variance; and (3) $t(y,x) = \mathbf{1}_A(y)$, the indicator function of the set A, in which case $\theta(x)$ is the conditional probability of the event A, given X = x. Examples of economic applications are Y a vector of consumer demands, and x the vector of income and prices; or Y a vector of firm net outputs and x a vector of levels of fixed inputs and prices of variable inputs.

Define the disturbance $\varepsilon = \varepsilon(y,x) \equiv t(y,x) - \theta(x)$. Then the setup above can be summarized as a *generalized regression model*,

$$t(y,x) = \theta(x) + \varepsilon,$$

where $\mathbf{E}(\varepsilon|x) = 0$. Econometric problems fitting this setup can be classified as *fully parametric*; *semiparametric*; or *nonparametric*. The model is fully parametric if the function θ and the distribution of the disturbance ε are both known *a priori* to be in finite-parameter families. The model is nonparametric if both θ and ε have unknown functional forms, except possibly for shape and regularity properties such as concavity or continuous differentiability. The model is semiparametric if it contains a finite parameter vector, typically of primary interest, but parts of θ and/or the distribution of ε are not restricted to finite-parameter families. This is a rather broad definition of semiparametric, which includes for example linear regression under Gauss-Markov conditions where the distribution of the disturbances is not restricted to a parametric family, and only the first two moments are parametric. Some econometricians prefer to reserve the term semiparametric for situations where the problem can be characterized as one with a finite-dimensional parameter vector that is the target of the analysis and an infinite-dimensional vector of nuisance parameters (which might, for example, determine an unknown function), for it is in this case that non-classical statistical methods are needed.

Where can an econometrician go wrong in setting out to analyze the generalized regression relationship $t(y,x) = \theta(x) + \varepsilon$? First, there is nothing in the formulation of this model *per se* that assures that $\theta(x)$ has any causal or invariance properties that allow it to be used to predict the distribution of values of $t(y,x)$ if the distribution of x shifts. Put another way, the model will by definition be descriptive of the conditional mean in the current population, but not necessarily

predictive under policy changes that alter the distribution of x. Because econometricians are often interested in conditional relationships for purposes of prediction or analysis of policy scenarios, this is potentially a severe limitation. The prescription for "robust" causal inference is to use statistical methods and tests that can avoid or detect joint or "wrong-way" causality (e.g., instrumental variables, Granger invariance tests in time series, exogeneity tests); avoid claiming causal inferences where confounding of effects is possible; and avoid predictions that require substantial extrapolation from the data. Second, when $\theta(x)$ is approximated by a parametric family, there will be a specification error if the parametric family fails to contain $\theta(x)$. Specification errors are particularly likely if the parametric family leaves out variables or variable interactions that appear in the true conditional expectation. Third, the only property that is guaranteed for the disturbances ε when $\theta(x)$ is correctly specified is the conditional first moment condition $\mathbf{E}(\varepsilon \mid x) = 0$. There is no guarantee that the conditional distribution of ε given x is independent of x, or for that matter that the variance of ε is homoskedastic. In addition, there is no guarantee that the distribution of ε has thin enough tails so that higher moments exist, or are sufficiently well behaved so that estimates are not unduly (and unstably) influenced by a small number of high influence observations. In these circumstances, statistical methods that assume well-behaved disturbances can be misleading, and better results may be obtained using methods that bound the influence of tail information. At minimum, it is often worth providing estimates of estimator dispersion that are consistent in the presence of various likely problems with the disturbances.

In statistics, there is a fairly clear division between *nonparametric statistics*, which worries about the specification of $\theta(x)$ or about tests of the qualitative relationship between x and *t*, and *robust statistics*, which worries about the properties of ε. In econometrics, both problems appear, usually together, and it is useful to refer to the treatment of both problems in economic applications as *robust econometrics*.

Despite the leading place of fully parametric models in classical statistics, elementary nonparametric and semiparametric methods are used widely without fanfare. Histograms are nonparametric estimators of densities. Contingency tables for data grouped into cells are one approach to estimating a regression function nonparametrically. Linear regression models, or any estimators that rely on a finite list of moment conditions, can be interpreted as semiparametric, since they do not require complete specification of the underlying distribution function.

2. HOW TO CONSTRUCT A HISTOGRAM

One of the simplest examples of a nonparametric problem is that of estimating an unknown univariate unconditional density $g(x)$, given a random sample of observations x_i for i = 1,...,n. Assume, by transformation if necessary, that the support of g is the unit interval. An elementary method of approximating g is to form a histogram: First partition the unit interval into K segments of length 1/K, so that segment k is $(c_{k-1},c_k]$ with $c_k = k/K$ for k = 0,...,K. Then estimate g within a segment by the share of the observations falling in this segment, divided by segment length. If you take relatively few segments, then the observation counts in each segment are large, and the variance of the sample share in a segment will be relatively small. On the other hand, if the underlying density is not constant in the segment, then this segment average is a biased estimate of the density at a point. This bias is larger when the segment is longer. Segment length can be varied to balance variance against bias. As sample size rises, the number of segments can be increased so that the contributions of variance and bias remain balanced.

Suppose the density g has the following smoothness property:

$$|g(x') - g(x)| \leq L|x' - x|,$$

where L is a positive constant. Then the function is said to satisfy a *Lipschitz condition*. If g is continuously differentiable, then this property will be satisfied. Let n_k be the number of observations from the sample that fall in segment k. Then, the histogram estimator of g at a specified argument x is

$$\hat{g}(x) = Kn_k/n \text{ for } x \in (c_{k-1},c_k].$$

Compute the variance and bias of this estimator. First, the probability that an observation falls in segment k is the segment mean of g, $p_k = K\cdot \int_{c_{k-1}}^{c_k} g(x)dx$. Then, n_k has a binomial distribution with probability p_k/K, so that it has mean np_k/K and variance $n(p_k/K)(1 - p_k/K)$. Therefore, for $x_o \in (c_{k-1},c_k]$, $\hat{g}(x_o)$ has mean p_k and variance $(K/n)p_k(1 - p_k/K)$. The bias is $B_{nK}(x) = p_k - g(x)$. The *mean square error* of the estimator equals its variance plus the square of its bias, or

$$MSE(x) = (K/n)p_k(1 - p_k/K) + (p_k - g(x))^2.$$

A criterion for choosing K is to minimize the mean square error. Looking more closely at the bias, note that by the theorem of the mean, there is some argument z_k in the segment $(c_{k-1},c_k]$ such that $p_k/K = \int_{c_{k-1}}^{c_k} g(x)dx = g(z_k) \int_{c_{k-1}}^{c_k} dx = g(z_k)/K$. Then, using the Lipschitz property of g,

$$|p_k - g(x)| = |g(z_k) - g(x)| \leq L|z_k - x| \leq L/K,$$

Then, the MSE is bounded by

$$MSE(x) \leq (K/n)p_k(1 - p_k/K) + L^2/K^2.$$

Approximate the term $p_k(1 - p_k/K)$ in this expression by g(x), and then minimize the RHS in K. The (approximate) minimand is $K = (2L^2n/g(x))^{1/3}$, and the value of MSE at this minimand is approximately $(Lg(x)/2n)^{2/3}$. Of course, to actually do this calculation, you have a belling-the-cat problem that you need to know $g(x)$. However, there are some important qualitative features of the solution. First, the optimal K goes up in proportion to the cube root of sample size, and MSE declines proportionately to $n^{-2/3}$. Compare this with the formula for the variance of parametric estimators such as regression slope coefficients, which are proportional to 1/n. Then, the histogram estimator is *consistent* for g, since the mean square error goes to zero. However, the cost of not being able to confine g to a parametric family is that the rate of convergence is lower than in parametric cases. Note that when L is smaller, so that g is less variable with x, K is smaller.

If you are interested in estimating the entire function g, rather than the value of g at a specified point x, then you might take as a criterion the Mean Integrated Square Error (MISE),

$$\text{MISE} = \mathbf{E}\int(\hat{g}(x) - g(x))^2dx = \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} \mathbf{E}(\hat{g}(x) - p_k + p_k - g(x))^2dx$$

$$= \sum_{k=1}^{K} \mathbf{E}(Kn_k/n - p_k)^2/K + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} (p_k - g(x))^2dx$$

$$= \sum_{k=1}^{K} (1/n)p_k(1 - p_k/K) + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} (g(z_k) - g(x))^2dx$$

$$\le K/n + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} L^2\cdot(z_k - x)^2dx \le K/n + L^2/3K^2.$$

The RHS of this expression is minimized at $K = (2L^2n/3)^{1/3}$, with $\text{MISE} \le (3L/2n)^{2/3}$. Both minimizing MSE at a specified x and minimizing MISE imply that the number of histogram cells K grows at the rate $n^{1/3}$. When $g(x) < 3$, the optimal K for the MISE criterion will be smaller than the optimal K for the MSE criterion; this happens because the MISE criterion is concerned with average bias and the MSE criterion is concerned with bias at a point. One practical way to circumvent the belling-the-cat problem is to work out the value of K for a standard distribution; this will often give satisfactory results for a wide range of actual distributions. For example, the triangular density $g(x) = 2x$ on $0 \le x \le 1$ has L = 2 and gives $K = 2(n/3)^{1/3}$. Thus, a sample of size n = 81 implies K = 6, while a sample of size n = 3000 gives K = 20.

3. KERNEL ESTIMATION OF A MULTIVARIATE DENSITY

One drawback of the histogram estimator is that it is estimating a continuous density by a step function, and the constancy of this estimate within a cell and the steps between cells contribute to bias. There would seem to be an advantage to using an estimator that mimics the smoothness that you know (believe?) is in the true density. This section describes the commonly used *kernel method* for estimating a multivariate density.

Suppose one is interested in estimating an unknown density $g(x)$ for $x = (x_1,...,x_m)$ in the domain $[0,1]^m$. Suppose that g is not known to be in a parametric family, but is known to be strictly positive on the interior of $[0,1]^m$ and is known to have the following smoothness property: g is continuously differentiable up to order p (where $p \ge 0$), and the order p derivatives satisfy a Lipschitz condition. Some notation is needed to make this precise. Let $r = (r_1,...,r_m)$ denote a vector of non-negative integers, and $|r| = \sum r_j$. Let $g^r(x) = \partial^{|r|} / \partial x_1^{r_1} \cdots \partial x_m^{r_m}$ denote the mixed partial derivative of g of order $|r|$ with respect to the arguments in r. The assumption is that $g^r(x)$ exists and is continuous for all r satisfying $|r| \le p$, and that there exists a constant L such that $|g^r(x) - g^r(y)| \le L|x - y|$ for any r satisfying $|r| = p$. In applications, the most common cases considered are p = 0, where one is assuming g continuous and not too variable (e.g., Lipschitz), and p = 2, where one is assuming g twice continuously differentiable. Define $\mathbf{z}^r \equiv z_1^{r_1}...z_m^{r_m}$. A function g that satisfies the smoothness condition above has a Taylor's expansion (in h) that satisfies

$$g(x - hz) = \sum_{q=0}^{p} \frac{(-h)^q}{q!} \sum_{|r|=q} g^r(x)\cdot z^r + \lambda\cdot \frac{h^{p+1}}{p!} \sum_{|r|=p} |g^r(x)\cdot z^r|\cdot L|z|$$

for some scalar $\lambda \in (-1,1)$.

Exercise 1. Verify that for m = 1, these smoothness conditions reduce to the requirement that *g* be p-times continuously differentiable, with $d^p g(x)/dx^p$ satisfying a Lipschitz condition, so the Taylor's expansion is a textbook expansion in derivatives up to order p.

Exercise 2. Show that in the case p = 0, the expansion reduces to $g(x - hz) = g(x) + \lambda h\cdot L|z|$.

Suppose you have a random sample x_i for i = 1,...,n drawn from the density *g*(x). In applications, it is almost always desirable to first do a linear transformation of the data so that the components of x are orthogonal in the sample, with variances that are the same for each component. Hereafter, assume that the x's you are working with have this property. Suppose that you estimate *g* using a kernel estimator,

$$\hat{g}(\mathbf{x}) = \frac{1}{nh^m} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) .$$

The function K(**z**) is the *kernel*, and the scalar h is the *bandwidth*. The kernel K is a function on $(-\infty,+\infty)^m$ with the properties that $\int K(z)dz = 1$, and for some integer s with $0 \le s \le p$, $\int z^r\cdot K(z)dz = 0$ for $|r| \le s$ and $\int z^r\cdot K(z)dz = k_r$ for $|r| = s+1$, where the k_r are constants that are finite and not all zero. In words, K is a "density-like" function which integrates to one, but which is not necessarily always non-negative. All the moments of this function up through order s vanish, and moments of order s + 1 exist and some do not vanish. This is called a *kernel of order* s. In applications, you will encounter mostly first-order kernels satisfying $\int z_i K(z)dz = 0$ and $\int z_i^2 K(z)dz > 0$; these are usually constructed as non-negative densities that are symmetric about zero. Higher-order kernels, for s > 1, will be used to take advantage of problems where *g* is known to be differentiable to higher order than two. Higher order kernels will necessarily sometimes be negative.

An example of a first-order kernel is $K(z) = (2\pi)^{-m/2}\cdot\exp[-z'z/2]$, a Gaussian kernel formed by the product of univariate standard normal densities. Forming products of univariate kernels in this fashion is a convenient way to build up multivariate kernels. Another example of a multivariate kernel is the multivariate Epanechnikov kernel, $K(z) = (½)c_m\cdot(m+2)\cdot(1 - z'z)\cdot\mathbf{1}(z'z < 1)$, where c_m is the volume of a unit sphere in $\mathbb{R}^m$, which can be calculated recursively using the formulas $c_1 = 2$, $c_2 = \pi$, and $c_n = c_{n-2}\cdot n/(n-1)$ for n > 2. An example of a second-order kernel derived from a first-order kernel K is

$$K^*(\mathbf{z}) = [K(\mathbf{z}) - \gamma^3 K(\gamma\mathbf{z})]/(1-\gamma^2),$$

where γ is a scalar in (0,1). (If K is symmetric about zero, then K* is actually a third-order kernel.) Kernels to any order can be built up recursively as linear combinations of lower order kernels.

Mean and Variance of the Kernel Estimator

The mean of the kernel estimator is

$$\mathbf{E}\hat{g}(\mathbf{x}) = \frac{1}{nh^m}\sum_{i=1}^{n} \mathbf{E}\, K\left(\frac{x - x_i}{h}\right) = \frac{1}{h^m}\int K\left(\frac{x - y}{h}\right)\cdot g(y)dy .$$

Using the fact that the observation x_i are independent, the variance of the kernel estimator is

$$\mathbf{V}\hat{g}(x) \equiv \mathbf{E}[\hat{g}(x) - \mathbf{E}\hat{g}(x)]^2 = \frac{1}{n^2h^{2m}}\sum_{i=1}^{n} \left\{ \boldsymbol{E}\, K\left(\frac{x - x_i}{h}\right)^2 - \left[\boldsymbol{E}\, K\left(\frac{x - x_i}{h}\right)\right]^2\right\}$$

$$= \frac{1}{nh^{2m}}\left\{\int K\left(\frac{x - y}{h}\right)^2 g(y)dy - \left[\int K\left(\frac{x - y}{h}\right)\cdot g(y)dy\right]^2\right\} .$$

Consistency, Bias, and Mean Square Error

Require $h \to 0$ and $n\cdot h^{2m} \to +\infty$. Then, $\mathbf{E}\hat{g}(x) \to g(x)$ and $\mathbf{V}\hat{g}(x) \to 0$, so that $\hat{g}(x)$ converges to $g(x)$ in mean square error, and is hence consistent. Note that for m large, these conditions require that h fall quite slowly as n rises. This is called the *curse of dimensionality*.

Next approximate the bias and variance of the estimator when h is small. Assume that the order of the kernel s is less than or equal to the degree of differentiability p. Introduce the change of variables $y = x - hz$ in the expressions for the mean and variance of $\hat{g}(x)$, and then use the Taylor's expansion for $g(x - hz)$ up to order s, to obtain

$$\mathbf{E}\hat{g}(x) = \frac{1}{h^m}\int K\left(\frac{x - y}{h}\right)\cdot g(y)dy = \int K(z)\cdot g(x - hz)dz$$

$$= g(x) + \sum_{q=0}^{p} \frac{(-h)^q}{q!}\, g^r(x) \int K(z)\cdot z^r dz + \lambda\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} g^r(x)\cdot \int K(z)\cdot z^r\cdot L|z|dz$$

$$= g(x) + \lambda'\cdot L\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |g^r(x)|\cdot C_r ,$$

where $C_r = \int |K(z)\cdot z^r|\cdot|z|dz$ is a positive constant determined by the kernel, and λ' is a scalar in (-1,1). Then,

$$\text{Bias}(x) = \lambda'\cdot L\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |g^r(x)|\cdot C_r .$$

From this formula, one sees that the magnitude of the bias shrinks at the rate h^{s+1}, where s is the order of the kernel, as long as $s \le p$. Thus, when one knows that *g* has a high degree of differentiability, one can use a higher order kernel and control bias more tightly. The reason this works is that when *g* is very smooth, you can in effect estimate and remove bias components that change smoothly with x; e.g., bias terms that are linear in deviations from the target x. However, if one uses a low order kernel, the bias is determined by the order of the kernel, and is not reduced even if the function *g* is very smooth. At the other extreme, the bias is of order h^{p+1} for any kernel of order $s \ge p$, since the Taylor's series cannot be extended beyond the order of differentiability of *g*, so nothing is gained on the bias side by going to a kernel of order $s > p$. For example, if $p = 0$, so that one knows only that *g* is Lipschitz, then one cannot reduce the order of bias by using a symmetric kernel.

Next consider the variance. Making the change of variables y = x - hz,

$$\mathbf{V}\hat{g}(x) = \mathbf{E}[\hat{g}(x) - \mathbf{E}\hat{g}(x)]^2 = \frac{1}{nh^m}\int K(z)^2 \cdot g(x - hz)dz - \frac{1}{n}\left(\int K(z)\cdot g(x - hz)dz\right)^2$$

$$= \frac{g(x)}{nh^m}\int K(z)^2 dz + \frac{D}{n\cdot h^{m-1}},$$

where D is a constant that depends on K and g. As $h \to 0$, the first term in the variance will dominate. Then, the mean square error of the estimator $\hat{g}$ at x is bounded by

$$MSE(x) = Bias(x)^2 + \mathbf{V}\hat{g}(x) = L^2 \cdot \frac{h^{2(s+1)}}{(s!)^2}\left(\sum_{|r|=s} |g^r(x)| \cdot C_r\right)^2 + \frac{g(x)}{nh^m}\int K(z)^2 dz + HOT,$$

where HOT stands for "Higher Order Terms". The *mean integrated square error* (MISE) is then

$$MISE = \int MSE(x)dx = L^2 \cdot \frac{h^{2(s+1)}}{(s!)^2} \cdot A + \frac{1}{nh^m}\int K(z)^2 dz + HOT,$$

where

$$A = \int \left(\sum_{|r|=s} |g^r(x)| \cdot C_r\right)^2 dx .$$

The optimal bandwidth h minimizes MISE:

$$h_{opt} = \left(\frac{m(s!)^2}{2(s+1)n\cdot A\cdot L^2}\int K(z)^2 dz\right)^{\frac{1}{m+2(s+1)}} .$$

Then, the bandwidth falls with n, at a slower rate the higher the dimension m or the higher the order of the kernel s. Intuitively, this is because when m is high, there are more dimensions where data can "hide", so the sample is less dense and one has to look more widely to find sufficient neighboring points. Also, when the order of the kernel s high, more distant points can be used without adding too much to bias because the function is smooth enough so that leading bias terms can be taken out. Increasing the order of derivatives typically increases A and/or L, and this also shrinks bandwidth. In an applied problem, direct application of the formula for h_{opt} is impractical because it depends on functions of g that one does not know.

Substituting the optimal bandwidth in MISE yields

$$MISE(h_{opt}) = n^{\frac{2(s+1)}{m+2(s+1)}} \cdot \left\{\frac{2(s+1)AL^2}{m(s!)^2}\right\}^{\frac{m}{m+2(s+1)}} \cdot \left\{\int K(z)^2 dz\right\}^{\frac{2(s+1)}{m+2(s+1)}} \cdot \frac{m+2(s+1)}{2(s+1)} .$$

Note first that MISE will always fall more slowly than 1/n. This is due to the nonparametric nature of the problem, which implies in effect that only local data is available to estimate the density at each point. Chuck Stone has shown that the rate above is not particular to kernel estimation, but is a best rate that can be obtained by any estimation method. Second, the higher the dimension m, the lower

the rate at which MISE falls with sample size, the *curse of dimensionality*. If the problem is very smooth, and one exploits this by using a higher-order kernel, one can offset some of the curse of dimensionality. In the limiting case, as s –→ +∞, the rate approaches the limiting 1/n rate. However, other terms in MISE also change when one goes to higher order kernels. In particular, $\int K(z)^2 d\mathbf{z}$ will increase for higher order kernels, and the constant A will typically increase rapidly because higher order derivatives are less smooth than lower order ones.

Least-Squares Cross-Validation

The idea behind cross-validation is to formulate a version of the MISE criterion that can be estimated from the data alone. Then, the bandwidth that minimizes this empirical criterion is close to the optimal bandwidth. The MISE criterion can be written

$$\text{MISE} = \mathbf{E} \int [\hat{g}(x) - g(x)]^2 dx = \mathbf{E} \int \hat{g}(x)^2 dx - 2\cdot\mathbf{E} \int \hat{g}(x)\cdot g(x) dx + \int g(x)^2 dx .$$

The approach is to obtain unbiased estimators of the terms involving $\hat{g}(x)$, and then to choose h iteratively to minimize this estimated criterion. Consider first the term $\mathbf{E} \int \hat{g}(x)^2 dx$. This expression can be estimated using the kernel estimator $\hat{g}$. To get a convenient computational formula, first define $K^{(2)}(z) = \int K(w - z)\cdot K(w) dw$. This is a convolution that defines a new kernel starting from K, and is an expression that can often be determined analytically. When K is a probability density, $K^{(2)}$ has a simple interpretation: if $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent random vectors with density K, then the density of $\mathbf{Z} = \mathbf{W}_1 - \mathbf{W}_2$ is $K^{(2)}$. For example, if K is a product of univariate standard normal densities, then $K^{(2)}$ is a product of univariate normal densities with mean 0 and variance 2. Using the definition of $K^{(2)}$, and making the transformation of variables $w = (x - x_i)/h$,

$$\int \hat{g}(x)^2 dx = \frac{1}{n^2 h^{2m}} \sum_{i=1}^{n} \int K\left(\frac{x - x_i}{h}\right) \cdot K\left(\frac{x - x_j}{h}\right) \cdot dx$$

$$= \frac{1}{nh^m} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(2)}\left(\frac{x_j - x_i}{h}\right) .$$

This statistic converges to its expectation as n –→ +∞.

Next consider the term $\int \hat{g}(x)\cdot g(x) dx = \frac{1}{n^2 h^{2m}} \sum_{i=1}^{n} \int K\left(\frac{x - x_i}{h}\right) g(x) dx$. Replace the unknown $g(x)$ in the expression $\int K\left(\frac{x - x_i}{h}\right) g(x) dx$ by the empirical density from the sample, excluding $\mathbf{x}_i$; this puts probability 1/(n-1) at each data point x_j for $j \neq i$. This gives an estimator $\frac{1}{nh^m} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right)$ for $\int \hat{g}(x)\cdot g(x) dx$.

Exercise 3. Show that $\int \hat{g}(x)\cdot g(x)dx$ and the estimator for it given above have the same expectation.

Putting together the estimators for the first two terms in the MISE, one obtains the empirical criterion

$$\text{MISE}'(h) = \frac{1}{n^2 h^m}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(2)}\left(\frac{x_j - x_i}{h}\right) - \frac{2}{nh^m}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} K\left(\frac{x_j - x_i}{h}\right)$$

$$= \frac{1}{n^2 h^m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K^{(2)}\left(\frac{x_j - x_i}{h}\right) - \frac{2n}{n-1}K\left(\frac{x_j - x_i}{h}\right)\right] + \frac{2K(0)}{(n-1)h^m}.$$

For application, use a nonlinear search algorithm to minimize this expression in h. The minimand h_{lsxv} is the optimal bandwidth estimated by the cross-validation method. An important theoretical result due to Chuck Stone is that if *g* is bounded, then $\text{MISE}(h_{opt})/\text{MISE}(h_{lsxv}) \to 1$ as $n \to +\infty$, so that asymptotically one can do as well using the bandwidth obtained by minimizing the empirical criterion MISE′(h) as one can do using the optimal bandwidth.

4. NONPARAMETRIC REGRESSION

Now consider the general problem of estimating $\theta(x)$ in the regression model $t_i = \theta(x_i) + \varepsilon_i$, where x_i is of dimension m, $t_i = t(y_i,x_i)$ is a known transformation, θ is an unknown function, ε_i is a disturbance satisfying $\mathbf{E}(\varepsilon_i | x_i) = 0$, but otherwise not restricted, and (y_i,x_i) for i = 1,...,n is a random sample. This is the general setup from the introduction. Consider *locally weighted* estimators of the form

$$T_n(x) = \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)t(y_i,x_i),$$

where the w_{ni} are scalars that put the most weight on observations with x_i near x. The weights do not have to be non-negative, but their sum has to approach one as $n \to +\infty$. Here are some examples of nonparametric estimation methods that are of this form, and their associated weight functions:

1. *Kernel Estimation*: Suppose K is a *kernel* function from $\mathbb{R}^m$ into $\mathbb{R}$, and h is a *bandwidth*. The function K will be large near zero, and will go to zero at arguments far away from zero; common examples for m = 1 are the *uniform* kernel, $K(v) = \mathbf{1}_{[-1,+1]}(v)$; the *normal* kernel $K(v) = \varphi(v)$, where φ is the standard normal density; the *triangular* kernel $K(v) = \text{Max}\{1-|v|,0\}$; and the *Epanechnikov* kernel $K(v) = (3/4)(1-v^2)\mathbf{1}_{[-1,+1]}(v)$, which turns out to have an efficiency property. The local weights are

$$w_{ni}(x;x_1,...,x_n) = \frac{1}{h_n^{\,m}} K\left(\frac{x-x_i}{h_n}\right) \Big/ \sum_{j=1}^{n} \frac{1}{h_n^{\,m}} K\left(\frac{x-x_j}{h_n}\right) ,$$

where the bandwidth h_n shrinks with sample size. The kernel estimator of $\theta(x)$ is

$$T_n(x) = \frac{\frac{1}{nh^{\,m}} \sum_{i=1}^{n} t(y_i,x_i) K\left(\frac{x-x_i}{h_n}\right)}{\frac{1}{nh^{\,m}} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)} .$$

The denominator of this expression can be interpreted as an estimator of $g(x)$, and the numerator as an estimator of $g(x)\mathbf{E}_{y|x}t(y,x) = g(x)\theta(x)$. The kernel function K is typically defined so that $\int K(v)dv = 1$, and is taken to be symmetric so that $\int vK(v)dv = 0$. If θ is known to be a smooth function, with Lipschitz derivatives of order p, then there turns out to be an advantage (in large enough samples) to using a *higher-order* kernel that satisfies $\int v^j k(v)dv = 0$ for $j = 1,...,p$.

2. *Nearest Neighbor Estimator*. For the given x, order the observations $(y_{(i)},x_{(i)})$ so that $|x - x_{(1)}| \le |x - x_{(2)}| \le ... \le |x - x_{(n)}|$. To simplify discussion, rule out ties. Define a sequence of scalars $w_{n,(i)}$ that sum to one, and define

$$T_n(x) = \sum_{i=1}^{n} w_{n,(i)} t(y_{(i)},x_{(i)}) .$$

If $w_{n,(i)} = 0$ for $i > r$, this is termed a *r-nearest neighbor* estimator. Examples of weights are uniform, $w_{n,(i)} = 1/r$ for $i \le r$ and zero otherwise, and triangular, $w_{n,(i)} = 2(r-i+1)/r(r+1)$. If θ is known to be a smooth function with Lipschitz derivatives of order p, then it is advantageous to run a *local regression*, in which $t(y_i,x_i)$ is regressed on all points of the form $\prod_{h=1}^{m} x_{ih}^{p_h}$ with $\sum_{h=1}^{m} p_h \le p$, with weights $w_{n,(i)}$, and the fitted value of this regression at x is the estimator of $\theta(x)$. This extension reduces bias by taking into account the fact that a smooth function must vary regularly in its arguments, allowing larger neighborhoods so that variance as well as bias can be reduced.

Uniform nearest neighbor and uniform kernel estimators have the following relationship: If the bandwidth in a uniform kernel estimator is chosen as a function of the data, a *variable kernel* method, so that exactly r observations fall in the interval where the kernel is positive, then this estimator is a uniform nearest neighbor estimator.

3. *Other Nonparametric Methods*. There are several widely used nonparametric estimation methods other than locally weighted estimators. First, the function $\theta(x)$ may be approximated by sums of standard functions, such as polynomials, with the number of terms in the sums growing with sample

size. A traditional form of these series approximations is the use of *Fourier* or *Laplace* approximations, or other series of *orthogonal polynomials*. These series are truncated at some point, depending on the sample size, the dimension of the problem, and the smoothness assumed on $\theta(x)$. Once this is done, the problem is effectively parametric, and ordinary regression methods can be used. (Judicious choice of the series so that the terms are orthogonal results in computational simplifications, as you do not have to invert very large matrices.) This approach to nonparametric regression is called, awkwardly, *semi-nonparametric estimation*. The traditional econometric practice of adding variables to regression models as sample sizes grow, and using some criterion based on t-statistics to determine how many variables to keep in, can be interpreted as a version of this approach to estimation. What nonparametric econometrics adds is a mechanism for choosing the number of terms in an "optimal" way, and an analysis that determines the statistical properties of the result.

More recently it has become common to use a functional approximation approach with functions whose determination is more local; popular functional forms are *splines*, *neural nets*, and *wavelets*. This approach is called the *method of sieves*. Loosely speaking, splines are piecewise polynomials, neural nets are nested logistic functions, and wavelets are piecewise trigonometric functions. Another approach to nonparametric estimation is *penalized maximum likelihood*, in which the log likelihood of the sample, written in terms of the infinite-dimensional unknown function, is augmented with a penalty function that controls the "roughness" of the solution.

All the nonparametric estimation methods listed above will be consistent, in the sense that the mean square error MSE(x) of $T_n(x)$ at a given point x converges to zero, with asymptotically normal distributions (although not at a root-n rate) under suitable regularity conditions and choices of estimation tuning features such as bandwidth. Further, the conditions on the underlying problem needed to get this result are essentially the same for all the methods. An important result, due to Chuck Stone, is that given sample size, the dimensionality of a problem, and the smoothness that can be assumed for the regression function, there is a maximum rate at which MSE(x) can decline. Any of the estimation methods listed above can achieve this maximum rate. Thus, at least in terms of asymptotic properties, one method is as good as the next. In practical sample sizes, there are no general results favoring one method over another. Kernel methods are usually the easiest to compute at a point, but become computationally burdensome when an estimator is needed for many points. Nearest neighbor estimators require large sorts, which are time-consuming. The method of sieves involves more computational overhead, but has the advantage of being "global" so that once the coefficients of the series expansion have been estimated, it is easy to produce forecasts for different points. The method of sieves is currently the most fashionable approach, particularly using neural net or wavelet forms which have been spectacularly successful in recovering some complex test functions. On the whole, nonparametric methods in finite samples place a considerable burden on the econometrician to decide whether nonlinearities in nonparametric estimators are true features of the data generation process, or are the result of "over-fitting" the data.

Consistency:

As in the case of the histogram estimator of a density, good large sample properties of a locally weighted estimator are obtained by giving sufficient weight to nearby points to control variance, while down-weighting distant points to control bias. As sample size increases, distant observations will be down-weighted more strongly, since there will be enough observations close by to control

the variance. The following theorem, adapted from C. Stone (1977), gives sufficient conditions for consistency of a locally weighted estimator.

Theorem 1. Assume (i) g(x) has a convex compact support $\mathbf{B} \subseteq \mathbb{R}^m$; (ii) $\theta(x)$ satisfies a Lipschitz property $|\theta(x') - \theta(x)| \leq L|x' - x|$ for all $x', x \in B$; (iii) the conditional variance of $t(y,x)$ given x, denoted $\Omega(x)$, satisfies $\Omega_o \leq \Omega(x) \leq \Omega_1$, where Ω_o and Ω_1 are finite positive definite matrices; (iv) a random sample i = 1,..,n is observed; and (v) as $n \to +\infty$ the local weights w_{ni} satisfy

$$\text{(a)} \quad E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2 \to 0$$

$$\text{(b)} \quad E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \to 0$$

$$\text{(c)} \quad E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x - x_i| \to 0 .$$

Then $T_n(x) - \theta(x)$ converges to zero in mean square.

Proof: The bias of the estimator is

$$B_n(x) = E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)[\theta(x_i) - \theta(x)] + \theta(x) \left\{ \boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \right\} ,$$

so that assumption (v), (b) and (c) imply

$$|B_n(x)| \leq L \cdot E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x_i - x| + \theta(x) \left\{ \boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \right\} \to 0.$$

The variance of the estimator is, by assumption (v), (a),

$$V_n(x) = E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2 \Omega(x_i) \leq \Omega_1 \; E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2.$$

Then, $MSE = V_n(x) + B_n(x)^2 \to 0$, completing the proof. ■

It is useful to work out conditions on nearest neighbor and kernel estimators that satisfy the sufficient conditions in Theorem 1. First, consider a uniform nearest neighbor estimator, with r_n

points included in the neighborhood at sample size n. Then, $w_{n(i)} = 1/r_n$ for the points in the neighborhood. The LHS of condition (v), (b) in Theorem 1 equals $1/r_n$, so the condition is satisfied if $r_n \to +\infty$. Next, we show that a sufficient condition for (v), (c) in Theorem 1 is $r_n/n \to 0$. Let $N_t(x)$ denote a neighborhood of x of radius t. For any $\lambda > 0$, define τ_n such that $g(N_{\tau_n}) = (1+\lambda)r_n/n$, and note that $r_n/n \to 0$ and $x \in \mathbf{B}$ implies $\tau_n \to 0$. Let $\mathbf{R}_n$ denote the (random) number of observations in the neighborhood N_{τ_n}; then $\mathbf{E}\mathbf{R}_n = n\ g(N_{\tau_n}) = (1+\lambda)r_n$ and $\mathrm{Var}(\mathbf{R}_n) = n\ g(N_{\tau_n})\ [1 - g(N_{\tau_n})\ ] \le (1+\lambda)r_n$. Let $\mathbf{T}_n$ denote the (random) radius of the neighborhood that contains exactly r_n of the observations x_i. Then

$$P(\mathbf{T}_n > \tau_n) = P(\mathbf{R}_n < r_n) = P(\mathbf{R}_n - \mathbf{E}\mathbf{R}_n < r_n - (1+\lambda)r_n) = P(R_n - \mathbf{E}R_n < -\lambda r_n)$$
$$\le \mathrm{Var}(\mathbf{R}_n)/\lambda^2 r_n^{\,2} \le (1+\lambda)/\lambda^2 r_n,$$

with the first inequality obtained by applying Chebyshev's inequality to the sum of the independent random indicators for the events $x_i \in N_{\tau_n}$; these indicators sum to $\mathbf{R}_n$. From this result, and a bound $|x - x'| \le M$ for $x, x' \in B$ implied by the compactness of B,

$$\mathbf{E}\mathbf{T}_n \le \tau_n \cdot P(\mathbf{T}_n \le \tau_n) + M \cdot P(\mathbf{T}_n > \tau_n) \le \tau_n + M(1+\lambda)/\lambda^2 r_n \to 0.$$

Then,

$$E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x - x_i| \le \mathbf{E}\mathbf{T}_n \to 0,$$

establishing that (v), (c) in Theorem 1 holds. The kernel estimator of $\theta(x)$ is

$$T_n(x) = \frac{\frac{1}{n \cdot h^m} \sum_{i=1}^{n} t(y_i,x_i) \cdot K\left(\frac{x - x_i}{h}\right)}{\frac{1}{n \cdot h^m} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}.$$

Note that this estimator is of the generic form $T_n(x) = \sum_{i=1}^{n} w_{in} t(y_i,x_i)$, where the w_i are weights that sum to one. Because the kernel $K\left(\frac{x - x_i}{h}\right)$ is small unless x_i is near x, the weights w_i will be concentrated on points with x_i near x. Then, this estimator corresponds to intuition on how a non-parametric estimator can be constructed. You will recognize the denominator in the formula for $T_n(x)$ is simply a kernel estimator of $g(x)$. The numerator is an estimator of $\int t(y,x) \cdot f(y|x) dy \cdot g(x)$. Then, $T_n(x)$ can be interpreted as an estimator of $\int t(y,x) \cdot f(y|x) dy = [\int t(y,x) \cdot f(y|x) dy \cdot g(x)]/g(x)$.

Now suppose that $\theta(x)$ and $g(x)$ are continuously differentiable to order p, with Lipschitz order p derivatives, and that the kernel is of order $s \le p$. Also assume that $\sigma^2(x)$ is finite and Lipschitz in

x. As in the case of density estimation, require that $h \to 0$ and $nh^m \to +\infty$ as $n \to +\infty$. This will ensure that the numerator of $T_n(x)$ converges in mean square error to $\theta(x)\cdot g(x)$ and that the denominator converges in mean square error to $g(x)$, so that the ratio is a consistent estimator of $\theta(x)$.

Arguments similar to those for density estimation are used to establish further statistical properties of $T_n(x)$. Treat the numerator and the denominator separately. The denominator is the earlier density estimator, where we found that the bias satisfied $\text{Bias}_{\text{denom}}(x) = C\cdot h^{s+1}$, where C is a constant. Make a Taylor's expansion of the function $q(x - hz) \equiv \theta(x - hz)\cdot g(x - hz)$ to order s:

$$q(\mathbf{x} - h\mathbf{z}) = \sum_{j=0}^{s} \frac{(-h)^j}{j!} \sum_{|r|=j} q^r(x)\cdot z^r + \lambda\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |q^r(x)\cdot z^r|\cdot L'|z| .$$

Then, the numerator satisfies

$$\mathbf{E}\ \frac{1}{nh^m}\sum_{i=1}^{n} t(y_i,x_i)\cdot K\left(\frac{x - x_i}{h}\right) = \int g(x-hz)\cdot\theta(x-hz)\cdot K(z)dz = g(x)\cdot\theta(x) - \lambda''\cdot A'\cdot h^{s+1},$$

where A' is a constant that depends on the order s derivatives of t, and on the Lipschitz constant L'. Then, $\text{Bias}_{\text{numer}}(x) = \lambda''\cdot A'\cdot h^{s+1}$.

The variance of the denominator, from the previous analysis, is $\frac{g(x)}{nh^m}\int K(z)^2dz + \text{HOT}$. An analogous argument applied to the numerator establishes that its variance is $\frac{\sigma^2(x)\cdot g(x)}{n\cdot h^m}\int K(z)^2dz$ + HOT. The covariance of the numerator and denominator is zero.

Consider a ratio α_n/β_n of random variables α_n and β_n that have finite second moments, satisfy $\alpha_n \to_p \alpha_0$ and $\beta_n \to_p \beta_0$ as $n \to +\infty$, and have β_n uniformly bounded and bounded away from zero. Then, $\mathbf{E}\alpha_n \to \alpha_0$, $\mathbf{E}\beta_n \to \beta_0$, and the ratio can be rewritten

$$\frac{\alpha_n}{\beta_n} - \frac{\alpha_0}{\beta_0} = \frac{\frac{\alpha_n - \boldsymbol{E}\alpha_n}{\boldsymbol{E}\beta_n} - \frac{\alpha_0}{\beta_0}\cdot\frac{\beta_n - \boldsymbol{E}\beta_n}{\boldsymbol{E}\beta_n} + \frac{\boldsymbol{E}\alpha_n - \alpha_0}{\boldsymbol{E}\beta_n} - \frac{\alpha_0}{\beta_0}\cdot\frac{\boldsymbol{E}\beta_n - \beta_0}{\boldsymbol{E}\beta_n}}{1 + \frac{\beta_n - \boldsymbol{E}\beta_n}{\boldsymbol{E}\beta_n}} .$$

The expectation of the square of this expression is the mean square error of α_n/β_n. For n large, the denominator is almost always very close to one, and is rarely close to zero. The expectation of the square of the numerator can be written

$$\frac{\boldsymbol{V}\alpha_n}{\beta_0^2} + \left(\frac{\alpha_0}{\beta_0}\right)^2\cdot\frac{\boldsymbol{V}\beta_n}{\beta_0^2} - \frac{2\alpha_0}{\beta_0}\cdot\frac{\boldsymbol{cov}(\alpha_n,\beta_n)}{\beta_0^2} + \left(\frac{bias_\alpha}{\beta_0} - \frac{\alpha_0\, bias_\beta}{{\beta_0}^2}\right)^2$$

Applying this formula to the numerator and denominator of $T_n(x)$, substituting the expressions just derived for variances and biases, the mean square error in $T_n(\mathbf{x})$ is

$$\text{MSE}(x) = \frac{\sigma^2(x)}{n\cdot h^m\cdot g(x)}\int K(z)^2dz + \frac{\theta(x)^2}{n\cdot h^m\cdot g(x)}\int K(z)^2dz + h^{2(s+1)}\cdot\frac{C}{g(x)^2} ,$$

where C is a constant depending on order s derivatives, Lipschitz constants, and K. The h_{opt} that minimizes MSE(x), or the integral MISE of MSE(x) over a domain where $g(x)$ is bounded positive, is proportional to $n^{-1/(m+2(s+1))}$, and the mean square error criterion is proportional to $n^{-2(s+1)/(m+2(s+1))}$, just as in the case of density estimation. Again, the precision of the estimator falls when dimensionality m rises, and high-dimension problems require immense sample sizes to achieve accurate estimators. A high degree of smoothness, exploited using high-order kernels can offset some of the negative impacts of dimensionality, but can never get mean square error to fall at a 1/n rate. As in the case of density estimation, a least squares cross-validation procedure can be used to determine an approximately optimal bandwidth in applications. W. Hardle and O. Linton (1994) give the formulas.

Optimal Rates

The number of observations included in a nearest neighbor estimator, or the bandwidth in a kernel estimator, can vary over considerable ranges and still produce consistent estimators. However, there are typically optimal values for these design parameters that minimize mean square error. These values depend on the properties of the function being estimated, but their qualitative properties are of interest. These notes mentioned earlier the result of Stone that there will be a best rate at which MSE(x) declines, for *any* nonparametric method, and that all the standard methods can achieve this rate. This best rate of decline turns out to be very slow when the dimension m of x is large. This is called the *curse of dimensionality*, and is a consequence of the fact that when dimensionality is high, data are more sparse. (This proposition can be made precise by considering the statistical problem of the expected radius of the largest sphere that can be circumscribed around a data point without encountering any other data points. For a given sample size, this expected radius rises with dimension m at a rate that corresponds to the curse of dimensionality.)

I will give a rough outline of an argument that determines the optimal bandwidth for kernel estimation in the case that $\theta(x)$ is Lipschitz, and after that a rough outline of an argument that determines the optimal number of neighbors for nearest neighbor estimation. These arguments draw heavily from the demonstrations following the proof of Theorem 1, and parallel the arguments for consistent kernel estimation of a multivariate density given earlier.

Kernel Estimation: From the earlier analysis, the variance of the estimator is approximately proportional to $K(0)/g(x)nh^m$, and the bias is approximately proportional to h. Then, the first-order-condition for minimization of variance plus squared bias is $h_n = D/n^{1/(m+2)}$ for a constant D, and the corresponding MSE declines at rate $n^{-2/(2+m)}$. For m = 1, this is the same $n^{-2/3}$ rate that was achieved by the optimal histogram estimator of a Lipschitz density.

Nearest Neighbor Estimation: From the earlier analysis, if there are r observations in the neighborhood, with $r \to +\infty$ and $r/n \to 0$, then the estimator is a (weighted) average of r observations, so that its variance is approximately D_0/r, where D_0 is a constant that does not depend on r. The volume of a sphere of radius t in $\mathbb{R}^m$ is $C_m t^m$, where C_m is a constant depending only on m. Then, for $g(x) > 0$, the radius τ_n of a neighborhood that is expected to contain $(1+\lambda)r$ points satisfies $(1+\lambda)r/n = g\ (N_{\tau_n}) \approx g(x)C_m\tau_n^m$ and the random radius $\mathbf{T}_n$ of a neighborhood that contains exactly r points satisfies $\mathbf{E}\mathbf{T}_n \le \tau_n + D_1/r \approx D_2(r/n)^{1/m} + D_1/r$ for some constant D_2. Suppose for the moment that we omit the D_1 term. Then, the first-order-condition for minimizing the sum of variance and squared bias is $D_0/r_n = (D_2/m)r_n \cdot n^{-2/m}$, which implies that the optimal r_n is proportional to $n^{2/(2+m)}$. Substituting

this into the formula for the bias shows that at this rate the D_1 term becomes negligible relative to the D_2 term, justifying its omission. Finally, when r_n is proportional to $n^{2/(2+m)}$, the MSE declines at the rate $n^{-2/(2+m)}$.

The common rate $n^{-2/(2+m)}$ at which MSE declines for the "best" nearest neighbor and kernel estimators of a Lipschitz nonparametric regression is in fact the maximum rate found by Stone for a problem of m dimensions with Lipschitz θ that has no further known smoothness properties. Hence the rates above for the number of neighbors and for bandwidth are also "best". Note that for m even moderately large, the rate of decline of MSE is agonizingly slow. When m = 8 for example, to reduce MSE by a factor of 10, it is necessary to increase sample size by a factor of 100,000. This is the curse of dimensionality in action. The only way to circumvent this problem is to assume (and justify the assumption) that θ is differentiable to high order, and use this in constructing the nonparametric estimator, or to assume that θ depends only on low-dimensional interactions of the variables, e.g., θ is a sum of functions of the variables taken two at a time.

Asymptotic Normality

Returning to the general family of locally weighted estimators, we look for conditions, in addition to those guaranteeing consistency, that are sufficient to establish that the nonparametric estimator is asymptotically normal. The following theorem gives a general result; the added conditions are (iv) and in (vi), strengthened conditions (b) and (c), and new conditions (d)-(f):

Theorem 2. Assume (i) g(x) has a convex compact support $B \subseteq \mathbb{R}^m$; (ii) $\theta(x)$ satisfies a Lipschitz property $|\theta(x') - \theta(x)| \le L|x' - x|$ for all $x', x \in B$; (iii) the conditional variance of $t(y,x)$ given x, denoted $\Omega(x)$, satisfies $\Omega_o \le \Omega(x) \le \Omega_1$, where Ω_o and Ω_1 are finite positive definite matrices; (iv) $\mathbf{E}_{y|x}|t(y,x) - \theta(x)|^3 \le A|\Omega(x)|^{3/2}$ for some constant A; (v) a random sample i = 1,..,n is observed; and (vi) as $n \to +\infty$ the local weights w_{ni} satisfy

(a) $$\sum_{i=1}^{n} E_{\{x_i\}}\ w_{ni}^2(x;x_1,...,x_n) \to 0$$

(b) $$\left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2(x;x_1,...,x_n)\Omega(x_i)\right)^{-1/2} \left\{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1\right\} \to 0$$

(c) $$\left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2(x;x_1,...,x_n)\Omega(x_i)\right)^{-1/2} E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)|\cdot|x - x_i| \to 0$$

(d) $$\frac{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)|^3}{\left\{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2\right\}^{3/2}} \to 0$$

(e) $$\frac{\left\{\boldsymbol{E}_{\{x_i\}}\sum_{i=1}^{n}|w_{ni}(x;x_1,...,x_n)|.|x-x_i|\right\}^2}{\boldsymbol{E}_{\{x_i\}}\sum_{i=1}^{n}w_{ni}(x;x_1,...,x_n)^2|\Omega(x_i)|} \to 0$$

(f) $$\frac{\left\{\boldsymbol{E}_{\{x_i\}}\sum_{i=1}^{n}w_{ni}(x;x_1,...,x_n) - 1\right\}^2}{\boldsymbol{E}_{\{x_i\}}\sum_{i=1}^{n}w_{ni}(x;x_1,...,x_n)^2|\Omega(x_i)|} \to 0$$

Then $\left\{\boldsymbol{E}_{\{x_i\}}\sum_{i=1}^{n}w_{ni}(x;x_1,...,x_n)^2\Omega(x_i)\right\}^{-1/2}$ $\{T_n(x) - \theta(x)\}$ converges in distribution to $N(0,\mathbf{I})$.

Proof: We make use of the following central limit theorem, which is a corollary of the Lindeberg-Feller theorem for triangular arrays; see Serfling (1980, 1.9.3, Corollary, p. 32): *For each n, let* ζ_{ni} *for* $i \le n$ *be independent random variables with mean zero, finite variances* σ_{ni}^2, *and for some* $\nu > 2$, $\left(\sum_{i=1}^{n}\boldsymbol{E}|\zeta_{ni}|^{\nu}\right) / \left(\sum_{i=1}^{n}\sigma_{ni}^2\right)^{\nu/2} \to 0$. *Then,* $\left(\sum_{i=1}^{n}\zeta_{ni}\right) / \left(\sum_{i=1}^{n}\sigma_{ni}^2\right)^{1/2} \to_d N(0,\mathbf{I})$.

Assume that $T_n(x)$ is a scalar, or else consider a fixed linear combination of components. Define $\zeta_{ni} = w_{ni}[t(y_i,x_i) - \theta(x_i)]$; then for each n, the ζ_{ni} are independent with finite variances $\sigma_{ni}^2 = w_{ni}^2\Omega(x_i)$. Hypotheses (iv) and (vi), (d) imply

$$\left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}|w_{ni}|^3|t(Y,x_i) - \theta(x_i)|^3\right) / \left(\sum_{i=1}^{n}\sigma_{ni}^2\right)^{3/2}$$

$$\le A\left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}|w_{ni}|^3|\Omega(x_i)|^{3/2}\right) / \left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}w_{ni}^2\Omega(x_i)\right)^{3/2}$$

$$\le A(|\Omega_1|/|\Omega_o|)\left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}|w_{ni}|^3\right) / \left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}w_{ni}^2\right)^{3/2} \to 0.$$

Finally, consider the scaled bias term

$$\left(\sum_{i=1}^{n}\boldsymbol{E}_{\{x_i\}}w_{ni}^2\Omega(x_i)\right)^{-1/2}\ [\ \boldsymbol{E}_{\{x_i\}}\ T_n(x) - \theta(x)]$$

$$= \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2 \Omega(x_i) \right)^{-1/2} \left\{ \sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni} [\theta(x_i) - \theta(x)] + \theta(x) [\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni} - 1] \right\} .$$

This converges to zero by (vi), (e) and (f). Then, the limiting distribution has mean zero. ■

Consider the "best" kernel and nearest neighbor estimators. The assumptions on these estimators made in the discussion of consistency and best rates, along with assumptions (i)-(v) in Theorem (ii), are sufficient to establish (vi), (a)-(d). These in turn are sufficient to establish consistency and asymptotic normality, but possibly with a non-zero mean. A device introduced by Herman Bierens allows one to get this asymptotic mean to zero while preserving the "best" rate. I will explain the trick for a nearest neighbor estimator. Suppose $r_n = Dn^{2/(2+m)}$ and $r_n' = 2^m r_n$ are two cutoff numbers for nearest neighbor estimation, both growing at the "best" rate, where D is some constant. Let $T_n(x)$ and $T_n'(x)$ be the corresponding estimators. Since $r_n' > r_n$, the estimator $T_n'(x)$ will have a larger bias and a smaller variance than $T_n(x)$. Now consider an estimator $T^*(x) = 2T_n(x) - T_n'(x)$. This estimator is also a locally weighted estimator, with weights that are the {2,-1} linear combination of the weights for the two original estimators. It is easy to check that these weights satisfy the same properties in Theorems 1 and 2 as do the original weights, so that $T^*(x)$ is consistent for $\theta(x)$. These combined weights increase at the "best" rate $n^{1/(2+m)}$, so that $T^*(x)$ is again a "best" estimator. Recall from the discussion of optimal rates that except for terms that are negligible in large samples, the bias for a nearest neighbor estimator with $r = Cn^{2/(2+m)}$ points is proportional to $(r/n)^{1/m} = C^{1/m} n^{-1/(2+m)}$. For $T_n(x)$, C = D, while for $T_n'(x)$, $C = 2^m D$. Therefore, except for higher-order terms, the bias in $T^*(x)$ is proportional to $2D^{1/m} n^{-1/(2+m)} - (2^m D)^{1/m} n^{-1/(2+m)} = 0$. Then, there is a "best" nearest neighbor estimator that is asymptotically normal with mean zero. The weights for the estimator $T^*(x)$ can be interpreted as "higher order" weights that remove more bias; note that these weights are sometimes negative. This trick has reduced bias, at the expense of increasing variance, since the variance of $T^*(x)$ is greater than that of $T_n(x)$, while leaving the "best" rate unchanged. A similar device works for kernel estimators, using a higher-order kernel that is a linear combination of two kernels whose bandwidths differ by a multiplicative constant.

Exercise 4. Find the appropriate constants for a second-order kernel that removes asymptotic bias from the estimator so that its asymptotic distribution is centered at zero.

.

5. SEMIPARAMETRIC ANALYSIS

Semiparametric methods provide estimates of finite parameter vectors without requiring that the complete data generation process be assumed in a finite-dimensional family. By avoiding bias from incorrect specification, such estimators gain robustness, although usually at the cost of decreased precision. The most familiar semiparametric method in econometrics is ordinary least squares, which estimates the parameters of a linear regression model without requiring that the distribution of the disturbances be in a finite-parameter family. The recent literature in econometric theory has extended semiparametric methods to a variety of nonlinear models. Four overlapping major areas are models for censored duration data (e.g., employment duration); limited dependent variable (partial observability) models for discrete or censored data (e.g., employment status or employment hours); models for data with (natural or intentional) endogenous sample selection (e.g., wage

determination among self-selected workers, or case-control sampling); and models for additive non-parametric effects. The following table summarizes some applications.

Model	**Applications**
Regression and Single Index Models for Censored Duration Data: $Y\|x \cong Y\|x'\beta$	Employment Duration, Innovation Lags, Mobility
Limited Dependent Variable Models (E.g., Discrete response or censored response) $Y^* = x'\beta - \varepsilon$, $\varepsilon\|x \sim F(\cdot)$, *observability transformation* $Y = \Psi(Y^*)$ E.g., Discrete: $Y = \text{sgn}(Y^*)$, Censored: $Y = \text{Min}(Y^c, Y^*)$	Discrete: Employment Status, Brand Choice Censored: Employment Hours, Expenditure Levels
Endogenous Sample Selection $Y = x'\beta - \varepsilon$, $\varepsilon\|x \sim f(\cdot)$, $x \sim g(\cdot)$, Natural: (Y,x) observed iff $Y > 0$ Intentional: (Y,x) sampled iff $Y > 0$ $P(Y,x\|\text{Obs}) = \dfrac{f(Y-x'\beta)g(x)\mathbf{1}(Y>0)}{\int_{z=-\infty}^{+\infty}\int_{y=0}^{+\infty} f(y-z'\beta)g(z)dydz}$ $P(Y\|x,\text{Obs}) = f(Y-x'\beta)/\int_{y=0}^{+\infty} f(y-x'\beta)dy$	Natural: Self-selected Workers, Self-selected Homeowners Intentional: Case-Control Sample Designs
Additive Non-Parametric Effects: $Y = x'\beta + H(z) + \varepsilon$	Robust policy analysis

In most cases, the primary focus of semiparametric analysis is estimation of coefficients of covariates that index the location of the distribution of a dependent variable; then, the unknown distribution is a (infinite-dimensional) nuisance parameter. There are also applications where some functional of the unknown distribution, such as the expectation of the dependent variable conditioned on covariates, is of primary interest. The final objective may be point estimates or confidence intervals for the objects of interest, or hypothesis tests involving these parameters. Usually, it is important to have measures of precision for the estimates of interest, including convergence rates, asymptotic distributions, and bootstrap or other indicators of finite-sample precision and accuracy of asymptotic approximations.

These notes will not survey the full range of semiparametric models in econometrics, or develop the properties of semiparametric estimator except for illustrative cases. A good survey of the foundations of semiparametric analysis can be found in Powell (1994). These notes will instead survey two areas of application. The first is the analysis of censored employment duration data, perhaps the leading case of applied semiparametric work. The second is the analysis of data on stated willingness-to-pay for natural resources.

Censored Employment Duration

The main focus of the literature on employment duration has been the effect of covariates such as sex, race, age, and education on the hazard of leaving a job. Data on employment duration is typically censored because employment spells start before a panel study is initiated (and the start date may not be recovered accurately using retrospective questions) and/or continue past the end of the panel study, or because of attrition from the panel. In this chapter, we consider only right-censoring before the end of a spell. Parametric analysis of the duration problem has typically used exponential or Weibull survival curves, or the Cox proportional hazards model, which qualifies as one semiparametric formulation.

Horowitz and Newmann (1987) make perhaps the first empirical application of semiparametric censored regression methods to data on employment duration. To provide some context for the economic application, consider the hazards that may lead to termination of a spell of employment. First, termination may be initiated either by the employee (quits), or the employer (layoffs, separations). The quit decision of an employee is presumably influenced by nonpecuniary job features (e.g., safety, variety, and work rules), wage opportunity cost, and worker characteristics such as education, race, and loyalty. The termination decision of the firm is influenced by the expected productivity of the worker, net of wages. The worker's job-specific human capital influences both wage opportunity cost and expected productivity. Wage opportunity cost is also influenced by expected unemployment insurance benefits and duration of unemployment. Macroeconomic and product cycles influence expected productivity. Several aspects of this description are important for modeling employment duration:

1. Quits and separations are competing risks, with overlapping but not identical covariates. Structural estimates of duration must distinguish these two hazards. Data on whether employment spells end in quits would greatly aid identification and estimation of the separate hazards.
2. Important covariates such as the level of macroeconomic activity and job-specific human capital vary in elapsed or chronological time, so a structural model must accommodate time-varying covariates. To do this is fairly easy in discrete time using heterogeneous Markov models, and quite difficult in continuous time.
3. Unobserved variables such as worker loyalty are heterogeneous in the population and are selected by survival. Thus, it is necessary for structural modeling of duration to determine the distribution of these unobservables. The presence of unobserved heterogeneity also selects the subpopulation that start employment spells during the interval of observation. The subpopulation starting employment spells near the beginning of the observation interval will be less loyal on average than all workers. Those whose first observed employment spell start comes near the end of the observation period will be more loyal on average if the panel is long enough.
4. In a structural model of employment duration, the hazard must depend solely on the history of economic variables, and not directly on elapsed time. Thus, models that postulate a reduced-form "baseline" hazard are removing variation that must have a structural source. From the standpoint of structural estimation of the economic determinants of duration, emphasis on the effect of covariates with the baseline hazard treated as a nuisance parameter is misplaced.
5. Economic theory provides neither a tight specification of functional forms or the distributions of unobservables; the assumption that observables enter in a parametric additive combination must be justified as an approximation. Consequently, analyses that assume observables appear in an exact additive combination within unknown transformations or distributions in fact assume too much on the structure of the additive combination, and perhaps too little on the unknown

transformations, which may be approximable to comparable accuracy using flexible finite-parameter families.

The duration data generation process can be characterized by a *survival curve* $q(t|x)$ stating the proportion of a population with spells starting at time zero who survive at elapsed time t, given an observed covariate process $x(\cdot)$. If there are unobserved covariates ξ distributed in the initial population with density $\nu(\cdot|x,0)$, and the "structural" survival curve is $q(t|x,\xi)$, then the data generation process satisfies

$$(1) \qquad q(t|x) = \int_{-\infty}^{+\infty} q(t|x,\xi)\cdot\nu(\xi|x,0)d\xi.$$

The density of the unobserved covariates, conditioned on survival, is modified over time by selection, satisfying

$$(2) \qquad \nu(\xi|x,t) = \nu(\xi|x,0)q(t|x,\xi)/q(t|x).$$

The survival curve can also be described by the *hazard rate*,

$$(3) \qquad h(t|x,\xi) = -\nabla_t \mathrm{Ln}(q(t|x,\xi)).$$

The *average hazard rate* in the surviving population is

$$(4) \qquad h^*(t|x) = -\nabla_t \mathrm{Ln}(q(t|x))$$

$$= \left(\int_{-\infty}^{+\infty} h(t|x,\xi)q(t|x,\xi)\nu(\xi|x,0)d\xi\right) /q(t|x) = \int_{-\infty}^{+\infty} h(t|x,\xi)\nu(\xi|x,t)d\xi\ .$$

Equation (3) can be inverted to obtain

$$(5) \qquad q(t|x,\xi) = \exp\left(-\int_0^t h(s|x,\xi)ds\right) \equiv \exp\left(-\Lambda(t|x,\xi)\right)\ ;$$

with $\Lambda(t|x,\xi)$ termed the *integrated hazard*. The mean duration of completed spells is

$$(6) \qquad \mathbf{E}(t|x,\xi) = -\int_0^\infty t\,\nabla_t\, q(t|x,\xi)dt = \int_0^\infty q(t|x,\xi)dt,$$

with the second formula obtained using integration by parts.

When the observation interval is finite, some spells are *interrupted* or *right-censored*; the survivor function defined up to the censoring point continues to characterize the data generation process. The mean duration of all spells whether ended naturally (at t) or by censoring (at t^c) is

(7) $$\mathbf{E}(\text{Min}(t,t^c)) = -\int_0^{t^c} t\,\nabla_t q(t|x,\xi)dt + t^c q(t^c|x.\xi) = \int_0^{t^c} q(t|x,\xi)dt.$$

Analogous formulas hold for the average hazard rate.

With sample attrition, the censoring time becomes a random variable, with an associated censoring survivor function $r(t^c|x,\xi)$. Then the probability that a spell is observed to extend to t is $q(t|x,\xi)r(t|x,\xi)$; the combined hazard rate for termination of an observed spell either naturally or by censoring is $h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi)$; for a spell ending at time t, the probability that it is censored is $h(t|x,\xi)/(h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi))$; and the mean duration of observed spells is

$$\int_0^{\infty} q(t|x,\xi)r(t|x,\xi)dt.$$

An example of a parametric duration model when x is time-invariant is the *Weibull* model, which specifies

(8) $$q(t|x) = \exp(-t^{\alpha}e^{-x'\beta}),$$

with α a positive parameter, β a vector of parameters, and x a vector of covariates. The associated hazard rate is

(9) $$h(t|x) = \alpha t^{\alpha-1}e^{-x'\beta}$$

and the mean duration of completed spells is

(10) $$\mathbf{E}(t|x) = e^{x'\beta/\alpha}\Gamma(1+1/\alpha),$$

where Γ is the gamma function. When $\alpha = 1$, this simplifies to the *exponential* duration model.

There are three strategies for statistical inference of censored duration data:

1. The fully parametric approach, with $q(t|x)$, or in the case of unobserved heterogeneity $q(t|x,\xi)$ and $\nu(\xi|x,0)$, assumed to be in a finite-parameter family.[14]

[14]Typical examples are a Weibull or log-normal distribution for $q(t|x)$, or an exponential distribution for $q(t|x,\xi)$ combined with a gamma distribution for ξ. The parameters of the distribution can be estimated by maximum likelihood.

2. The fully nonparametric approach, in which $q(t|x)$ is estimated without parametric restrictions, using for example a Kaplan-Meier estimator.[15]

3. The single-index semiparametric approach, in which $q(t|x)$ depends on x through a scalar function $V(x,\beta)$ that is known up to a finite parameter vector β, but $q(t|v)$ is not confined to a parametric family. In the case of unobserved heterogeneity, either $q(t|v,\xi)$ or $\nu(\xi|v,t)$ may be nonparametric (but not both, without further restrictions, due to identification requirements).[16]

We survey some of the alternative semiparametric problems that have been discussed in the literature. Let x be a vector of covariates, assumed now to be *time-invariant*. Let β be a vector of unknown parameters, $V(x,\beta) \equiv x'\beta$ be a single index function known up to β, and $q(t|x'\beta)$ the survivor function. Let T^* be the random variable denoting completed duration, and T^c the censoring time, so observed duration is $T = \text{Min}(T^*,T^c)$. Four alternative models for T are

1. *Regression model*: $\text{Ln } T^* = x'\beta + \varepsilon$, with $\varepsilon|x$ distributed with an unknown density $f(\varepsilon)$ with zero mean. The density f is often assumed symmetric and homoskedastic. This model yields the survivor function

$$q(t|x'\beta) = 1 - F(\text{Ln } t - x'\beta), \tag{11}$$

where F is the cumulative distribution function of f. The associated hazard rate is

$$h(t|x'\beta) = f(\text{Ln } t - x'\beta)/t[1 - F(\text{Ln } t - x'\beta)]. \tag{12}$$

A generalized version of this model allows ε to be heteroskedastic, with variance depending on the index $x'\beta$, or more generally on some other function of x. The *censored regression model* is simply

$$\text{Ln } T = \text{Min}(\text{Ln } T^c, x'\beta+\varepsilon); \tag{13}$$

---

[15]The classical Kaplan-Meier estimator is formulated for duration data without covariates. Suppose that in a data set spells starting at a common time 0 are observed to end (naturally or by censoring) at times $t_1 < ... < t_J$. Let n_j denote the number that end naturally at time t_j, and let m_j denote the number that are censored at this time. The total number "at risk" at time t_j is $N_j = \sum_{i=j}^{J} (n_i+m_i)$. The Kaplan-Meier estimate of the hazard rate at t_j is $h^*(t_j) = n_j/N_j$. A corresponding estimate of the survival function is $q^*(t_j) = (1-h^*(t_j))q^*(t_{j-1})$, or $q^*(t_j) = \prod_{i=1}^{j} (1-n_j/N_j)$. In the presence of categorical covariates, the Kaplan-Meier estimator obviously applies cell-by-cell for each configuration of the covariates. Using the nearest neighbor idea from non-parametric regression, the Kaplan-Meier estimator can be adapted to the general case of non-categorical covariates. In the case of unobserved heterogeneity, it is not possible in general to identify the structural survivor functions and the density of the unobserved covariates when both are non-parametric. Heckman and Singer (1984) establish this result, and also establish semiparametric methods for estimation of a parametric structural survivor function $q(t|x,\xi,\beta)$ in the presence of a non-parametric heterogeneity density $\nu(\xi|x,0)$.

[16]Other semiparametric approaches include multiple-index models and methods that parameterize quantiles without fully parameterizing the distribution.

it has the property in the case of non-stochastic censoring that

(14) $$\mathbf{E}(\text{Ln } T|x) = \int [1 - F(y - x'\beta)]dy$$

is an increasing function of $x'\beta$.

2. *Transformation* (Generalized Box-Cox) *model*: Suppose G is an unknown monotone increasing transformation from $(0,+\infty)$ onto the real line, and assume

(15) $$G(T^*) = x'\beta + \varepsilon,$$

with $\varepsilon|x$ distributed with a known or unknown density $f(\varepsilon)$. The associated survivor function is

(16) $$q(t|x'\beta) = 1 - F(G(t) - x'\beta),$$

and the associated hazard rate is

(17) $$h(t|x'\beta) = G'(t)f(G(t) - x'\beta)/[1 - F(G(t) - x'\beta)].$$

Again, the model can be generalized to allow heteroskedasticity depending on $x'\beta$.

3. *Projection Pursuit* (single index) *regression*: Suppose H is a unknown transformation from the real line into the real line. Assume

(18) $$\text{Ln } T^* = H(x'\beta) + \varepsilon,$$

with $\varepsilon|x$ distributed with a known or unknown density $f(\varepsilon)$. The associated survivor function is

(19) $$q(t|x'\beta) = 1 - F(\text{Ln } t - H(x'\beta)),$$

and hazard rate is

(20) $$h(t|x'\beta) = f(\text{Ln } t - H(x'\beta))/t[1 - F(\text{Ln } t - H(x'\beta))].$$

The error distribution is usually assumed homoskedastic, but some estimators for this model permit heteroskedasticity depending on $x'\beta$.

4. *Proportional Hazards model*: Let $h_o(t)$ be an unknown nonnegative "baseline hazard" function, and assume the covariates exert a proportional effect on the hazard, so that

(21) $$h(t|x) = h_o(t)\exp(-x'\beta).$$

Define the integrated baseline hazard

(22) $$\Lambda_o(t) = \int_0^t h_o(s)ds.$$

Then the survivor function is

(23) $$q(t|x'\beta) = \exp\left(-\Lambda_o(t)\ e^{-x'\beta}\right) ,$$

and

(24) $$\text{Ln}\ \Lambda_o(T^*) = x'\beta + \varepsilon,$$

where ε has the extreme value cumulative distribution function

(25) $$F(\varepsilon) = 1 - \exp\left(-e^{-\varepsilon}\right) .$$

Other error distributions may result from a proportional hazards model with unobserved heterogeneity. For example, following Lancaster (1979), assume

(26) $$h(t|x,\xi) = h_o(t)\exp(-x'\beta)\xi,$$

with ξ having a gamma density, $\nu(\xi|x,0) = \xi^{\theta-1}e^{-\xi}/\Gamma(\theta)$. Then, applying the relation (1),

(27) $$q(t|x) = \left(1 + e^{\Lambda_o(t) - x'\beta}\right)^{-\theta} ,$$

which implies that (15) holds with ε having a generalized logistic distribution (or, e^{ε} having a Pareto distribution),

(28) $$F(\varepsilon) = 1 - (1+e^{\varepsilon})^{-\theta}.$$

The average hazard for (26),

(29) $$h^*(t|x) = \theta h_0(t)e^{\Lambda_o(t)} / \left(e^{\Lambda_o(t)} + e^{x'\beta}\right) ,$$

is no longer of the proportional hazards form. The conditional distribution of the unobserved covariates given survival $\nu(\xi|x,t)$ remains Gamma with parameter θ, but in the transformed variable

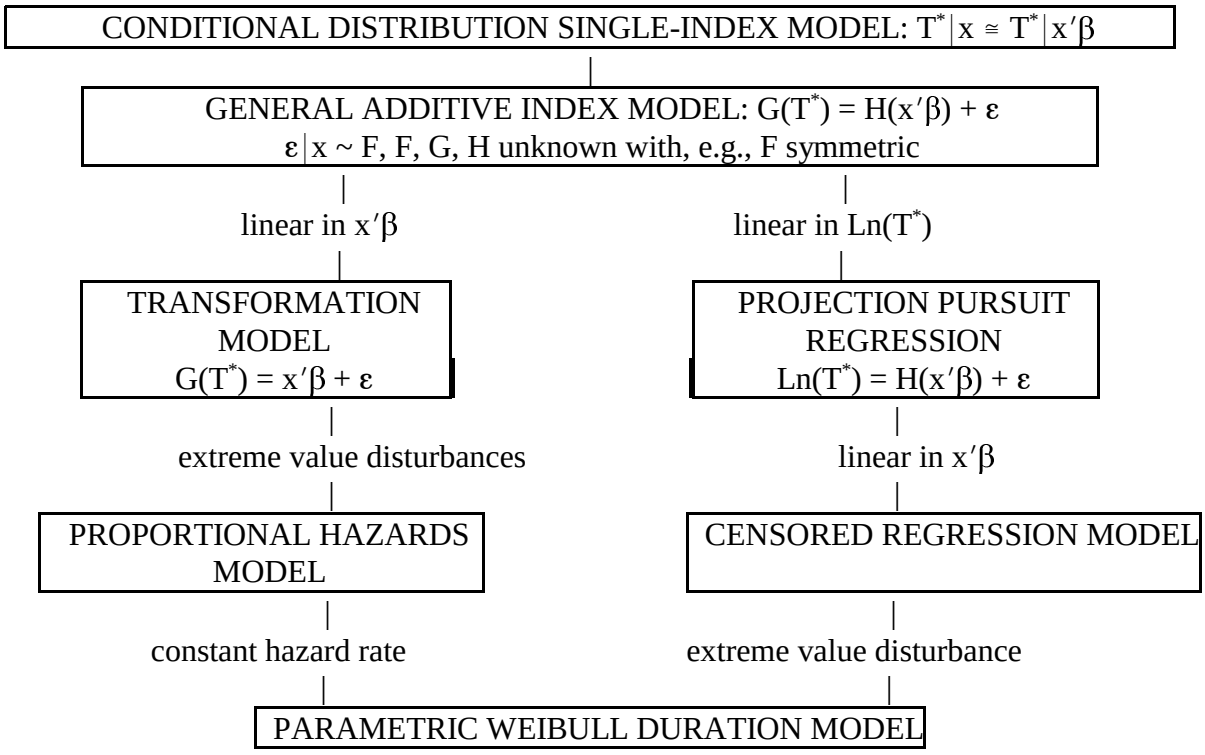
$$(1+e^{\Lambda_o(t) - x'\beta})\xi .$$

The proportional hazards model (21) is a special case of the transformation model where the disturbance has the distribution (25). The proportional hazards model with heterogeneity (26) is also a specialization of the transformation model. When the baseline hazard varies with a power of t, $h_o(t) = \alpha t^{\alpha-1}$, (21) specializes to the parametric Weibull duration model, and also can be interpreted as a censored regression model with extreme value distributed disturbances.

FIGURE 1. SINGLE-INDEX MODELS

Observation Rules: $T = \text{Min}(T^c, T^*)$ for right-censored data
$T = \text{sgn}(\text{Ln}(T^*))$ for binomial discrete response data

(Specificity Increases as You Move Down the Table)



A common "generalized additive single-index" model in which the four models above are nested is

$$G(T^*) = H(x'\beta) + \varepsilon, \tag{30}$$

with ε distributed with cumulative distribution function F. The associated survivor function is

$$q(t | x'\beta) = 1 - F(G(T) - H(x'\beta)). \tag{31}$$

Figure 1 shows the logical relationship between these models. All the models are special cases of *single-index sufficiency* where the conditional distribution of the dependent variable depends on covariates x solely through the index $x'\beta$. The proportional hazards model and the censored regression model are logically distinct, except when both specialize to the common Weibull parametric model. Both are specializations of the transformation model. The censored regression model is a specialization of the projection pursuit regression model. The transformation model can be rewritten as a heteroskedastic projection pursuit model: If $G(T^*) = x'\beta + \varepsilon$ with G monotone

increasing, then Ln $T^* = H(x'\beta) + \zeta$, where $H(x'\beta) = \mathbf{E}_\varepsilon \text{Ln } G^{-1}(x'\beta + \varepsilon)$, and ζ has the distribution function $F(G(\exp(\zeta + H(x'\beta)) - x'\beta)$, which in general is heteroskedastic.

The statistical issues that arise in application of these models are the (large sample and, potentially, small sample) distributional properties of estimators that are available under various assumptions, and the efficiency of alternative estimators. Most of the work to date has concentrated on finding computationally feasible estimators, establishing consistency and asymptotic normality, and establishing asymptotic efficiency bounds.

Horowitz and Newmann use two estimators for the censored regression model, a quantile estimator (Powell, 1986) and one-step semiparametric generalized least squares estimator (SGLS) (Horowitz, 1986). Other estimators that have been proposed for this problem include flexible parametric approximation of the cumulative distribution function (e.g., Duncan, 1986, who considers spline approximations--the "method of sieves"). Chamberlain (1986) and Cosslett (1987) have established for the censored regression problem the existence of a positive information bound on the parametric part. This suggests that the it is adequate to use relatively crude estimators of the nonparametric part in order to achieve |n asymptotically normal estimation of the parametric part. The Powell and Horowitz estimators have been shown |n asymptotically normal. Neither achieves the information bound for i.i.d. errors, and in general neither is efficient relative to the other.

Estimation of the proportional hazards model with an unknown baseline hazard function has been studied extensively; see Kaplan and Meier (1968), Cox (1972), Kalbfleisch and Prentice (1982), and Meyer (1990). A particularly useful "semiparametric" method for this model, applicable to the case where duration is measured in "weeks", is to flexibly parameterize the baseline hazard; Meyer (1990) shows this method is root-n asymptotically normal.

Estimators for the projection pursuit (single index) model have been proposed by Ichimura (1987), Ruud (1986), Stoker (1986), and Powell, Stock, and Stoker (1989). The Ichimura estimator chooses β to minimize the conditional variance of Ln T given $x'\beta$, using a kernel estimator of the conditional mean to form an estimate of the conditional variance. This estimator is consistent even if the disturbances are heterogeneous in the index function, so it can also be applied to the transformation model. The Ichimura estimator is $n^{1/2}$ asymptotically normal, and has recently been argued to achieve the semiparametric information bound for the homoskedastic projection pursuit problem with normal disturbances. It is almost certainly not efficient for the transformation model. The Ruud and Stoker estimators rely on the fact that under suitable conditions the regression of Ln T on x is proportional to β; these are also |n asymptotically normal.

An estimator for the transformation model, applicable also to the proportional hazards model, is the maximum rank correlation method of Han (1987) and Doksum (1985).

Newey (1990) has established the asymptotic efficiency of some kernel and quantile estimators for the censored regression model when error distributions are symmetric. The status of these estimators under some other information conditions remains unresolved. A problem requiring further work is construction of reliable and practical covariance estimators for the semiparametric estimators. An interesting empirical question is whether the censored regression model or the proportional hazards models can be accepted as restrictions on the transformation model (and what are appropriate and practical test statistics)?

Stated Willingness-to-Pay for a Natural Resource

A method for eliciting *Willingness-to-Pay* (WTP) for natural resources is a *referendum contingent valuation* experiment: Survey respondents are asked if they are willing to pay an amount *b*, where *b* is a bid set by experimental design. Let *d* denote a dummy variable that is one for a "Yes"

response, zero otherwise. A sample of n observations are collected on (*b*,*d*) pairs, plus covariates x characterizing the respondent. Suppose WTP is distributed in the population as $w = x\beta - \varepsilon$, where ε has a cumulative distribution function $G(\varepsilon)$ that is independent of x. Then, $\Pr(d=1|x'\beta) = G(x\beta - b)$, or

$$d = G(x\beta - b) + \varepsilon, \tag{32}$$

Suppose β and the function G are unknown. The econometric problem is to estimate β and, if necessary, G, and use these to estimate a measure of location of the distribution of WTP, conditional on x or unconditional. This is an example of a projection-pursuit regression model.

Contingent valuation experiments are controversial because they are very sensitive to psychometric context effects, such as anchoring that leads respondents who are unsure about their preferences to take the offered bid as a cue to the "politically correct" range of values. Some subjects also appear to misrepresent their responses strategically, giving extreme values that they would not practically pay, but which express "protest" positions. These effects make WTP estimates imprecise, and their connection to welfare economics tenuous.

Why do contingent valuation experiments use the referendum elicitation format, rather than a format in which subjects would be asked to give an open-ended WTP response? One answer is that the open-ended format produces a much higher non- response rate, so that the referendum method reduces selection bias caused by non- response. Another is that psychologically the referendum and open-ended methods elicit quite different behaviors. Some argue that the referendum format is closer to the voting mechanisms used elsewhere to make social decisions, and there is a virtue in mimicking this mechanism for social decisions on natural resources.

One issue that enters the contingent valuation experimental design is the location of the bid levels *b*. Alternatives are to randomize *b*, or to choose *b* on a grid with a specified mesh. In practice, coarse meshes have been used, which limits the accuracy of semiparametric estimates. Let $h(b|x)$ be the density from which the bid level *b* is drawn, given x. Since this is chosen by experimental design, it is known to the analyst.

Econometric analysis of referendum WTP data can use the fact that (32) is a binary response model and a single-index model (that is heteroskedastic, but with the heteroskedasticity depending on the index). Then, available methods to estimate β are the Manski (1978) maximum score estimator, the Cosslett (1987) semiparametric maximum likelihood estimator, the Ichimura (1986) estimator that minimizes expected conditional variance, the Horowitz (1992) estimator that is a smoothed version of maximum score, and the Klein-Spady (1993) estimator. The key result for the binomial response model is that under some smoothness conditions, there are root-n consistent estimators β_n for β; i.e., $n^{1/2}(\beta_n - \beta)$ is asymptotically normal. A nonparametric estimator of G can be obtained jointly with the estimation of β, as in the Cosslett procedure, or by conventional kernel methods in a second step after the β estimate is plugged in to form the index; it can be estimated only at a nonparametric rate less than root-n.

One particularly simple estimator for the index parameters β has been proposed for this problem by Lewbel and McFadden (1997): Carry out a least squares regression on the model,

$$(d_i - 1(b_i<0))/h(b_i|x_i) = x_i\beta + \zeta_i. \tag{33}$$

The authors show that the coefficients from this regression are consistent for β, and are asymptotically normal at a $n^{1/2}$ rate. The estimates are not particularly efficient, but their simplicity

makes them an excellent starting point for analysis of model specification and construction of more efficient estimators.

Exercise 5. Prove that the estimator based on (33) is consistent. Apply a law of large numbers to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} x_i'(d_i - 1(b_i<0))/h(b_i|x_i) \rightarrow_p \mathbf{E}_x\mathbf{E}_{b|x}\, x'(G(x\beta-b) - \mathbf{1}(b<0))/h(b|x).$$

Then apply integration by parts to conclude that

$$\mathbf{E}_{b|x}\, x(G(x\beta-b) - \mathbf{1}(b<0))/h(b|x) = x\cdot \int_{-\infty}^{0} (G(x\beta-b) - 1)\cdot db + x\cdot \int_{0}^{+\infty} G(x\beta-b)\cdot db$$

$$= x' \int_{-\infty}^{+\infty} b\cdot G(x\beta-b)\cdot db = x'x\beta.$$

From this conclude that the least squares coefficients converge to $(\mathbf{E}x'x)^{-1}(\mathbf{E}x'x\beta) = \beta$.

The authors also establish that the r-th moment of WTP, conditioned on $x = x_o$, can be estimated consistently at a root-n rate by

$$M_r = (x_o\beta)^r + r \sum_{i=1}^{n} (b_i+(x_o-x_i)\beta)^{r-1}\cdot \frac{d_i - \mathbf{1}((x_i\beta>b_i)}{\sum_{j=1}^{n} h(b_i+(x_j-x_i)\beta|x_j)} \quad . \tag{34}$$

The estimators (33) and (34) are good examples of statistical procedures for a semiparametric problem that are "robust" in the sense that they do not depend on parametric assumptions on the distribution of WTP, and provide an easily computed alternative to use of a kernel-type nonparametric estimator.

6. SIMULATION METHODS AND INDIRECT INFERENCE

Econometric theory has traditionally followed classical statistics in concentrating on problems that yielded analytic solutions. This explains the emphasis on the linear model, and on asymptotic approximations in situations where nonlinearities or other factors make exact sample analysis intractable. Increased computational power, and better understanding of the uses and limitations of numerical analysis, have greatly expanded the ability of econometricians to explore the characteristics of the methods they use under realistic conditions. The idea is straightforward. The economist can write down one or more trial data generation processes, perhaps after an initial round of econometric analysis, and use these data generation processes to generate simulated or virtual samples. If a comparison of a real sample with these virtual samples reveals inconsistencies, this is evidence that the trial data generation process is unrealistic. Conversely, if the econometrician has discovered the true data generation process, then the virtual samples generated from it should not differ systematically from the real sample. Computers and Monte Carlo simulation methods come in at the stages of drawing the virtual samples and comparing the real and virtual samples.

If the kinds of comparisons just described are done casually, without attention to statistical properties, they can mislead the analyst. Traditional calibration exercises in economics and other disciplines often suffer from this deficiency. However, it is possible to develop a statistical theory to support these comparisons, and use this theory to consistently identify the real data generation process, or good approximations to it. In various manifestations, this theory has been developed by

Hendry, Mizon, and Richard under the name *encompassing*, by Gourieroux and Monfort under the name *indirect inference*, and by McFadden under the name *simulation-assisted inference*.

Consider two parametric families of data generation processes, H_f containing models $f(y|x,\alpha)$ for parameter vectors α in a set **A**, and H_g containing models $g(y|x,\beta)$ for parameter vectors β in a set **B**. Both of these families have the same dependent variable y, and are conditioned on the same explanatory variables x. It may be the case that one of these families is nested within the other; this is the situation in classical hypothesis testing where the null hypothesis (say H_g) is a subset of the universe (say H_f), and the true data generation process is a member of H_f and under the null a member of H_g. However, we will now consider more general situations where the two families are not necessarily nested, and the true data generation process may not be in either.

Example. The family H_f is the family of linear models $y = x\gamma + \varepsilon$, where x is a vector of explanatory variables and ε is a normal disturbance with variance σ^2. This family is parameterized by $\alpha' = (\gamma,\sigma^2)$. H_g is the family $y = z\delta + \eta$, where z is a vector of explanatory variables and η is a normal disturbance with variance λ^2, parmeterized by $\beta = (\delta,\lambda^2)$. The vectors x and z may have some variables in common, but in the most general case will each contain some distinct variables so that neither is contained (nested) within the other. $y = x\alpha + \varepsilon$ and the family H_g of linear models $y = z\beta + \eta$, where x and z may have some variables in common, but also contain distinct variables corresponding to alternative theories of the determination of y. The families are said to be *non-nested* when neither can be written as a linearly restricted case of the other.

A proximity measure between densities is the Kullback-Leibler Information Criterion (KLIC),

$$K_{fg}(\alpha,\beta,x) = \int \log(f(y|x,\alpha)/g(y|x,\beta))\cdot f(y|x,\alpha)dy.$$

The KLIC is always non-negative, and is zero only if *f* and *g* coincide. This measure depends on exogenous variables x. We could alternately take its expectation with respect to x,

$$K_{fg}(\alpha,\beta) = \mathbf{E}_x K_{fg}(\alpha,\beta,x)$$

and approximate this expectation by a sample average

$$K_{fgn}(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K_{fg}(\alpha,\beta,x_i).$$

For the model *g*, define the *pseudo-true value* $\beta_f(\alpha)$ to be the $\beta \in B$ that minimizes $K_{fg}(\alpha,\beta)$, and the *conditional pseudo-true value* $\beta_{fn}(\alpha)$ to be the $\beta \in B$ that minimizes $K_{fgn}(\alpha,\beta)$. Then, $g(y|x,\beta_f(\alpha))$ is the data generation process in the *g* family closest to $f(y|x,\alpha)$, and

$$J_f(\alpha,B) \equiv K_{fg}(\alpha,\beta_f(\alpha))$$

is the proximity of *f* and the $g \in H_g$ that is closest to *f*. In an earlier chapter, where $f(y|x,\alpha_o)$ was identified as the true data generation process, we called $g(y|x,\beta_f(\alpha_o))$ the *least misspecified* model in H_g. However, we will now consider more general situations where the *f* family may not contain the true data generation process.

Exercise 6. In the linear model example, Show that

$$\log(f/g) = 0.5\cdot\{\log(\lambda^2/\sigma^2) - (y-x\gamma)^2/\sigma^2 + (y-z\delta)^2/\lambda^2\}$$

$$= 0.5\cdot\{\log(\lambda^2/\sigma^2) - (y-x\gamma)^2(1/\sigma^2 - 1/\lambda^2) + 2(y-x\gamma)(x\gamma-z\delta)/\lambda^2 + (x\gamma-z\delta)^2/\lambda^2\},$$

and hence that $K_{fg}(\alpha) = 0.5\cdot\{\log(\lambda^2/\sigma^2) + \sigma^2/\lambda^2 - 1 + \mathbf{E}(x\gamma-z\delta)^2/\lambda^2\}$. The pseudo-true values in the model H_g are the values $\beta_f(\alpha)$ that minimize $K_{fg}(\alpha)$. Show that the pseudo-true value for δ is $(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\gamma$ and the pseudo-true value for λ^2 is $\sigma^2 + \gamma'\{\mathbf{E}x'x - (\mathbf{E}x'z)(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\}\gamma$. Show that the minimum distance from f to H_g is

$$J_f(\alpha) = 0.5\cdot\log(1 + \gamma'\{\mathbf{E}x'x - (\mathbf{E}x'z)(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\}\gamma/\sigma^2).$$

The distance is zero if z can be written as a linear combination of the variables in x

A model $f(y|x,\alpha)$ is said to *encompass* the family *g* if *f* can account for, or explain, the results obtained with the *g* family. Operationally, this concept says the *g* family will fit similarly the observed sample data and virtual data generated by the model $f(y|x,\alpha)$. If we define

$$b_n = \text{argmax}_\beta \sum_{i=1}^{n} g(y_i|x_i,\beta);$$

to be the maximum likelihood estimate from the family H_g for the observed sample, and $f(y|x,\alpha)$ encompasses H_g, then b_n should converge to the pseudo-true value $\beta_f(\alpha)$. Conversely, if $b_n - \beta_f(\alpha)$ converges to a non-zero limit, $f(y|x,\alpha)$ fails to encompass H_g. This is the same as saying that as judged from the family H_g, samples generated by the model $f(y|x,\alpha)$ look like samples generated by the true data generation process.

Exercise 7. In the linear model example with n observations, write the models H_f and H_g as $y = X\gamma + \varepsilon$ and $y = Z\delta + \eta$ respectively. Show that the maximum likelihood estimates in the family H_g are $\delta_e = (Z'Z)^{-1}Z'y$ and $\lambda_e^2 = y'[I - Z(Z'Z)^{-1}Z']y/n$, and in the H_f family are $\gamma_e = (X'X)^{-1}X'y$ and $\sigma_e^2 = y'[I - X(X'X)^{-1}X']y/n$. Suppose the model $y = X\beta + \varepsilon$ with parameters α is true. Show that the differences of the maximum likelihood estimates in the H_g family and the corresponding pseudo-true values for this family, evaluated at α, converge in probability to zero.

If $f(y|x,\alpha_o)$ is the true data generation process, then by definition it encompasses any other family of models H_g. It is possible for a member of H_g to encompass the true data generation process $f(y|x,\alpha_o)$; this means that the member of *g* can generate data that looks like data drawn from $f(y|x,\alpha_o)$. This could obviously happen if H_g contains one or more models that are observationally equivalent to *f*, but could also occur if H_g contains models that are more "structural" than *f* so that they potentially can explain the same phenomena as *f*, and more.

In the theory of *tests of non-nested hypotheses*, the setup is to have two families of data generation processes, H_f and H_g, which are not nested, with the true data generation process assumed to be in one of the two families. Then, the family containing the true data generation process will encompass the other, but not vice versa (except in the unidentified case where there are models in either family that can mimic the true data generation process). Let a_n be the maximum likelihood estimator of α from the model $f(y|x,\alpha)$. Then $b_n - \beta_{fn}(a_n)$ converges to zero if and only *f* encompasses *g*, and $a_n - \beta_{gn}(b_n)$ converges to zero if and only if g encompases *f*. These observations form the basis for practical test statistics for non-nested hypotheses; see Pesaran (1987) and Gourieroux & Monfort (1994). These ideas also form the basis for an estimation method called *indirect inference*, or in a more general but less focused form, *method of simulated moments*: If the family H_f contains the true

data generation process $f(y|x,\alpha_o)$, then this model encompasses g and one has $b_n - \beta_{fn}(\alpha_n)$ converging to zero if α_n converges to α_o, and with an assumption of identifiability, to a non-zero limit if α_n converges to something other than α_o. Then, choosing α_n to make $b_n - \beta_{fn}(\alpha_n)$ small will under some regularity conditions make these estimators consistent for α_o. The reason to consider these indirect estimates, rather than direct maximum likelihood estimates of α from the model $f(y|x,\alpha)$, is that the true model may be very complex or very difficult to work with computationally. For example, $f(y|x,\alpha)$ may involve a complex structural model, or may involve probabilities that require high-dimensional numerical integration to evaluate. Then, the indirect inference may utilize a simpler family of models H_g that are easier to compute or more "robust". For example, g may be a reduced form model and indirect inference may involve choosing structural parameters so that their transformation to reduced form parameters gives the same values as direct least squares estimation of the reduced form. Or, indirect inference may utilize a select list of moment conditions that you are confident hold in the population. The reason simulation methods enter is that the practical way to calculate $\beta_{fn}(\alpha_n)$ is to use Monte Carlo methods to draw virtual samples from the data generation process $f(y|x,\alpha)$ for various trial α, and select α_n to minimize the distance between the estimator b_n from the observed sample and estimators $b_n(\alpha)$ obtained from a virtual sample from $f(y|x,\alpha)$ by estimating β by maximum likelihood estimation applied to this virtual sample. Because this process can also be interpreted as matching the "moments" b_n from the virtual sample with simulated "moments" $b_n(\alpha)$ from the simulated virtual sample by varying α, it is also called the *method of simulated moments*.

Encompassing is a limited concept when comparing the true data generation process with an alternative, since the true data generation process will encompass any alternative model. However, it becomes more general and more interesting under two circumstances: (1) the true data generation process may fail to lie in either H_f or H_g, or (2) the results from H_f and H_g are based on limited information, such as GMM estimates that rely on specific orthogonality conditions, rather that a full parametric specification of a data generation process. Then, encompassing can be a useful approach to model selection.

We will not attempt to provide any general introduction to simulation and Monte Carlo methods in these notes. However, there a few key concepts that are important enough to introduce at this stage. First consider the problem of drawing a virtual sample from the data generation process $f(y|x,\alpha)$ for a trial value of α. Consider the simplest case when y is one-dimensional. The corresponding CDF $U = F(Y|x,\alpha)$ has a uniform distribution, and a Monte Carlo draw of y for observation i is $y^* = F^{-1}(u_i|x_i,\alpha)$, where u_i is a draw from a uniform distribution. This is a practical method of drawing a realization of a random variable if F^{-1} can be determined analytically or efficiently evaluated numerically. When it is impractical to calculate F^{-1}, one may be able to use *Monte Carlo Markov Chain* (MCMC) methods. A *Metropolis-Hastings* (MH) *sampler* for $f(y|x,\alpha)$ is defined by a conditional density $q(y'|y,x)$ chosen by the analyst and kernel $w(y,y',x) = \text{Min}\{q(y'|y,x), f(y'|x,\alpha)\cdot q(y|y',x)/f(y|x,\alpha)\}$. This kernel is associated with a transition process in which y' is sampled from $q(y'|y,x)$, then the process moves to y' with probability $p(y,y',x)$, and otherwise stays at y, where $p(y,y',x) = \text{Min}\{1, q(y|y',x)\cdot f(y'|x,\alpha)/q(y'|y,x)\cdot f(y|x,\alpha)\}$. A simple choice for $q(y'|y,x)$ is a density $q(y')$ independent of y and x from which it is computationally easy to draw and which has the property that $f(y|x,\alpha)/q(y)$ is never too large, a key determinant of the efficiency of the sampling process. The MH sampler is a generalization of what are called *acceptance/rejection* methods.

The Metropolis-Hastings sampler starts from an arbitrary point, and proceeds recursively. Suppose at step t-1, the draw is y^{t-1} and $f_{t-1} = f(y^{t-1}|x,\alpha))$. Draw y' from the conditional density

$q(\cdot|y^{t-1})$, and define $q_{t+} = q(y'|y^{t-1})$ and $q_{+t} = q(y^{t-1}|y')$, Calculate $\alpha(y^{t-1},y') = \text{Min}\{1, q_{+t}f_t/q_{t+}f_{t-1}\}$. Draw a uniform [0,1] random number ζ. If $\zeta \leq p(y^{t-1},y',x)$, set $y^t = y'$; otherwise, set $y^t = y^{t-1}$. Once it is "burned in", the sequence y^t behaves like a sample drawn from $f(\cdot|x,\alpha)$. Note that the terms in the sequence are not statistically independent. When one needs to form expectations with respect to $f(y|x,\alpha)$, these can be approximated by means over the y^t draws.

In indirect inference or method of simulated moments, one searches iteratively for parameter values that satisfy some criterion, such as minimizing the distance of $b_n - \beta_{fn}(\alpha)$ from zero, using simulation to approximate $\beta_{fn}(\alpha)$. It is important in doing this that the simulated value of $\beta_{fn}(\alpha)$, considered as a function of α, have a property called *stochastic equicontinuity*. Informally, this means that the simulator does not "chatter" as α varies. The way to accomplish this is to keep the Monte Carlo draws that drive the simulation fixed as α changes. For example, when a virtual sample from $f(y|x,\alpha)$ is drawn by the inverse method $y^* = F^{-1}(u|x,\alpha)$, keeping the uniformly distributed draws u fixed as α is varied does the job.

Further reading on simulation methods and indirect inference can be found in McFadden (1989), Gourieroux & Monfort (1994), and Hajivassiliou & Ruud (1994).

7. THE BOOTSTRAP

The idea fundamental to all of statistical inference is the principle that a statistical sample forms an *analogy* to the target population, and to estimate the results of an operation on the target population, one can complete the analogy by carrying out the same operation on the statistical sample. Thus, the sample mean is analogous to the population mean, and hence has decent statistical properties as an estimate of the population mean. Manski (1994) shows how this principle can guide the construction of estimators.

Extending the analogy principle, if one is interested in the relationship between a target population and a given sample drawn from this population, one could form an analogy by starting from the given sample, drawing subsamples from it, and forming analogous relationships between the original sample and the subsamples. When the subsamples are drawn with replacement and are the same size as the original sample, this is called the *bootstrap*.

To illustrate the operation of the bootstrap, suppose you have an estimate a_n of the parameter in a data generation process $f(y|x,\alpha)$, obtained from a sample of size n from the target population. You would like to know the variance of the estimator a_n. Note that this is a property of the relationship between the population and the sample that could in principle be determined by drawing repeated samples from the population, and estimating the variance of a_n from the repeated samples. The bootstrap idea is to start from the observed sample, draw repeated subsamples from it (with replacement), and complete the analogy by forming the estimator a^* for each subsample, and computing the sample variance of these estimators. The bootstrap process is computationally intensive, because it involves the subsampling process and the computation of a^*, repeated many times. Under very general regularity conditions, the analogy principle applies and the estimate of the variance of a_n formed in this way will have good statistical properties. Specifically, the bootstrap estimate of the variance of a_n will have the same properties as the first-order asymptotic approximation to the variance, without the effort of determining analytically and computing the asymptotic approximation. Further, the bootstrap estimator will under some conditions pick up higher order effects, so that it is a better finite sample approximation that the first-order asymptotic approximation. In particular, if the expression being studied has a limiting distribution that is independent of the parameters of the problem, as for example when one is interested in the finite

sample distribution of the ratio of a parameter estimate to its standard error which has a limiting T-distribution, the bootstrap will be more accurate for finite samples that the first-order asymptotic approximation. A statistic with the last property is called *pivotal*.

Bootstrap methods can often be used to estimate the distribution of statistics, for purposes of estimating moments or critical levels, in situations where asymptotic analysis is intractable or tedious. The bootstrap is itself one member of a broad class of techniques called *resampling methods*. There are various pitfalls to be avoided in application of resampling methods, and a variety of shortcuts and variants that can speed calculation or make them more accurate. For further reading, see Efron & Tibshirani (1993), Hall (1994), and Horowitz (1999).

REFERENCES

Chamberlain, G. (1986) "Asymptotic Efficiency in Semiparametric Models with Censoring" *J. of Econometrics* **29**, 189-218.

Chamberlain, G. (1992) "Efficiency Bounds for Nonparametric Regression" *Econometrica* **60**, 567-96.

Cosslett, S. (1987) "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models" *Econometrica* **55**, 559-585.

Cox, D. (1972) "Regression Models and Life Tables" *Journal of the Royal Statistical Society* **B, 34**, 187-220.

Delgado, M. and P. Robinson (1992) "Nonparametric and Semiparametric Methods for Economic Research" *J. of Economic Surveys* **6**, 201-49.

Doksum, K. (1985) "An Extension of Partial Likelihood Methods for Proportional Hazard Models to General Transformation Models" U. of California, Berkeley working paper.

Duncan, G. (1986) "A Semiparametric censored Regression Estimator" *Journal of Econometrics* **29**, 5-34.

Gourieroux, C. and A, Monfort (1994) "Testing Non-Nested Hypotheses," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2585- 2640.

Hajivassiliou, V. and P. Ruud (1994) "Classical Estimation Methods for LDV Models using Simulation," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2384-2443.

Hall, P. (1994) "Methodology and Theory for the Bootstrap," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2342- 2383.

Han, A. (1987) "Nonparametric Analysis of Generalized Regression Models: The Maximum Rank Correlation Estimator", *J. of Econometrics* **35**, 303-16.

Hardle, W. and O. Linton (1994) "Applied Nonparametric Methods," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2297- 2341.

Heckman, J. and B. Singer (1984) "A method for minimizing the impact of distributional assumptions in econometric models for duration data" *Econometrica* **52**, 271-320.

Horowitz, J. (1986) "A Distribution-Free Least Squares Method for Censored Linear Regression Models" *J. of Econometrics* **29**, 59-84.

Horowitz, J. (1989) "Semiparametric M-Estimation of Censored Linear Regression Models" in G. Rhodes and T. Fomby (eds) *Nonparametric and Robust Inference*, Advances in Econometrics **7**, 45-83.

Horowitz, J. (1992) "A Smoothed Maximum Score Estimator for the Binary Response Model" *Econometrica* **60**, 505-31.

Horowitz, J. and G. Newmann (1987) "Semiparametric Estimation of Employment Duration Models" *Econometric Reviews* **6**, 5-40.

Horowitz, J. and G. Newmann (1989) "Computational and Statistical Efficiency of Semiparametric GLS Estimators" *Econometric Reviews* **8**, 223-25.

Ichimura, H. (1986) "Estimation of Single Index Models" PhD Dissertation, MIT.

Kalbfleisch, J. and R. Prentice (1980) *The Stochastic Analysis of Failure Time Data*, New York: Wiley.

Kaplan, E. and P. Meier (1958) "Nonparametric estimation from Incomplete Observations" *J. American Statistical Association* **53**, 487-491.

Klein, R. and R. Spady (1993) "An Efficient Semiparametric Estimator for Binary Response Models" *Econometrica* **61**, 387-422.

Lancaster, T. (1979) "Econometric methods for the duration of unemployment" *Econometrica* **47**, 141-165.

Manski, C. (1978) "Maximum Score Estimation of the Stochastic Utility Model of Choice" *Journal of Econometrics* **3**, 205-228.
Manski, C. (1994) "Analog Estimation of Econometric Models," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2560- 2584.
McFadden, D. (1989) "A Method of Simulated Moments for Estimation of Multinomial Probit Models without Numerical Integration," *Econometrica,*.
Meyer, B. (1987) "Unemployment Insurance and Unemployment Spells" *Econometrica* **58**, 757-82.
Newey, W. (1990) "Semiparametric Efficiency Bounds" *J. of Applied Econometrics* **5**, 99- 135.
Newey, W. and D. McFadden (1994) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2113-2247.
Pesaran, M. (1987) "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power," *Econometric Theory* 3, 69-79.
Powell, J. (1986) "Censored Regression Quantiles" *J. of Econometrics* **29**, 143-155.
Powell, J., J. Stock, and T. Stoker (1989) "Semiparametric estimation of weighted average derivatives" *Econometrica* **57**, 1403-30.
Powell, J. (1994) "Semiparametric Econometric Methods" *Handbook of Econometrics*, Vol. 4, R. Engle and D. McFadden (eds.), North Holland.
Ritov, Y. (1985) "Efficient and unbiased estimation in nonparametric linear regression with censored data" U. of California, Berkeley working paper.
Robinson, P. (1986) "Semiparametric Econometrics: A Survey" London School of Economics working paper.
Ruud, P. (1986) "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution" *J. of Econometrics* **29**, 157-187.
Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley: New York.
Silverman, B. (1986) *Density Estimation*. Chapman and Hall: London.
Stoker, T. (1986) "Consistent Estimation of Scaled Coefficients" *Econometrica* **54**, 1461-1481.
Stone, C. (1977) *Annals of Statistics*.