The Craft of Labormetrics

Daniel S. Hamermesh

Industrial and Labor Relations Review, Vol. 53, No. 3 (Apr., 2000), 363-380.

Stable URL:
http://links.jstor.org/sici?sici=0019-7939%28200004%2953%3A3%3C363%3ATCOL%3E2.0.CO%3B2-1

Industrial and Labor Relations Review is currently published by Cornell University, School of Industrial & Labor Relations.

THE CRAFT OF LABORMETRICS

DANIEL S. HAMERMESH*

Using a wide array of examples from the literature and from original estimates, the author examines the pitfalls that make good empirical research in labor economics at least as much craft as statistical technique. Among the subjects discussed are the appropriateness and cleanliness of data; problems of extreme observations; the validity of attempts to produce exogeneity using instrumental variables, "natural experiments," and structural models; and the treatment of selectivity and unobservable individual effects. The author stresses the importance of testing empirical results to ensure that they make sense, and of presenting them clearly and economically.

Instruction on a massive array of topics covering nearly the entire panoply of econometric technique is widely available in textbooks and surveys. There is, however, more to good empirical work than technique: there is the craft of knowing what makes economic sense and of emphasizing those clever ideas that will make the largest substantive contribution to one's work. The discussion here makes no claim to technical originality. Instead, its purpose is to provide a feel for what might be important, what to do, and, perhaps most of all, what not to do in empirical work. The examples are almost exclusively from the literature of labor economics, but they also illustrate difficulties in empirical work in other sub-specialties of economics and in the statistical analysis of labor issues using approaches of other disciplines.

Several underlying themes connect the wide-ranging discussion that follows. Empirical research in labor economics ideally focuses on the interpretation of behavior. Discovering the facts—cross-section and time-series patterns of wages and time use and their correlates, how various policies affect labor-market outcomes, and so on—is crucial. But labor-market outcomes change remarkably rapidly; and the Sgt. Friday approach to studying labor markets ("Just the facts, Ma'am") condemns us to endlessly repeated reportage. The best labormetric research documents outcomes

*The author is Edward Everett Hale Centennial Professor of Economics, University of Texas at Austin, and Research Associate, National Bureau of Economic Research. He thanks Julian Betts, Jeff Biddle, Charles Brown, Barry Hirsch, Jacob Klerman, Steve Pischke, Daniel Slesnick, and participants in seminars at several universities for comments and suggestions on earlier drafts, and Steven G. Allen for comments and for the term "labormetrics."

The data sets underlying the estimates in Tables 1 and 2 are available upon request from the author at the Department of Economics, University of Texas at Austin, Austin, TX 78712-1173.

Industrial and Labor Relations Review, Vol. 53, No. 3 (April 2000).
0019-7939/00/5303 $01.00

and uses economic theory to infer the behavior that generated them, allowing us to understand why the outcomes change and to predict their paths.

The payoff to cleverness in labormetrics is huge. The biggest rewards in our field have gone to those who have developed new approaches that solve old, often ill-perceived problems in inferring behavior from data. Their innovations diffuse rapidly among other applied economists, often too rapidly, for they are adopted because they are available, not because they are necessarily appropriate. The availability of a new technique is not its own justification; and those using it should ask themselves whether the technique enhances or clouds the attempt to infer behavior. Cleverness in labormetrics at least as often involves using standard techniques in novel ways to increase our understanding of some phenomenon.

Even before using sophisticated techniques, it is crucial not to misapply what have become fairly standard techniques. In what follows I illustrate many of my admonitions about such misapplications and misinterpretations using examples from the recent literature of empirical labor economics and new calculations based on a variety of sets of data. I draw an embarrassingly large number of these examples from my own research, not because that research is particularly important, but because I am most familiar with the data sets that underlie it.

Data Cleanliness ahead of Econometric Godliness

Before we worry about clever technique we must obtain data on which to exercise our technique, generate estimates of effects, and infer the behavior that caused them. Too often we accept the data that are given to us as representing the economic concept that we seek to include in our estimates. We need to ask whether we have found the best *available* data for the purpose and, more important, whether those data offer any hope of representing the concept. Research on labor supply, for example, has difficulties measuring the unearned income that is essential to identifying income effects; in studying labor demand, the wage rates typically used are very far from the full marginal cost of labor. The issue is whether we can match empirical proxies to our theoretical constructs. If not, collecting our own data, or at least assembling data sets from existing records, is the only alternative to abandoning empirical work. While data collection is often expensive and time-consuming, the payoffs can be very large, as the new literature matching employer and employee data has begun to show (for example, Abowd et al. 1999).

Another way we reduce the chance of answering our research question is by restricting our samples so that they cannot answer the question. Pick up any recent issue of a good labor journal, or examine the labormetric articles in the top general journals, and consider whether they meet this criterion. For example, in a clever paper Baker (1997) examined the time-series structure of men's earnings, a crucial question for inferring the nature of earnings inequality. Yet by excluding from his 20-year sample all men who were not household heads or who did not work for pay *in each year,* he selected his sample in a highly nonrandom fashion. Shin (1997) was interested in the relative importance of micro and macro shocks to employment. While he laudably based the study on firm-level data, the restriction was to manufacturing (which in the United States accounted for 15% of total employment and 17% of GDP in 1996). This "manucentrism" pervades the much-emulated research on idiosyncratic employment changes (for example, Davis and Haltiwanger 1992). While we can obtain useful knowledge about this relatively small sector, the approach has given idiosyncratic impressions about the relative importance of different sources of job growth and their differing cyclical variation. Just because the data are readily available does not mean that they will answer the research question we are studying.

Once we are satisfied that the data *may* provide answers, the main issue concerns

measurement error of the sort:

$$(1) \quad X_{it} = X^*_{it} + \theta_{it}, \; i = 1,\ldots, N; \; t = 1,\ldots, T,$$

where X^*_{it} is the true measure, X_{it} is what we observe, and θ_{it} is the error. (Throughout, I will generalize and write both i and t subscripts, implying that we may have observations on units at a point in time and on some or all of them over time.) Here one cannot assume that $E(\theta_{it}) = 0$. That would not be a problem if this expectation were constant; but it need not be. There are cases where $E(\theta_{it})$ is changing, perhaps even trending in t; there are other cases where it is not independent of which i we examine.

Consider the substantive problem that has probably occupied more labor economists' efforts than any other, the measurement and explanation of hours of work. We wish to examine the amount of some time interval (day, week, year, lifetime) devoted to market production. Nearly all our research is based on retrospective data, in which respondents to a survey are asked to describe their activities in some past period. Such data consistently overstate work time in comparison with measures of market hours from diaries that record actual time use (Juster and Stafford 1991). Worse still, the overstatements differ with observed characteristics (differ predictably across i). Even more troublesome, the differences across i suggest that the $E(\theta_{it})$ may also be trending, although the regrettable absence of repeated cross-sections of time diaries in most countries makes direct inference impossible. All of these problems mean that our inferences about trends in hours of market work and about its changing responsiveness to prices and nonwage incomes contain nonrandom errors whose direction may be known, but whose size is not.

Slightly longer-term problems of retrospection about time use are underscored when we try to analyze spells of unemployment (Akerlof and Yellen 1985). It seems clear that the quality of recollections of unemployment deteriorates as they recede into the past, and this deterioration is systematically related to the amount of unemployment experienced and to demographic characteristics. The problem becomes even more severe with discrete events recollected many years later. The Displaced Worker Surveys, biennial supplements to the monthly U.S. Current Population Surveys, ask workers whether they lost a job in the past five years, an event that respondents apparently remember nonrandomly depending on its temporal distance and severity (Evans and Leighton 1995). The nonrandomness implies not only errors, but also biases to estimates of the average effects of displacement on wages and other outcomes, to the extent that the relationships are nonlinear.

Most econometrics texts treat measurement error as God-given. It is not: data may be dirty, but in many cases the dirt is more like mud than Original Sin. In a study of the spending behavior in 1972–73 of 655 American households with some income from unemployment insurance (UI), I estimated

$$(2) \quad C_t = 0.55\,UI_t + 0.75\,YD_t + \varepsilon_t,$$

where the constant was insignificantly different from 0, and YD is the household's other income (Hamermesh 1982). An unpublished estimate of this model yielded a propensity to spend out of UI benefits of 0.15! The reason was simple, albeit not detected for some time: in one of the households, annual UI benefits were coded on the data tape as $22,184. This outlier clearly resulted from an extra first digit.[1] The moral is that one must check the descriptive statistics, especially the minima and maxima, of all series prior to any estimation.

In some cases the mud is purposely smeared on the data to preserve the respondents' confidentiality. Topcoding responses to questions about earnings or incomes creates measurement errors whose presence generates problems for standard regression techniques. In other cases we

[1] In most states at that time, unemployed workers could receive up to 26 weeks of benefits, and $84 was a typical weekly benefit amount.

create subtle but important measurement errors ourselves, as when we use as dependent variables measures of wage rates constructed by dividing annual or weekly earnings by hours (Borjas 1980). In still others, the errors are not subtle at all, and when they are in the dependent variable, even a few such errors can be catastrophic. One author sought to explain monthly coital frequency using a longitudinal sample, but mistakenly assigned 4 of over 2000 observations a value of 88 (in a data set with the missing-data indicator 99). As her critics pointed out (Kahn and Udry 1986), these few erroneous outliers generated a sign reversal on a crucial independent variable.

Other measurement errors arise from careless survey responses. While these are neither conceptual nor induced, their effects on our estimates can be inferred, and in some cases the errors can be reduced. One approach, which has been used in several different contexts, is to design the survey to obtain several measures of the same indicator of some status or outcome. Elliott and Sandy (1997), for example, estimated hedonic wage equations using employers' and employees' assessments of job risks, and Card (1996) used employers' and employees' responses in a BLS CPS Supplement to infer a better measure of union status. To the extent that the parties' responses contain uncorrelated measurement errors, this approach enables the researcher to reduce the noise-signal ratio in the variable of interest. Another approach (Philipson 1997) is to model and perhaps alter the quality of the responses of the underlying agents.

All of these considerations should alert us that measurement error is not merely something to which we need only bow before we proceed with estimation. Physicians bury their medical mistakes in the ground; we bury mistakes in our data under a welter of econometric technique. Neither group is honest about the extent of deaths that are caused; but at least physicians can usually tell when their colleagues' patients are dead.

Away from the Mean

Because of its assumption that the sum of squared errors is minimized to derive parameter estimates, the Gauss-Markov theorem implies that least-squares regression weights outliers very heavily. The question is whether the outliers are providing us with information or are merely the result of measurement errors. If the former, least squares is a sensible (not merely convenient) way to infer effects at sample averages; if the latter, it is misleading, and we would get better estimates of the true relationship if we chose a technique that avoided weighting outliers so heavily. A variety of estimators of the α in

$$Y_{it} = \alpha X_{it} + \varepsilon_{it} \quad (3)$$

are possible (Manski 1991); minimizing the sum of the $|\varepsilon_i|$ is one common approach (equivalent to estimating a regression line so that the median residual is zero). While one may have beliefs about whether the theory applies at the means or the medians, the more typical concern is how much information is conveyed by extreme observations.

Does this concern about outliers matter in practice? Clearly it will matter most when one has reason to believe that there are true outliers in the data (*not* outliers reflecting dirty data that we created or failed to clean), perhaps as diagnosed by the descriptive statistics on the dependent variable. To examine the practical importance of this concern, I take four data sets as examples, three from my previous published research. Part I of Table 1 reports LS (least squares) and LAD (least absolute deviations) estimates of a bivariate regression relating the average salary of economics full professors in 17 major public universities in 1996–97 to the average rankings of the departments' quality.[2] Even in this very

[2]This is an average of the 1992 National Research Council rating, which is probably in part retrospective, and the rankings of the departments based on pages published during 1990–94 in leading economics journals. A lower number denotes a higher ranking.

Table 1. LS and LAD Estimates from Four Data Sets.

Example	*Dep. Var.: (Mean; s.d.) (Min; Max)*	*Coefficient (Std. Error)* LS	LAD
I. Salaries of full professors in public-university economics departments, effect of reputational ranking (1 is highest), N = 17:	96274; 8949 80512; 115021	−662 (187)	−626 (315)
II. Ln(Hourly Earnings), Quality of Employment Survey 1977, 700 men, effect relative to average of:	1.83; .48 .05; 3.24		
Below-average looks		−.164 (.046)	−.171 (.055)
Above-average looks		.016 (.033)	.010 (.040)
III. Weekly minutes sleeping, resting and napping, Time Use Survey, 1975–76, effect of minutes of work, N = 706:	3383; 499 1335; 6110	−.199 (.020)	−.182 (.028)
IV. Ln(1+Estate), wealthy decedents, Connecticut, 1939–76, N=149, effect of:	12.23; 1.88 7.15; 17.34		
Expected Longevity		.160 (.033)	.154 (.031)
Unexpected Years of Life		−.015 (.020)	−.006 (.018)

small sample the estimates differ remarkably little between LS and LAD, perhaps because the distribution around the mean is fairly tight and more or less symmetric. Part II of Table 1 reproduces from Hamermesh and Biddle (1994) the estimated effects of physical appearance, with indicator variables denoting whether the respondent is in the bottom 15% of physical appearance or the top 30%, on the logarithm of men's earnings adjusted for a wide array of standard variables. Here too, but with much larger samples, the two approaches yield almost identical estimates of the two crucial parameters, even though the range of Y is much larger relative to its standard deviation than in the first example.

The third example (from Biddle and Hamermesh 1990, Table 3, column 1) uses LS to explain weekly sleep time by a variety of demographic variables as well as minutes of market work. A comparison of the third and final columns in the third part of the table shows that the LAD estimates differ little from the LS estimates. The last part of Table 1 examines the effects of subjective life expectancy and unexpected years of life on the size of bequests (data from Hamermesh and Menchik 1987). The expected positive coefficient of the former (more time to accumulate to satisfy a bequest motive) is almost identical with the two estimators, and the expected negative, but statistically insignificant coefficient of the latter is hardly altered by LAD.

These results suggest that the extreme weight that LS attaches to outliers does not greatly affect the parameter estimates in an admittedly nonrandom sample of typical data sets of different sizes and substantially different properties that seems typical of the kinds labormetricians use. The distributions of most monetary outcomes (wages, incomes, and wealth) can often be transformed in reasonable ways that make them conform to the criteria for using LS (see Koenker and Bassett 1978). No doubt, however, there are data sets for which this choice of technique may matter, either because the data cannot readily be trans-

formed to satisfy the assumptions of the Gauss-Markov theorem, or because even with transformations there are important outliers. Since statistical packages have made it increasingly easy to examine the sensitivity of one's estimates to the least-squares assumptions, checking out this possibility is a low-cost test of robustness. In addition, examining the effects of influential observations, perhaps by trimming the sample, is another good way to handle this potential problem.[3]

Without some loss function based in theory or policy concerns, the choice among LS, LAD, and other estimators has no basis in the economic behavior one is modeling, being purely a matter of one's beliefs about the informational content of outliers in the data. A related issue, whether underlying behavior differs at different points of the distribution of Y, is economic. A typical case involves estimating wage equations that test whether an institution or policy affects wages differently over their distribution. For example, where in the distribution of wages are the wage gains from trade-union membership or from public-sector employment greatest, adjusted for workers' skills (Card 1996; Mueller 1998)?

In other cases the theory implies systematic differences in the α, making the use of quantile regression an essential tool in hypothesis testing. For example, one might believe that the labor of recent immigrants is increasingly easily substituted for that of natives as we move down the level of skill, both measurable and unmeasurable, with both presumably reflected in native workers' wages (Reimers 1998). If we let Y be natives' wages and let one component of X be the fraction of recent immigrants in an area, we should expect $\hat{\alpha}_q < \hat{\alpha}_{q+1}$, where q is some quantile of the wage distribution. Whether we are merely testing for variations in the impact of some X or have some behavioral hypothesis that implies that the impact varies, there is no reason to assume that the α are constant. As with LS, with more observations one can estimate more precisely the relationship at more quantiles. With larger data sets and low-cost methods of estimating quantile regressions, examining the constancy of α over the distribution of Y makes sense.

[3]A very careful comparison using all these approaches is provided by Anderson and Meyer (2000).

Rerum Cognoscere Causas

"To understand the causes of things," the motto of the London School of Economics, should be the central goal of labormetrics. Since the development of modern econometrics we have been concerned with possible violations of the assumption of the Gauss-Markov theorem that $E(X\varepsilon) = 0$. From the 1940s through the 1960s that concern led to intense concentration on developing the simultaneous-equation methods that are familiar now to all first-year econometrics students. One might interpret the growth of time-series econometrics in the 1970s and 1980s, particularly the attention to inferring causality in time-series models, as resulting from the same concern.

A few examples can show that these recent concerns are sensible. One of the staples of labormetrics is

$$\log(W_i) = \alpha S_i + \beta X_i + \varepsilon_i, \tag{4}$$

where W is worker i's wage rate or earnings, S is her education, and X is a (large) vector of her characteristics. One wishes to identify $\hat{\alpha}$ as the rate of return to marginal investments in schooling by a randomly chosen member of the population. The agents' behavior confounds the estimate in a variety of ways, including nonrandomness induced by differential access to funds for additional schooling, by intergenerational and other transmission of tastes for additional schooling and of unmeasured productivity-augmenting factors, and others. The trick is to purge the estimates of these factors so as to infer how additional schooling would raise earnings (and presumably productivity). Another example is the problem of inferring the supply of labor L from equations such as

$$L_{it} = \alpha W_{it} + \beta X_{it} + \varepsilon_{it}, \tag{5}$$

where X here is a vector of variables describing other factors that shock the labor supply of workers i. The difficulty is the standard one of identifying shifts in demand so that we can treat $\hat{\alpha}$ as the slope of the labor-supply curve.

One approach to these problems takes off directly from the studies of simultaneous systems and their identification, but goes beyond that by deriving the estimating equations explicitly from the agents' utility maximization and is often interested in the parameters of more than one structural relation. The appeal of the structuralist approach (see, for example, Keane and Wolpin 1997) is that the estimating equations can be linked to underlying behavioral parameters. Ideally one generates estimates of the deep structural parameters that underlie the agents' maximization and that generate the higher-level outcomes we observe. To the extent that identification is achieved, this approach has the virtue of providing a much tighter link between economic theory and labormetric estimation than other approaches do.

Unfortunately, in many cases the identification restrictions required to interpret the estimates as being structural lack credibility, sometimes because the data sets used prevent the construction of the appropriate identifiers, but more usually because the identifying restrictions require heroic assumptions about heterogeneity and nonlinearity to obtain structural parameters that are consistent with the theory. In any case, in the past decade relatively few studies have relied on a structural approach. Instead, the 1990s saw labor economists playing a pioneering role in research addressing endogeneity and causality, and a huge amount of attention was devoted to developing new approaches to account for causality in labormetric relationships. These new directions have provided substantial gains over structural approaches in terms of addressing causality issues, at the cost of weakening the link between the economic theory underlying the agents' behavior and the estimates that are produced.

The new line of proposed solutions to the simultaneity problem in labor economics has involved a two-pronged approach: examinination of "natural experiments," and discovery of clever instrumental variables. A natural experiment consists of some event occurring between times $t = 1$ and $t = 2$ that shifts the RHS variable of interest. By calculating $Y_{i2} - Y_{i1}$, where Y is the dependent variable of interest and i are observations where the shock occurred, one can infer the impact of the shock to the variable that we wish to treat as exogenous, *provided* nothing else shifted Y on the time interval [1,2]. The common way of conditioning on other determinants of Y is to identify a set of observations j in which the shock did not occur, but in which the X either can also be measured or can be assumed to have changed identically and to have had the same marginal effects on Y as in observations i. That allows the double-difference $\Delta^2 = [Y_{i2} - Y_{i1}] - [Y_{j2} - Y_{j1}]$ to be calculated as an unbiased estimate of the effect of the exogenous shift in the RHS variable.[4]

The most obvious difficulty with this approach is that Δ^2 alone may not control for the changes in Y_i that occurred during this interval. One solution proposes finding additional observations i' similar to i but unaffected by the "experiment," and other j' similar to j, and calculating the triple-difference

$$\Delta^3 = \{[Y_{i2} - Y_{i1}] - [Y_{j2} - Y_{j1}]\} - \{[Y_{i'2} - Y_{i'1}] - [Y_{j'2} - Y_{j'1}]\}. \quad (6)$$

In one example, Gruber (1994) inferred the impact of state-mandated coverage of maternity benefits on wages of young women, using as "experimental" units those states that passed legislation before the federal mandate of 1978, as "controls" matched other states, and as i' and j' single

[4]In its simplest form this approach is, of course, the same as the intercity method of inferring the effect of unions on relative wages that was used by Gregg Lewis's students in a number of masters and doctoral dissertations completed in the 1940s and 1950s (discussed by Lewis 1963).

men and older workers. In another example, Hamermesh and Trejo (2000) inferred the impact of a rise in the penalty on overtime work by identifying a change in California in 1980 that extended to male workers a daily penalty that had applied to women only, letting i be men and j be women in California, and letting the (') be outside California.

Differencing ignores the strong possibility that other variables affecting the Y have changed differentially over time across areas and groups and indeed makes no attempt to theorize about any determinants of Y other than the shock. Given the difficulty of claiming that groups j, i', and j' are otherwise identical to group i, one should replace the Y_{kt} in (6), or in the calculation of Δ^2, by $E(Y_{kt} \mid X_{kt})$, conditioning on as many *theoretically based* components of X in i and j (and i' and j') as are available in the data. Double- and triple-differencing without additional conditioning variables should be viewed as the equivalent of calculating descriptive statistics prior to actual estimation, or as a last resort when one cannot obtain information on the components of X.[5]

An important issue is whether the observations Y_{i1} and Y_{i2} measure the outcome before and after the shock occurred. Is $t = 1$ sufficiently distant from the shock to give us confidence that agents had not yet begun adjusting to an event that may have been partly expected? Obversely, if we are interested in long-run effects of the change (which is what most theories discuss), does $t = 2$ sufficiently post-date the shock to give us confidence that agents have made all the adjustments to the shock? Answering these questions requires the researcher to think about agents' behavior. The difficulty with lengthening the real time between $t = 1$ and $t = 2$ is that when we do so, other factors that are unaccounted for but that affect Δ^2 (or Δ^3) are increasingly likely to have changed.

In Hamermesh and Trejo (2000) there is little problem with the choice of $t = 2$; but $t = 1$ was 1973, by which time agents may have begun adjusting their input demands in reaction to the already widespread discussion of extending the overtime penalty (which occurred in 1980). Both problems may be important in Gruber's (1994) study, as wages are unlikely to have adjusted fully in 18 months in the experimental states, and the passage of state mandates may have been expected within two years of the legislation. Observing at $t = 1$ and $t = 2$ too close to the event biases the estimated impact toward zero. A good way to circumvent problems associated with the choices of $t = 1$ and $t = 2$ is to use as many values for each as the data will allow.

The most difficult issue is whether the change that is supposed to identify the effect of the RHS variable of interest is truly exogenous. One must be able to argue that its timing and size are independent of the past history of Y_i (which is likely to be correlated with Y_{i1}); otherwise, Y_{i1} is correlated with the magnitude of the shock, and the approach has not solved the exogeneity problem. The best claim for exogeneity can be made for "acts of God" or acts originating outside the economy being evaluated and unaffected by events in it. Studies of the impact of migration (to the United States from Cuba—Card 1990; and to France from Algeria—Hunt 1992) are examples of fairly convincing claims of exogeneity. Treating legal changes as exogenous, on the other hand, is much less convincing, except where such changes are imposed on many subunits by a higher level of government, as in the second part of Gruber's (1994) study.

The recently revived instrumental-variables approach relies instead on finding clever instruments, ones that are correlated with the shock variable of interest (S in (4), W in (5)) but uncorrelated with the

[5]Conditioning on a vector of X variables is *not* an admission that one has failed to select the proper control group (despite Meyer 1995). We are never studying laboratory experiments, so other things may very well change differentially. At the very worst, if one has chosen a control group perfectly and has reproduced laboratory conditions—a highly unlikely event—conditioning on the X will be an empty exercise.

error term. Much of the focus has been on measuring the returns to schooling (essentially α in (4)), using as an instrument the date of birth—because compulsory schooling requirements have cut-off dates that impose exogenous and discrete constraints on schooling decisions (Angrist and Krueger 1991)—or siblings' sex composition—because it affects women's schooling and may not be related to earnings except through schooling (Butcher and Case 1994).

Clever instruments have also been devised to circumvent endogeneity in estimating labor supply equations (5). For example, to infer the impact of fertility (a component of the vector X in (5)) on mothers' labor supply, Iacovou (1996), noting that third births are more likely among couples whose first two children are of the same sex, instrumented fertility (a third child) by the exogenous sex composition of the first two children. In inferring the supply elasticity of effort by ballpark vendors, Oettinger (1999) instrumented their pay by various known exogenous factors that shift attendance at ballgames.

These innovations have substantial appeal; but, as with the earlier literature on instrumental variables, whether we can safely use an instrument to infer the impact of a hypothetical exogenous shock on the outcome rests in part on whether the instrument explains much of the variation in the supposed endogenous variables. Bound et al. (1995) discussed this in the context of using birth date as an instrument for education, and they illustrated clearly the problems that arise when the instrument is only weakly correlated with the variable for which it is instrumenting. Unrelated but equally crucial is the recognition that the behavioral effect of interest in the population as a whole may differ from the effects across differing values of the instrument.[6] Finally, an instrument's validity also rests on whether the behavior adapts to it in such a way as to render the instrument's exogeneity suspect.

The new experimental-instrumental literature is more precise than its predecessors in thinking about the conditions for identification, but its proponents are often too quick to assume that the chosen instrument is exogenous and generates a consistent estimate of the population parameter. In the case of date of birth, for example, parents choose whether to "hold back" from starting in school a child whose birthday barely makes the starting deadline. A different mix of offspring leads parents to change the amount of pre-school time they spend with daughters, affecting subsequent wages and rendering questionable the instrument's exogeneity to the schooling decision. In the case of using prior sex composition to instrument fertility, women who already have two children are not representative of all married women, or even of all mothers, in their labor-supply responses. One must be able to argue that the instrument is beyond the decision-makers' control and that it describes behavior that is randomly distributed in the population one wishes to describe.

Selected Unobservables and Not-So-Fixed Effects

Since the late 1970s two economic/technical issues have captivated labormetricians: sample selectivity and unit-specific effects. Both deal in some way with problems generated by behavioral effects in our main relations (equation 3) that produce subtle biases in the estimates of the α on the X variable(s) of interest. Selectivity problems arise because there may be unobserved correlates of Y that bias estimates of α by determining whether a data point is included in the sample. Problems with individual effects result from our belief that a bias is induced because unobservables are correlated with both Y and X. These are powerful ideas that have led to ingenious solutions. By the early 1990s canned statistical packages enabled labormetricians to apply these solutions to their own research problems at very low cost.

[6]This issue and related ones are discussed in Angrist et al. (1996) and in the comments thereon.

Consider first the selectivity issue. The classic selectivity problem (implied by Gronau [1974], analyzed and solved by Heckman [1976]) consists of the model

$$(7) \quad Y1_{it} = \alpha X_{it} + \beta Y2_{it} + \varepsilon_{it}, \text{ observed if:}$$

$$(8) \quad Y2_{it} \geq \gamma Z_{it} + v_{it},$$

where the Yk are endogenous variables, the α and β are parameters, and the ε and v are error terms. The example that generated the initial interest in this problem, the unobservability of the wages ($Y2$) of nonparticipants in the labor force who are thus excluded from estimates of the effect of wages on hours of work ($Y1$) in (7), had a very clear economic interpretation, with the variables in Z representing the value of time in the home, and those in X representing the nonwage variables that shift labor supply.

The solution (the so-called Heckman correction) continues to be applied frequently.[7] Aside from the ever-present possibility of the nonnormality of the disturbances in (7) and (8) (discussed by Newey and Powell 1993), these applications have a severe problem that should stand as a warning to those tempted by the presence of an easily available computer routine. Unless there are several observable variables that can be rationalized as belonging in Z but not in X and that vary independently of X, identification of the relationships is achieved only through the nonlinearity in the inverse Mills' ratio that is included in the estimation of (7). Researchers using this correction should present a good theoretical justification for excluding the Z from (7) and the X from (8), and should either present estimates of (7) with and without this correction or report in a footnote that the other approach yielded different (or similar) estimates of the crucial parameters in α and β.

A finding that the selectivity term is statistically insignificant in (7) may be evidence that the model is underidentified, not that selectivity is unimportant. Obversely, even if the correction "matters" statistically, one must have an *economic* theory justifying (7) and thus the inclusion of a selectivity correction. Some uses of the correction rest on the mere fact that observations are excluded; others rely on the faith that the user's problem is the same as the original motivation for the technique, even though the new problem may lack the sound microtheoretic basis of the original problem. Without an explicit justification for the auxiliary equation, it is not clear that the correction will improve estimates of α and β.

The typical individual-effects model specifies a time-invariant unobservable φ_i that affects Y_{it}:

$$(9) \quad Y_{it} = \alpha X_{it} + \varphi_i + \varepsilon_{it}.$$

Greater availability of longitudinal data sets has enabled labormetricians to use indicator variables for each observation i in the panel to remove these unobservables and thus free the estimated α from potential contamination from them. As with selectivity corrections, randomly chosen volumes of journals specializing in labor economics yield many applications of this technique.[8] The assumption that all the individual-specific variation not captured by the variables in X arises from the unobservable is implicit in these applications, while the assumption that the unobservable is unchanging over time (is fixed) is explicit.

The former assumption generates a problem if most of the true variation in the X of interest is cross-sectional (if the X are highly

[7] One recent year's editions of the *Journal of Labor Economics* and the *Journal of Human Resources* contained seven articles employing this technique.

[8] Individual-effects estimators are less frequently used than are selectivity corrections: recent volumes of the *Industrial and Labor Relations Review*, the *Journal of Labor Economics*, and the *Journal of Human Resources* typically contain one or two articles per year using the former.

autocorrelated) and there is measurement error in the series, since applying the fixed-effects estimator then removes the true variation, leaving mainly variations in errors of measurement (Griliches and Hausman 1986). The approach then generates estimates of α that are very close to zero and have large standard errors. Consider an equation describing the (logarithm of) real compensation in a balanced panel of 100 full professors of economics at six major American public universities observed in 1979–80 and 1985–86 (Hamermesh 1989). LS estimates of the coefficients of a quadratic on recent citations by others, and on prior administrative experience, are shown in column (1) of Table 2. Even in this partly administrative data set the estimates are two to three times the size of the fixed-effects estimates shown in column (2), and the standard errors are larger. The differences reflect the mistaken equation of the *persistent* impact of citations on salaries to unobservables that we cannot identify and possible measurement error in the citation counts.

The assumption that unobservables are time-invariant is extremely difficult to credit. (After all, if the variables that we do observe vary over time, why shouldn't those that we cannot observe?) The classic example is the exclusion of unmeasured ability in an equation explaining wages. Even there, while ability may be time-invariant, its interaction with other characteristics may change with time. One partial solution if $T > 2$ is to include individual-specific time trends as well as both individual and time effects. Even that solution, however, may moderate but fail to eliminate the problem, since individual trends impose a particularly rigid structure on the nature of the time-series changes in the individual effects. There is no "quick-fix" econometric solution; all one can do is recognize the nature of the problem, find more variables to include in X, attempt to reduce measurement error, and have a good economic justification for including the individual effect even when substantial cross-section variation is captured in X.

Table 2. The Impact of Citations on Real Compensation, 100 Full Professors in 1979–80 and 1985–86.[a]

Variable	*Pooled LS*	*Fixed Effects*
Citations (t-1,...,t-5)	.00482	.00245
	(.00063)	(.00096)
Citations^2/100	−.00138	−.00041
	(.00033)	(.00032)
Administrator	.1270	.0745
	(.0244)	(.0369)

[a]Each equation also contains a quadratic in experience (years since Ph.D.), and the LS equation contains the time-invariant indicator variable, theorist.

Time-Series Analysis in Labormetrics—Gone, Forgotten, but Perhaps Not Dead

Of the 28 empirical studies in labor economics published in the *American Economic Review* during 1967–72, 57% were based on time series with $T > 10$. Of the 24 published during 1992–96, a significantly lower 33% were so based. This difference confirms impressions that labor economists have shifted their interest away from data sets with small N and relatively large T. Partly this may arise from rational behavior on the part of labormetricians, who are responding to the increased abundance of and ease of access to micro-based cross-sectional and short longitudinal data sets.

Part of the shift may also stem from increased concerns about how much we can learn about behavior using typically available time series. There are two problems. First, the time series may represent such highly aggregated forms of units i that they are incapable of reflecting the structure of the microeconomic behavior we are trying to examine. This is an increasing problem as the specifications suggested by theory lead beyond easily aggregated linear approximations to general functions that are difficult to aggregate. Second, the integrating and cointegrating restrictions imposed on variables and their relationships by modern time-series analysis strain our belief that the time-series labormetrics of the 1950s through 1970s can be informative.

Is time-series labormetrics dead, or merely moribund? Hopefully the latter, because there are questions that can be answered only by examining time-series variation. How workers respond to transitory shocks to opportunities is best studied by examining patterns of earnings, wage rates, and hours *at the individual level* using fairly long time series. Similarly, studying the dynamics of labor demand inherently requires analyzing frequently observed and long time series in order to obtain sufficient information on temporal patterns of *firm-specific* shocks to allow us to separate out general patterns of dynamics from idiosyncratic behavior (for example, Caballero et al. 1997).

With the growth of long annual sets of data on households in several countries, and the possibility of studying relatively long time series on firms' employment, investment, and other characteristics, labormetricians will have to pay more attention to time-series econometrics. Of course these are panels, and contemporary methods of handling panel data are relevant; but to the extent that we wish to study dynamics in these data (as in, for example, Baker 1997), we should be paying attention to the statistical properties of the time-series relationships among them and applying the techniques that our colleagues in macroeconomics and finance have developed for these purposes. This requires thinking about problems of causality and stationarity in time-series estimation and learning the strengths and weaknesses of the techniques time-series econometricians have developed—with diminishing attention from labor economists—over the past few decades.

The Reasonableness of Results

The previous sections have dealt with a variety of issues in applying econometric technique in situations that labormetricians confront. In this and the next section I depart from this focus to examine issues that are less technical, but no less important. Here I consider how to test estimates for their reasonableness and how to avoid creating situations that might generate unreasonable results. Throughout our own empirical research and our evaluation of others' we should consider whether the research meets "the sniff test": does it make economic sense, or does the analysis simply reflect our enchantment with some new technique, our delight at some surprising result, or our infatuation with a new set of data? In evaluating the credibility and novelty of the findings in a piece of empirical work, a useful approach is to ask oneself whether, if the findings were carefully explained to a thoughtful layperson, that listener could avoid laughing. A good inoculation against laughter is to make sure that the empirical work is grounded in economic theory.

One sniff test in studies of the impact of labor-market policies consists in bounding the economic effects. This can be done by, for example, comparing the implications of the economic effects to the sizes of the programs under study. For example, Parsons (1980) generated cross-section estimates of the impact of U.S. Disability Insurance on the labor-force participation of older men and used them to simulate the impact of actual time-series changes in those benefits. He showed that they fully accounted for the decline in participation that occurred from 1955 to 1976. The implied growth in the number of men receiving Disability Insurance benefits over that period was less than that in nonparticipation, however, so that readers might question the validity of the cross-section estimates of the elasticities.[9] Many of the studies in a large time-series literature relating higher unemployment benefits to changing aggregate unemployment (for example, Grubel and Maki 1976) imply that a 10-percentage-point decrease in replacement rates would reduce the unemployment rate to below zero! Robert

[9]The rate of nonparticipation among men 45–54 rose from 3.5% of the population in 1955 to 7.9% in 1975 (Parsons 1980:132). The percentage on DI grew from 0 to 3.9% (Bound 1989:483).

Moffitt's (1997) demonstration that a recent estimate of the extent of consumption smoothing produced by transfer programs is far too high to be consistent with the sizes of the programs and their other effects is a good application of the sniff test. At the very least, in evaluating studies of the impact of a policy one must simulate reasonable changes in it to see if the estimated effects on the outcome of interest are absurd.

Another sniff test applicable in studying a labor-market policy or institution is to use the estimates of its effects to infer the behavioral parameters that are generating them. The employment effects of a higher minimum wage, for example, should be linked to the interaction of the relative size of the low-wage work force whose wages are affected and the demand elasticity for low-wage workers. Changes in hours of work induced by changing requirements on the overtime penalty can be converted to labor-demand elasticities and compared to elasticities that have been directly estimated in other studies (Hamermesh and Trejo 2000). The estimated impact on employment fluctuations of experience-rated taxes to finance unemployment insurance yields estimates of the relative sizes of the costs of adjusting employment across industries (Anderson 1993) that should accord with interindustry differences in relative hiring and firing costs. Estimates of the impact of the U.S. Earned Income Tax Credit on hours and participation imply supply elasticities that can be compared to those generated in the huge literature that estimates them directly (Eissa and Liebman 1996).

Our data usually come ready-made, which makes our life much easier; but they reflect observations aggregated temporally over intervals that may fail to mirror the frequency of the decisions generating the behavior we wish to examine. This difficulty means, for example, that studies attempting to infer the dynamics of some economic process will generate estimates that, while plausible, have nothing to do with the underlying behavior. For example, in the 1990s a laudable innovation in studying employment dynamics was the use of panel data on firms. Unfortunately, while substantial evidence based on industry and other more aggregated data suggests that employment dynamics are fairly rapid, most of these micro panels contain only annual data that cannot identify the temporal path of adjustment of employment in response to shocks.

Leamer (1978) made a major contribution to methodology with his critique of what he believed was the common practice of reporting the last of a long line of results (optionally stopping when the results were deemed satisfactory, presumably when they rejected the relevant null hypothesis). For a variety of reasons the "fishing expeditions"—the specification searches—that Leamer deplored have become less important in labor economics since the late 1970s. Because of the large individual variation in outcomes, the practice of including ancillary variables in our equations in the hope of altering the estimated effects of the variables of interest is less worthwhile with the micro data that we increasingly use. Also, given the large size of the micro data sets, adding a variable that we think might be important for some sample respondents is not often likely to affect behavior inferred over most of the sample. Finally, one can hope that the development of economic theory and prior empirical work has improved labormetricians' ability to specify the other variables that form the controls enabling us to study the particular variable of interest.

Old-fashioned fishing is much rarer now, although people still present the results of equations reestimated after deleting all variables whose coefficients did not achieve some desired significance level in earlier specifications.[10] The low cost of applying

[10]Thomas (1997) estimated hazard rates out of unemployment in specifications from which variables whose coefficients had t-statistics below 1.4 had been purged. While this practice probably does not create major difficulties, it does prevent the reader from inferring the effects of the variables of interest in a fully specified and presumably theoretically based model.

ever-more sophisticated techniques and the professional returns to that activity have, however, led us instead to hope that what is not readily visible in the data might stand out if, for example, kernel estimation or competing-risks hazard models with unobserved heterogeneity can be applied. Technique search has replaced specification search as the fishing tackle of choice. These and other sophisticated techniques should be used if the underlying theory warrants it or if the data are obviously analyzed best by them, but not as the rationalization for a fishing expedition. Even where such techniques are called for, the careful labormetrician should first examine whether the relationships of interest are apparent in cross-tabulations or perhaps in ordinary least squares regressions that include a theoretically based set of covariates.

Perhaps the best way to avoid all the pitfalls mentioned here is to base one's claims on several *independent* sets of data (ideally covering different geographical units—hopefully with different institutional structures—from different countries, different time periods, or even different phenomena illustrating the general issue being analyzed). There is little or no reward to replication in labormetrics; but the credibility of a new finding that is based on carefully analyzing two data sets is far more than twice that of a result based only on one.[11]

Presenting Results—Light out from under a Bushel

Searching for statistical significance—"95-percent confidence interval fetishism"—should not be our goal.[12] Even if our test is powerful and generates statistically significant results, to be interesting the estimates must be discussed from the viewpoint of whether or not they are *economically important* (McCloskey and Ziliak 1996). A large effect, albeit one that is statistically insignificantly different from zero, still tells us that the best estimate is that the impact on behavior is economically significant. This approach has the additional virtue of tying the presentation of our results to a sniff test—it requires us to focus on whether the results make economic sense, not merely whether they pass muster statistically.

The majority of labormetric results are shown in tabular form. The questions are: which results, and how to present them? Constraints on journal space and the proliferation of coefficients have increasingly led authors to include in their tables only the parameter estimates of central interest. This is a welcome trend—acknowledgment in a footnote that large vectors of other variables were included usually suffices. Tables that run over a page almost always contain information that is at best secondary to the author's main point. If some standard variable does generate unusual estimates, the anomaly is worth reporting. Even better, it should alert the author before publication that something may be severely wrong with the underlying data or specification.

Most of our raw data contain three or occasionally four digits. To report six-digit parameter estimates, even though they are readily "cut-and-pasted" from one's statistical log, is thus silly—three significant digits (after zeros) are the most that one can meaningfully discuss. When the first significant digit is the fourth after the decimal point, one simple way to save space and avoid inducing blindness in the reader is to redefine the variable so that the estimate rises by several orders of magnitude (for example, Citations^2/100 in Table 2). One should not report a coefficient, standard error, or p-value as being, for example, .000 (especially since no standard error or p-value could ever be zero).

Whether one is presenting the effects of indicator variables or others, the reader should be able to tell what the variable is.

[11]Milton Friedman noted, "I have long had relatively little faith in judging statistical results by formal tests of statistical significance. I believe that it is much more important to base conclusions on a wide range of evidence coming from different sources over a long period of time." (1987, quoted in Hammond [1996:202].)

[12]I am indebted to Jeff Biddle for this term.

Too many authors list variable names in mnemonic form (and too many journals indulge this bad habit). Even if the variables are defined elsewhere in the study, no reader should be required to search back repeatedly through the paper or to memorize their definitions. The variables should be referred to clearly in each table where estimates relating them to the dependent variable are presented.

Often the parameter estimates that are presented in labormetric publications have little or no intuitive *economic* meaning of their own. This includes parameters from probit or ordered probit models, estimates of structural parameters in systems of demand equations or cost/production tableaux, and others. Authors serve themselves and their readers far better by presenting economically interesting transformations of the parameter estimates. Thus, rather than present probit parameters associated with some variable X, it is better to present an estimate of $\partial \Pr\{Y = 1\}/\partial X$; and with ordered probits (so long as the number of categories is small), it is better to present the effects of a one-unit increase from the mean of X on the probability of Y being in each category. Similar good sense should apply in presenting results based on other techniques. With LS estimates, unless the variables are in logarithms, the reader should be given means of the crucial variables in order to use the parameter estimates to compute elasticities.

Reasonable people may differ about whether presenting t-statistics or standard errors is more useful, but I prefer standard errors. Most readers can divide by 2 (or 1.64, or 1.28) to obtain significance levels; and standard errors facilitate making cross-equation comparisons, calculating the partial effects of combinations of the variables, or testing pairwise constraints on the variables (with assumptions about the estimated covariances). Perhaps most important, in many cases the null hypothesis of interest is that the parameter equals some specific nonzero value (for example, unit marginal effect on consumption of some income flow).

We list estimates of parameters relating each of a vector of indicator variables X to Y, arbitrarily excluding one of the components of X. The reader can interpret results more easily if the parameter estimates are all of one sign, so the estimates should be recomputed before publication to achieve this end. Indeed, except where there is some natural order to the categories (as opposed to occupation or industry), listing the categories in the order of the indicators' effects on Y facilitates interpretation.

We should always aim to present our findings in such a way as to provide the clearest and strongest impression of what we have discovered. Labormetricians have always published their results (and sometimes their data too) in tabular form; but since the late 1980s the use of graphics has boomed due to technical changes in personal computing.[13] When should one follow this trend and use graphical instead of tabular presentation? A simple answer is to read Tufte (1983) and follow its guidelines. In many cases where we would heretofore have had to present long, repetitious tables listing numerous parameter estimates whose particular values are not crucial to our conclusions, today we can present them more succinctly and more potently in graphic form. For example, the results of estimating hazards are almost always much more clearly presented as graphs. Indeed, in most cases where there are many parameter estimates that can be arrayed temporally, a graph is more enlightening than a large table.

Like our increasingly accessible high-powered econometric tools, our increased graphics capabilities have led to misapplications. One should follow Tufte's dicta: (1) avoid graphs depicting one or two time series, since they are visually boring and

[13] In 1988 and 1989, 15% of the 113 empirical articles published in regular issues of the three leading American labor journals (*Industrial and Labor Relations Review, Journal of Human Resources,* and *Journal of Labor Economics*) included graphs of results or data. The corresponding figure for the 188 empirical articles published in 1995 and 1996 was 34%.

their content can be presented more concisely verbally; and (2) avoid graphs that are so cluttered with written descriptions as to hide the message of the data. New technology has also led to overkill: the reader need not be assaulted with pages of graphs depicting each set of results for a large number of industries; a few typical ones, plus a footnote referring to the others, suffice. Similarly, graphs showing the results of simulations describing how the results would be altered by a large variety of small changes in policy parameters have remarkably soporific effects.

Conclusion—A Warning about Wizardry

In 1978 I was studying the relationship between unemployment insurance benefits and retirement behavior. The crucial variable provided information on whether the individual was fully retired, partly retired, or working. Due to the unavailability of statistical packages, my research assistant and I spent substantial effort modifying a program to estimate the appropriate multinomial logit. In 1994 a major general journal accepted an article of mine with the admonitory proviso that I replace the ordered-probit estimation of educational attainment (which the data classified into intervals) by least squares to aid the readers' comprehension. I complied with the request.

In both cases I was wrong. Spending so much effort on what was for the 1970s quite sophisticated applied econometrics was a poor way to allocate time for someone who is basically an economist. No reasonable person could have expected the more esoteric technique to produce results very different from those generated by simpler techniques, and my time would have been more wisely spent gathering better data and trying to understand the economics of the behavior I was studying. By the same token, complying with the editor's request in 1994 represented a step backward from the frontier of knowledge, was foolish in light of most readers' fluency with the technique, and detracted from the story that the article had to tell.

The moral is clear: apply a benefit-cost calculation to the use of econometric technique. Labormetric research is not a cadenza designed to show off the sophistication of our tools. Its sole purpose should be to provide an empirical description of labor-market outcomes that helps to illuminate economic behavior. Sophisticated techniques can enhance description and shed additional light on behavior, but they should be pursued just shy of the point at which their time costs begin to eclipse their benefits in terms of data quality and characterizing the economic relationships that generate the outcomes. Before we resort to wizardry we should be certain that we do not add confusion by making mistakes with simpler techniques; and we should make very sure that the sophisticated technique is apropos our research question.

REFERENCES

Abowd, John, Francis Kramarz, and David Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica,* Vol. 67, No. 2 (March), pp. 251–334.

Akerlof, George, and Janet Yellen. 1985. "Unemployment through the Filter of Memory." *Quarterly Journal of Economics,* Vol. 100, No. 3 (August), pp. 747–74.

Anderson, Patricia. 1993. "Linear Adjustment Costs and Seasonal Labor Demand: Evidence from Retail Trade Firms." *Quarterly Journal of Economics,* Vol. 108, No. 4 (November), pp. 1015–42.

Anderson, Patricia, and Bruce Meyer. 2000. "The Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims, and Denials." *Journal of Public Economics,* forthcoming.

Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association,* Vol. 91, No. 434 (June), pp. 444–72.

Angrist, Joshua, and Alan Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics,* Vol. 106, No. 4 (November), pp. 979–1014.

Baker, Michael. 1997. "Growth Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings." *Journal of Labor Economics,* Vol. 15, No. 2

(April), pp. 338–75.
Biddle, Jeff, and Daniel Hamermesh. 1990. "Sleep and the Allocation of Time." *Journal of Political Economy,* Vol. 98, No. 5 (October), pp. 922–43.
Borjas, George. 1980. "The Relationship between Wages and Weekly Hours of Work: The Role of Division Bias?" *Journal of Human Resources,* Vol. 15, No. 3 (Summer), pp. 409–23.
Bound, John. 1989. "The Health and Earnings of Rejected Disability Insurance Applicants." *American Economic Review,* Vol. 79, No. 3 (June), pp. 482–503.
Bound, John, David Jaeger, and Regina Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variables Is Weak." *Journal of the American Statistical Association,* Vol. 90, No. 430 (June), pp. 443–50.
Butcher, Kristin, and Anne Case. 1994. "The Effect of Sibling Sex Composition on Women's Education and Earnings." *Quarterly Journal of Economics,* Vol. 109, No. 3 (August), pp. 531–64.
Caballero, Ricardo, Eduardo Engel, and John Haltiwanger. 1997. "Aggregate Employment Dynamics: Building from Microeconomic Evidence." *American Economic Review,* Vol. 87, No. 1 (March), pp. 115–37.
Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review,* Vol. 43, No. 2 (January), pp. 245–257.
____. 1996. "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica,* Vol. 64, No. 4 (July), pp. 957–79.
Davis, Steven, and John Haltiwanger. 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." *Quarterly Journal of Economics,* Vol. 107, No. 3 (August), pp. 819–62.
Eissa, Nada, and Jeffrey Liebman. 1996. "Labor Supply Response to the Earned Income Tax Credit." *Quarterly Journal of Economics,* Vol. 111, No. 2 (May), pp. 605–37.
Elliott, Robert, and Robert Sandy. 1998. "Adam Smith May Have Been Right after All; A New Approach to the Analysis of Compensating Wage Differentials." *Economics Letters,* Vol. 59, No. 1 (April), pp. 127–31.
Evans, David, and Linda Leighton. 1995. "Retrospective Bias in the Displaced Worker Surveys." *Journal of Human Resources,* Vol. 30, No. 2 (Spring), pp. 386–96.
Griliches, Zvi, and Jerry Hausman. 1986. "Errors in Variables in Panel Data." *Journal of Econometrics,* Vol. 31, No. 1 (February), pp. 93–118.
Gronau, Reuben. 1974. "Wage Comparisons: A Selectivity Bias." *Journal of Political Economy,* Vol. 82, No. 6 (November–December), pp. 1119–43.
Grubel, Herbert, and Dennis Maki. 1976. "The Effect of Unemployment Benefits on U.S. Unemployment Rates." *Weltwirschaftliches Archiv,* Vol. 112, No. 2, pp. 274–99.
Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review,* Vol. 84, No. 3 (June), pp. 622–41.
Hamermesh, Daniel. 1982. "Social Insurance and Consumption: An Empirical Inquiry." *American Economic Review,* Vol. 72, No. 1 (March), pp. 101–13.
____. 1989. "Why Do Individual-Effects Models Perform So Poorly? The Case of Academic Salaries." *Southern Economic Journal,* Vol. 56, No. 1 (July), pp. 39–45.
Hamermesh, Daniel, and Jeff Biddle. 1994. "Beauty and the Labor Market." *American Economic Review,* Vol. 84, No. 5 (December), pp. 1174–94.
Hamermesh, Daniel, and Paul Menchik. 1987. "Planned and Unplanned Bequests." *Economic Inquiry,* Vol. 25, No. 1 (January), pp. 55–66.
Hamermesh, Daniel, and Stephen Trejo. 2000. "The Demand for Hours of Labor: Direct Evidence from California." *Review of Economics and Statistics,* Vol. 82, No. 1 (February), pp. 38–47.
Hammond, J. Daniel. 1996. *Theory and Measurement: Causality Issues in Milton Friedman's Monetary Economics.* Cambridge: Cambridge University Press.
Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement,* Vol. 5, No. 4 (Fall), pp. 475–92.
Hunt, Jennifer. 1992. "The Impact of the 1962 Repatriates from Algeria on the French Labor Market." *Industrial and Labor Relations Review,* Vol. 45, No. 3 (April), pp. 556–72.
Iacovou, Maria. 1996. "Fertility and Female Labor Force Participation." Diss., University College–London.
Juster, F. Thomas, and Frank Stafford. 1991. "The Allocation of Time: Empirical Findings, Behavioral Models, and Problems of Measurement." *Journal of Economic Literature,* Vol. 29, No. 2 (June), pp. 471–522.
Kahn, Joan, and J. Richard Udry. 1986. "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review,* Vol. 51, No. 5 (October), pp. 734–37.
Keane, Michael, and Kenneth Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy,* Vol. 105, No. 3 (June), pp. 473–522.
Koenker, Roger, and Gilbert Bassett. 1978. "Regression Quantiles." *Econometrica,* Vol. 46, No. 1 (January), pp. 33–50.
Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Non-Experimental Data.* New York: Wiley.
Lewis, H. Gregg. 1963. *Unions and Relative Wages in the United States.* Chicago: University of Chicago Press.
Manski, Charles. 1991. "Regression." *Journal of Economic Literature,* Vol. 29, No. 1 (March), pp. 34–50.
McCloskey, Deirdre, and Stephen Ziliak. 1996. "The Standard Error of Regressions." *Journal of Economic Literature,* Vol. 34, No. 1 (March), pp. 97–114.
Meyer, Bruce. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics,* Vol. 13, No. 2 (April), pp. 151–62.
Moffitt, Robert. 1997. "Comment on Stephen

Dynarski and Jonathan Gruber, 'Can Families Smooth Variable Earnings?'" *Brookings Papers on Economic Activity,* pp. 285–92.

Mueller, Richard. 1998. "Public-Private Sector Wage Differentials in Canada: Evidence from Quantile Regressions." *Economics Letters,* Vol. 59, No. 2 (August), pp. 229–35.

Newey, Whitney, and James Powell. 1993. "Efficiency Bounds for Some Semiparametric Selection Models." *Journal of Econometrics,* Vol. 58, Nos. 1–2 (July), pp. 169–84.

Oettinger, Gerald. 1999. "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors." *Journal of Political Economy,* Vol. 107, No. 2 (April), pp. 360–92.

Parsons, Donald. 1980. "The Decline in Male Labor Force Participation." *Journal of Political Economy,* Vol. 88, No. 1 (February), pp. 117–34.

Philipson, Tomas. 1997. "Data Markets and the Production of Surveys." *Review of Economic Studies,* Vol. 64, No. 1 (January), pp. 47–72.

Reimers, Cordelia. 1998. "Unskilled Immigration and Changes in the Wage Distributions of Black, Mexican American, and Non-Hispanic White Male Dropouts." In Daniel Hamermesh and Frank Bean, eds., *Help or Hindrance? The Economic Implications of Immigration for African Americans.* New York: Russell Sage Foundation, pp. 107–48.

Shin, Kwanho 1997. "Inter- and Intrasectoral Shocks: Effects on the Unemployment Rate." *Journal of Labor Economics,* Vol. 15, No. 2 (April), pp. 376–401.

Thomas, Jonathan. 1997. "Public Employment Agencies and Unemployment Spells: Reconciling the Experimental and Nonexperimental Evidence." *Industrial and Labor Relations Review,* Vol. 50, No. 4 (July), pp. 667–83.

Tufte, Edward. 1983. *The Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press.