NBER WORKING PAPER SERIES

THE ART OF LABORMETRICS

Daniel S. Hamermesh

Working Paper 6927
http://www.nber.org/papers/w6927

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 1999

Edward Everett Hale Centennial professor of economics, University of Texas at Austin, and research associate, National Bureau of Economic Research. I thank Julian Betts, Jeff Biddle, Charlie Brown, Barry Hirsch, Jacob Klerman, Dan Slesnick and participants in seminars at several universities and conferences for comments and suggestions on earlier drafts, Steve Allen for comments and for the term "labormetrics," and a number of the authors cited here who clarified their research for me. The views expressed here are those of the author and do not reflect those of the National Bureau of Economic Research.

The Art of Labormetrics
Daniel S. Hamermesh
NBER Working Paper No. 6927
February 1999
JEL No. J00, C1

ABSTRACT

Using a wide array of examples from the literature and from original estimates, this essay examines the pitfalls that make good empirical research in labor economics as much art as science. Appropriateness and cleanliness of data are considered, as are problems of extreme observations and interactions. The validity of attempts to produce exogeneity using instrumental variables and "natural experiments" is examined, as are the treatment of selectivity and unobservable individual effects. Testing empirical results to ensure that they make sense is stressed along with the importance of clear, economical and useful presentation of those results.

Daniel S. Hamermesh
Department of Economics
University of Texas
Austin, TX 78712-1173
and NBER
hamermesh@eco.utexas.edu

I. Introduction

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 art of knowing what makes economic sense and of emphasizing those clever ideas that will make the largest substantive contribution to one's work. My discussion here makes no claim to technical originality. Instead, its purpose is to provide younger applied economists with a feel for what might be important, what to do and, perhaps most of all, what not to do in empirical work. While my examples stem almost exclusively from the literature of labor economics, they also illustrate difficulties in empirical work in other subspecialties of economics and in the statistical analysis of labor issues from the approaches of other disciplines. All the discussion presupposes the possibly subversive notion that analyzing data is a central part of applied economics -- and of social science more generally.

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, etc. -- 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 but 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 rightly gone to those who have developed new approaches that solve old, often ill-perceived problems of

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 their application is as appropriate as the original, or whether instead the technique is just clouding 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 economic 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 the admonitions about such misapplications and misinterpretations with examples from the recent literature of empirical labor economics and with 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 it and with the data sets that underlay it.

II. 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 impacts, and infer the behavior that caused them. Too often we mindlessly 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 ourselves whether we have found the best <u>available</u> data for the purpose and, more important, whether those data offer any hope of representing the concept. If they do not, we must either collect our own data, or failing that revert to doing applied theory, since we are deluding ourselves if we use such data for purposes of inferring behavior. In research on labor supply, for example, we have 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.[1] This is not a matter of classical measurement error, although I deal with that later in this section. Rather, the issue is whether we are able to match empirical proxies to our theoretical constructs.

Another way in which we obviate our chances 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) examines 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. In the same issue of the same journal Shin (1997) is concerned about the relative importance of micro and macro shocks to employment. While he laudably bases the study on firm-level data, the restriction is to manufacturing (which in the U.S. accounted for 15 percent of total employment and 17 percent of GDP in 1996). Indeed, this "manucentrism" pervades the much-emulated research on idiosyncratic employment changes (e.g., Davis and Haltiwanger 1992) and 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 can at least hope to provide answers, the main issue concerns measurement error of the sort:

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

where $X^*_{i\,t}$ is the true measure, $X_{i\,t}$ is what we observe, and $\theta_{i\,t}$ is the error. (Throughout this discussion 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 $E(\theta_{i\,t}) = 0$. That would not be a problem if this expectation were constant; but it need not be. There are cases where $E(\theta_{i\,t})$ 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, namely the measurement and explanation of hours of work. The concept we wish to examine is the amount of some time interval (day, week, year, lifetime) devoted to market production. Nearly all our research is based upon retrospective data, in which respondents to a survey are asked to describe their activities in some past period. Comparing such data to measures of market hours from diaries that record actual time use (Juster and Stafford 1991), the former consistently overerestimate work time. Worse still, the overestimates differ with observed characteristics (differ predictably across i). Even more problematic, the differences across i suggest that the $E_t(\theta_{i\,t})$ may also be trending, although the regrettable absence of repeated cross-sections of time diaries in most countries makes direct inference impossible. All of this means that our inferences about trends in hours of market work and about the changing responsiveness of market work 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.

It is easy to bemoan the impacts of what one might call conceptual measurement error, but we are still in the business of measuring as best we can. When finding data that solve this problem is impossible, we are obligated at the least to acknowledge it, to estimate its impacts on the descriptive statistics of the phenomenon we are measuring, and to deduce the direction of the bias that it imparts to the behavioral relationships we are estimating.

Our econometrics texts treat classical measurement error as God-given, something that is there in the data and the implications of whose existence for our estimates must be inferred. It is not. In some cases we ignore measurement errors that, if simply noticed and cleaned, enhance the quality of our results. 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 (Hamermesh 1982):

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

where the constant was insignificantly different from 0, and YD is the household's other income. An unpublished estimate of the same 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.[2] The moral of this anecdote is that one must check the descriptive statistics, especially the minima and maxima, of all series prior to any estimation.[3]

In other cases we induce the measurement errors ourselves. One (here unnamed) researcher submitted a paper in which he estimated a log-earnings equation designed to infer the rate of return to schooling. The estimated return seemed unusually low. It transpired that the author had mistakenly assumed that the data source reported actual years of schooling, which was correct for most observations; but for those for which data were unavailable education was coded as 99! We often unthinkingly create error-ridden data out of cleaner raw data, such as when we use as dependent variables measures of wage rates constructed by dividing annual or weekly earnings by hours (Borjas 1980).

Other measurement errors arise from careless responses to surveys. While these are neither conceptual nor induced, their impacts on our estimates can be inferred, and in some cases the errors can be reduced, if we can model and perhaps alter the behavior of the underlying agents whose response are error-ridden (Philipson 1997). Work in this area is just getting underway; but already it 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.

III. Extreme Observations

Because of its assumption that the sum of squared errors is minimized to derive parameter estimates, the Gauss-Markov theorem implies that we weight 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 way to infer impacts at sample averages (aside from being convenient); 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:

(3) $$Y_{it} = \alpha X_{it} + \epsilon_{it},$$

are possible (Manski 1991), with minimizing the sum of the $|\epsilon_i|$ one common approach (equivalent to estimating a regression line through the sample medians). 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 (<u>not</u> 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. On the first two I present earnings regressions. Part I of Table 1 reports LS (least squares) and LAD (least absolute deviations) estimates of a simple bivariate equation relating the average salary of full professors of economics in 17 major public universities in 1996-97 to the average ranking of quality of the department.[4] Even in this very small sample the estimates of the impact of reputation on average salary differ remarkably little between LS and LAD.

This may be because the distribution around the mean is both fairly tight and more or less symmetric. In Part II of Table 1 I reproduce from Hamermesh and Biddle (1994) the estimated impacts of physical appearance, indicator variables denoting whether the respondent is in the bottom 15 percent of physical appearance or the top 30 percent, on 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 the standard deviation than it was in the first example.

The third example explains weekly sleep time by a variety of demographic variables as well as minutes of market work. The LS estimates on this last variable are reported in Biddle and Hamermesh (1990, Table 3, column 1) and reproduced in Part III of Table 1. The final column in the Table shows that the LAD estimates differ little from these 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 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. My experience with other sets of data suggests the same thing. No doubt, however, there are data sets where this choice of technique may matter; and, since statistical packages have made it increasingly easy to examine the sensitivity of one's estimates to the least-squares assumption, 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.[5]

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. The related issue, whether underlying behavior differs at different points of the distribution of Y, is economic. In some cases we are just interested in discovering the impact of some RHS variable at different points of the distribution of Y. A typical case involves estimating wage equations where we wish to test whether an institution or policy affects wages differently over their distribution. For examples, where in the distribution of wages is the wage gain from trade-union membership or from public-sector employment greatest, adjusted for workers' skills (Card 1996; Mueller 1997)?

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 workers 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 generally no reason to assume that the α are constant. With more observations one can estimate the relationship at more quantiles. (There is no rule relating the number of quantiles to sample size; but I have found that having several hundred observations per quantile is generally the minimum necessary to make this technique worth pursuing.)

With larger data sets and low-cost methods of estimating quantile regressions, examining the constancy of α over the distribution of Y makes sense.

While in many cases we are interested in the X→Y relationship over the range of Y, in others we are interested in how the relationship is modified by variations in some other exogenous variable Z. For examples, the complementarity of formal and informal training leads us to expect "fanning-out" in age-earnings profiles, and we can test this by interacting a measure of education with measures of experience in a wage equation. Income and substitution effects in female labor supply should differ in the presence of young children, so that estimating them requires interacting indicator variables for family structure with measures of wage rates and other family income in equations describing women's labor-force participation and hours of work.

In these and many other cases our theory indicates that we should reestimate (3) as:

$$(4) \qquad Y_{it} = \alpha X_{it} + \beta X_{it} Z_{it} + \epsilon_{it} .$$

This approach is dangerous and wrong, for it restricts $\partial Y/\partial Z$ to operate only through X . No matter how tightly held is the belief that the interaction is important, one must estimate the equation as:

$$(5) \qquad Y_{it} = \alpha X_{it} + \beta X_{it} Z_{it} + \gamma Z_{it} + \epsilon_{it} .$$

Main effects must be present for every variable included in interaction terms. If a rigorously derived prediction implies only an interaction, at the very least one must test the main effect.[6] Interactions should also be included with all terms that make up a nonlinear main effect, not merely with the linear term.[7]

To understand the pitfalls of mistakenly estimating (4), consider the equation presented in Part II of Table 1 that related men's wages to indicators of their looks. A reasonable hypothesis is that employers pay attention to looks when hiring new workers, but that the effect of beauty diminishes

as workers acquire firm-specific experience and are increasingly rewarded for their skills rather than their ascriptive characteristics.[8] In this data set 22 percent of the sample has more than 10 years of job tenure, with the rest described by indicator variables for less than 1 year, 1-3 years, or 3-10 years of tenure in the firm. The equation underlying Part II of Table 1 included these three indicator variables as the vector X. The two indicators for looks make up Z.

The estimates in the first column of Table 2 show what happens when we mistakenly exclude the main effect for Z. There is a negative wage penalty for bad looks, all else equal, but we would interpret it as existing only among workers who have been on the job fewer than 10 years. The results would thus confirm the hypothesis that led us to include the interactions. In fact, this finding and the hypothesis are quite wrong: When the correct equation (5) is estimated, the results in the second column of Table 2 show that there is no significant interaction of looks and job tenure. The estimates of β in (4) mistakenly attributed the effect of beauty on wages to the interaction of beauty with job tenure.

Failure to include main effects for all interactions is happily fairly uncommon in published work.[9] Much more common is the specification:

$$Y_{it} = \alpha X_{it} + \beta Z1_{it} Z2_{it} + \epsilon_{it} .$$

In the recent literature Z1 has been the coverage of the minimum wage, the amount of unemployment insurance benefits a worker receives, or a geographic indicator. Z2 has correspondingly been the level of the minimum wage relative to some average wage, the inverse of the worker's wage, or an indicator for the presence of some law. In all of these cases the researcher cannot infer from the $\hat{\beta}$ whether Z1 or Z2 is generating the effect on Y. Good econometric practice suggests that the

equation should be estimated with main effects of Z1 and Z2 as well as the interaction, and the results should at least be in a footnote.

A related issue occurs in the extremely common case when one includes higher-order terms in some of the variables in X in (3), either because the theory leads us to expect a nonlinear relation, or simply because we wish to test for one. If, for example, one includes terms such as $\alpha_1 X_{i\,t} + \alpha_2 X^2_{i\,t}$ in (3), it is essential to calculate $\partial Y/\partial X = \alpha_1 + 2\alpha_2 X$ and to discuss the direction, size and significance of the relation between Y and X at various values of X, particularly at its mean.[10] Similarly, in the case of interactions and nonlinear terms, having estimated the correct equation (5) one must present, or at least discuss, estimates of $\partial Y/\partial X$ and $\partial Y/\partial Z$ at various values of X and Z, as well as their standard errors calculated based on the variances and covariances of the parameter estimates. In that way one can infer the mean effect as well as the impacts of X and Z on Y over their ranges.

IV. Experiments and Instruments — Circumventing Endogeneity?

Since the development of modern econometrics we have been concerned with possible violations of the assumption of the Gauss-Markov theorem that $E(X\epsilon) = 0$. From the 1940s through the 1960s that concern led to intense concentration on developing the simultaneous-equation methods, beginning with instrumental variables techniques, 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. In the 1990s labor economists played a pioneering role in research addressing this concern, and a huge amount of attention was devoted to attempts to account for causality in labormetric relationships.

A few examples are useful in motivating recent concerns. One of the staples of labormetrics is:

(6) $\log(W_i) = \alpha ED_i + \beta X_i + \epsilon_i$,

where W is worker i's wage rate or earnings, ED 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 demand for labor L from equations such as:

(7) $L_{it} = -\alpha W_{it} + \beta X_{it} + \epsilon_{it}$,

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

One line of proposed solutions to this type of simultaneity problem in labor economics has been what their proponents call "natural experiments." Some event occurs 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, <u>provided</u> 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 where the shock did not occur, but where the X can either also be measured, or, more

commonly, assumed to have changed identically and have the same marginal impacts on Y as in observations i, so that the double-difference $\Delta^2 = [Y_{i\,2} - Y_{i\,1}] - [Y_{j\,2} - Y_{j\,1}]$ can be calculated as an unbiased estimate of the effect of the exogenous shift in the RHS variable.[11]

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. This problem is hardly unique to the natural experiment approach; but that approach has generated a novel solution, namely finding additional observations i' similar to i but unaffected by the "experiment," and others j' similar to j, and calculating the triple-difference:

$$(8) \qquad \Delta^3 = \{[Y_{i\,2} - Y_{i\,1}] - [Y_{j\,2} - Y_{j\,1}]\} - \{[Y_{i'\,2} - Y_{i'\,1}] - [Y_{j'\,2} - Y_{j'\,1}]\} .$$

In one example Hamermesh and Trejo (1997) attempt to infer 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 males and j be females in California, and letting the (') be observations outside California. Even this approach ignores the strong possibility that other variables that affect the Y have changed differentially over time across areas. An essential extension, given the difficulty of claiming that groups j, i' and j' are otherwise identical to group i, is thus to replace the $Y_{k\,t}$ in (8), or in the calculation of Δ^2, by $E(Y_{k\,t} \mid X_{k\,t})$, conditioning on as many theoretically-based components of X in i and j (and i' and j') as are available in the data at hand. From this perspective double- and triple-differencing without additional conditioning variables should be viewed as a last resort when information on the components of X is absent.[12]

An important issue is whether the observations $Y_{i\,1}$ and $Y_{i\,2}$ measure the outcome before and after the shock occurred. Is t=1 sufficiently distant from the shock that agents had not yet begun adjusting to an event that may have been partly expected? Obversely, if one is interested in long-run

impacts of the change (which is what most of our theories discuss), is t=2 sufficiently long after the shock 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 other factors that are unaccounted for but that affect Δ^2 (or Δ^3) are increasingly likely to have changed.

There is probably no problem with t=1 (1980) in Klerman and Leibowitz's (1997) study of the impact of state maternity leave laws passed in the late 1980s; but t=2 is 1990, making it highly unlikely that agents had sufficient time to adjust their fertility and labor supply to the new choice sets facing them. The opposite potential problem arises in Hamermesh and Trejo (1997): t=2 (1985) is sufficiently long after the effective date of the change to have allowed adjustments in the input of overtime hours; 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 (even though the actual extension did not occur until 1980). That this timing problem generates difficulties is shown by the effect on the estimates if we let t=1 be 1978 instead of the more appropriate 1973: The impact of the shock on the fraction of male workers putting in more than 8 hours per day drops from -0.052 to -0.022; that on the fraction working exactly 8 hours drops from 0.077 to 0.024, and that on average overtime hours drops from -0.206 to -0.135. In this example, and I believe typically, observing at t=1 too close to the event biases its estimated impact toward zero. A good way of circumventing the 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 $Y_{i\,1}$); otherwise, $Y_{i\,1}$ 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. Some studies of the impact of migration (to the United States from Cuba, Card 1990; and to France from Algeria, Hunt 1992) are cases where claims of exogeneity can be fairly convincing. Except where legal changes are imposed on many subunits by a higher level of government, however, treating them as exogenous is much less convincing.

The "natural-experiment" approach is hardly a panacea for circumventing endogeneity. The user of this approach should make every effort to obtain data on as many X variables as possible, to observe the Y sufficiently before and after the "experiment" and to be convincing that the "experiment" represents an exogenous shock. A healthy skepticism about the potential for success along all of these dimensions is in order.

The other recently revived approach to endogeneity in labor economics is similar, but relies instead on finding clever instrumental variables that meet the criteria that they are correlated with the shocking variable of interest (ED in (6), W in (7)) but uncorrelated with the error term. Much of the focus has been on measuring the returns to schooling (essentially α in (6)), with instruments chosen being such items as date of birth, because compulsory schooling requirements have cut-off dates that impose exogenous and discrete constraints on schooling decisions (Angrist and Krueger 1991); siblings' sex composition, because it affects women's schooling and may not be related to earnings except through schooling (Butcher and Case 1994); and smoking, because it may reflect individuals' discount rates but be unrelated to access to funds for schooling (Evans and Montgomery 1994).

These innovations have a surface appeal; but, as with the earlier literature on instrumental variables, their validity in part rests on whether the instrument explains much of the variation in the supposed endogenous variables. Bound et al (1995) cover this very well in the context of using birth date as an instrument for education, and they illustrate clearly the problems that arise when the instrument's correlation with the variable for which it is instrumenting is low. It also rests on whether behavior adapts to them in such a way as to render the instrument's exogeneity suspect. The new literature is often more precise than it predecessors in thinking about the conditions for identification, but proponents of searching for instruments are often too quick to assume that the chosen instrument is exogenous. In the case of date of birth, for example, parents can choose whether to "hold back" from starting 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, thus affecting subsequent wages and rendering the instrument's exogeneity to the schooling decision questionable. Substantial evidence indicates a positive intergenerational correlation of smoking; coupled with the negative relationship between wealth and smoking, a youth's smoking is thus not a proper instrument for his education. As with the natural-experiment approach, one must in the end be able to argue that the instrument itself is beyond the decision-makers' control, that it describes behavior that is randomly distributed in the population one wishes to describe, and that the environment in which the outcome arises does not affect the RHS variable through their behavior.

V. Selected Unobservables and Not-So-Fixed Effects

Since the late 1970s two economic/technical issues have captivated labormetricians, sample selectivity and the importance of unit-specific effects. These two are related in that both deal in some way with problems generated by behavioral effects in our main relations (equations (3)) that produce

subtle biases in the estimates of the α on the X variable(s) of interest. Selectivity problems arise because we believe that there are unobserved correlates of Y that bias the α because they determine whether a data point is included in the sample. Problems with individual effects result from our beliefs that a bias is induced because unobservables are correlated with both Y and X. These are powerful ideas that have led some of the best minds in econometrics to generate solutions that account for them. By the early 1990s canned statistical packages enabled labormetricians to apply these solutions to their own research problems at very low cost.

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

(9) $Y1_{it} = \alpha X_{it} + \beta Y2_{it} + \epsilon_{it}$, observed if:

(10) $Y2_{it} \geq \gamma Z_{it} + \nu_{it}$,

where the Yk are endogenous variables, the α and β are parameters and the ϵ and ν 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 (9), had a very clear economic interpretation, with the variables in Z representing the value of time in the home, those in X representing the nonwage variables that shift labor supply.

The solution (the so-called Heckman correction) has been applied repeatedly and increasingly, as a perusal of recent issues of labor journals or major general journals shows.[13] 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, the inverse Mills' ratio included in the

estimation of (9) is essentially a nonlinear function of the variables in X. The user of this correction should present a good rationalization for excluding the Z from (9) and the X from (10), and should either present estimates of (9) 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.

Finding that the selectivity term is insignificant in (9) may be evidence that the model is underidentified, not that selectivity is unimportant. Even if the correction "matters," one must have an economic theory justifying (9) 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 often lacks 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 the α and β.

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

$$(11) \qquad Y_{it} = \alpha X_{it} + \phi_i + \epsilon_{it} .$$

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.[14] 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 assuming that the unobservable is unchanging over time (is fixed) is explicit.

The former assumption generates a problem if most of the variation in the X of interest is cross-sectional (if the X are highly autocorrelated), since applying the fixed-effects estimator then removes that variation.[15] Not surprisingly, the approach then generates estimates of α that are very close to zero. Consider estimating an equation describing the (logarithm of) real compensation received by a balanced panel of 100 full professors of economics at six major American public universities observed in 1979-80 and 1985-86 (Hamermesh 1989). The least-squares estimates of the coefficients of a quadratic on recent citations by others, and on prior administrative experience, are shown in column (1) of Table 3. The estimates range from twice to three times the size of the fixed-effects estimates shown in column (2).[16] The attenuated estimates reflect the mistaken equation of the persistent impact of citations on salaries to unobservables that we cannot identify. In general there is no way of dealing with this issue; but this technique is best suited to cases where the intraunit autocorrelation in X is low relative to the cross-section variation. A good check on the technique is thus to decompose the variance in each X in the hope that relatively little is due to individual or time-specific effects.

The assumption that the 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 vitiate 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, and have a good economic justification for including the individual effect even when substantial cross-section variation is captured in X.

VI. 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 percent were based on time series with T>10. Of the 24 published there during 1992-96, a significantly lower percentage, 33, were so based. Obviously this is a nonrandom survey, but it 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 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. The time series may be aggregates of units i to the point that they are incapable of reflecting the structure of the microeconomic behavior that 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. The second potential difficulty is our new awareness that modern time-series analysis imposes integrating and cointegrating restrictions on the variables and their relationships that make it harder to believe that the time-series labormetrics of the 1950s through 1970s can be informative.

Is time-series labormetrics dead, or merely moribund? I hope it is the latter, because there are questions that are inherently 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 (e.g. 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 the panel-data methods that today's graduate students in labor economics learn are relevant; but to the extent that we wish to study dynamics in these data, 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 us to think about problems of causality and of stationarity in time-series estimation and to learn (the strengths and weaknesses of) the techniques that time-series econometricians have developed during the decades that it has increasingly escaped our attention.

VII. Applying the "Sniff Test"

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. The next section considers how to present results so as to make them comprehensible to the reader.

Throughout our own empirical research and the evaluation of others' we should think whether the research meets "the sniff test": Does it make economic sense, or does the analysis simply reflect our enchantment with some new technique that we have created or happened upon, our delight at some surprising result, or our infatuation with a new set of data? In evaluating the innovation of a piece of empirical work a useful approach is to ask oneself whether, if the result 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 way of applying the sniff test in studies of the impact of labor-market policies is to bound the economic effects. This can, for example, be done by comparing their implications 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, so that readers might question the validity of the cross-section estimates of the elasticities.[17] A similar problem arises in a large international time-series literature from the 1970s that related higher unemployment benefits to changing aggregate unemployment rates. Many of those studies (e.g., Grubel and Maki 1976) imply that a 10-percentage-point decrease in replacement rates, well within the range of policy choices, would reduce the unemployment rate below zero! Robert 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 impacts 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 impacts 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 impact to infer the behavioral parameters that are generating them. Estimates of 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 1997). 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 our notions of interindustry differences in relative hiring and firing costs. Estimates of the impact of the U.S. Earned Income Tax Credit (essentially an income-tax rebate for low-wage workers based on their earnings) 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 that we wish to examine. This difficulty means, for example, that studies that attempt 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 has been 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 and are thus incapable of identifying 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 one had produced in a research project (optionally stopping when the results were deemed satisfactory, presumably when they rejected the desired null hypothesis). For a variety of reasons the "fishing expeditions" -- the specification searches -- that Leamer deplored are likely to have become less important in labor economics after the late 1970s. Because of the large individual variation in outcomes, including additional ancillary variables in our equations in the hopes of altering the estimated impacts of the variables of interest is less worthwhile with the micro data that we increasingly use. Also, the size of the micro data sets generally means that adding a variable that we think might be important for some sample respondents is not 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 that allow us to study the particular novelty 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.[18] The low cost of applying 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 kernel estimation or competing-risks hazard models are applied,

or if some (usually unspecified) heterogeneity can be accounted for. 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 in those cases the careful labormetrician is obligated to examine first whether the relationships of interest are apparent in cross-tabulations or perhaps in simple regressions. If they do not exist there in large sets of micro data, the sophisticated techniques required to elicit them may very well be generating inappropriate inferences.

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. This multiplied credit makes sense, for, as 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, p. 202)]

VIII. Presenting Results -- Light Out from Under a Bushel

In presenting the results of our research, searching for statistical significance -- "95-percent confidence interval fetishism," should not be our goal.[19] Even if our test is powerful and generates 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 likely to be 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 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 (e.g., $\text{Citations}^2/100$ in Table 3). 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 impact of indicator variables or others, the reader should be able to tell what the variable is. 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 presenting probit parameters associated with some variable X, better to present an estimate of $\partial \Pr\{Y=1\}/\partial X$. With ordered probits (so long as the number of categories is small), 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.

Many authors still insist, and many journals still allow them to present parameter estimates with sequences of asterisks, daggers, crosses, crosses of Lorraine, and the like to denote significance levels. Where the test being applied is unfamiliar this makes sense; but where, as in most cases, these marks annotate t-statistics or standard errors describing regression-type parameters, they are quite superfluous: I am sure that readers of labormetric articles know the significance tables for the distribution of Student's t. 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; 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); and the null hypothesis is not always that the parameter equals zero.

We often list estimates of parameters relating each of a vector of indicator variables X to Y, with one of the components of X arbitrarily excluded. The reader can interpret results more easily if the parameter estimates are all of one sign, so that the estimates should be recomputed before publication to achieve this end. Indeed, except where there is some natural order to the categories (as with years of job tenure in Table 2, but not with occupation or industry) listing the categories in the order of the indicators' effects on Y facilitates interpretation.

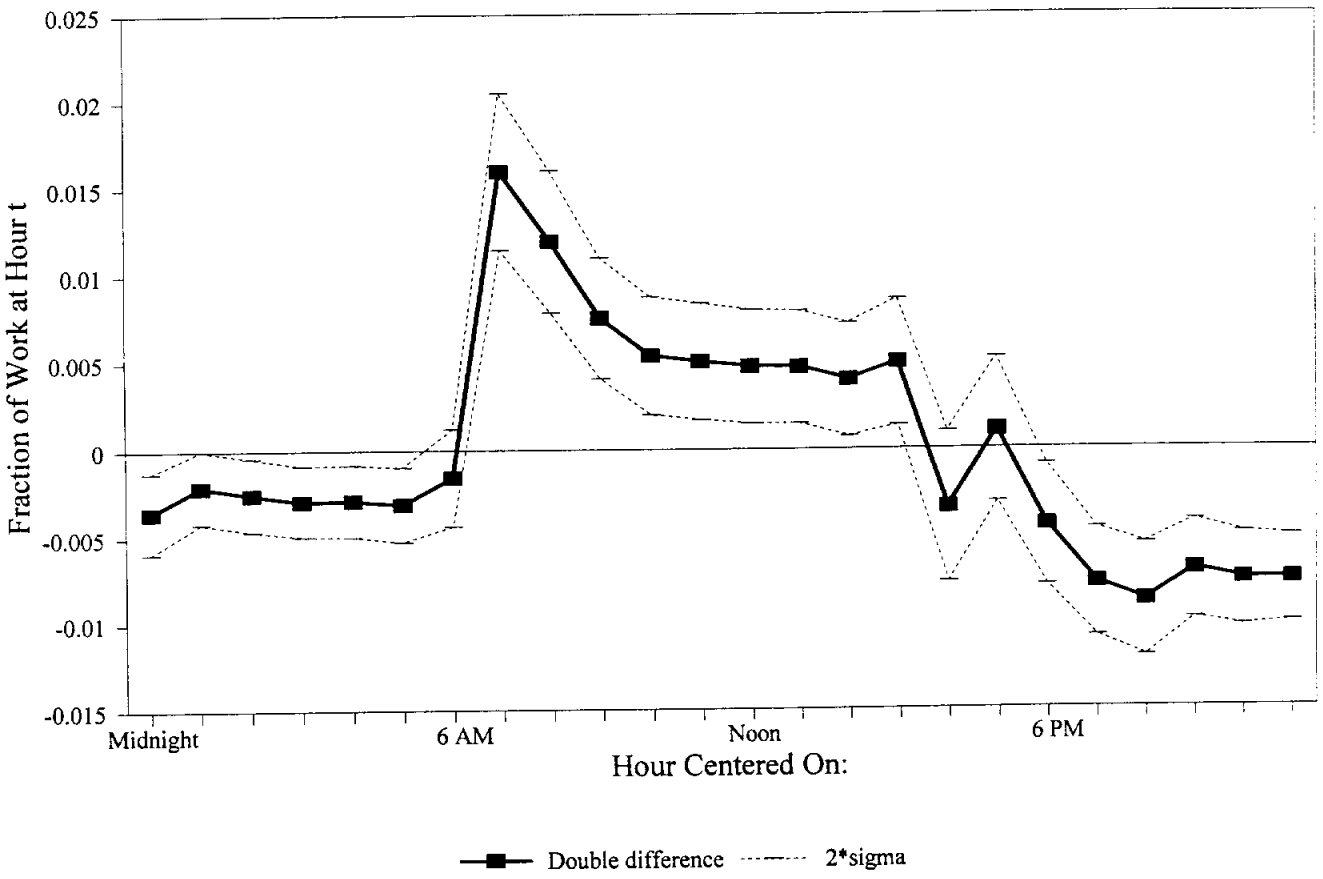
We should always aim to present our findings in such a way as to provide the clearest and strongest impression of what we have discovered. Labormetricians have always published their results (and sometimes their data too) in tabular form. Since the late 1980s, however, there has been a boom in presenting both using graphics.[20] No doubt this arose from the easy access to computer graphics in spreadsheet and statistical programs on personal computers and the ability of high-resolution personal printers to reproduce the figures. When should one follow this trend and use graphics instead of tabular presentation?

A simple answer is to read Tufte (1983) and follow its guidelines. Specifically, the increased ease of using graphical presentations has occurred simultaneously with the expansion of our ability to generate masses of results on relationships in large data sets. Thus 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 graphically. For example, the results of estimating hazards are almost always much

more clearly presented as graphs. As a related example, consider Figure 1 (taken from Hamermesh 1999), which shows estimates of the double-differences between the fraction of work done at each particular hour of the day by men in the top quartile of the earnings distribution and those in the bottom, comparing 1991 to 1973. The estimates could have been tabulated as 24 double-differences with the associated standard errors. I am convinced that the graph illustrates much more clearly the salient result, that the burden of evening and night work shifted toward the bottom of the earnings distribution, while day work shifted toward the top.

As with 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 their content can be presented more concisely verbally; and 2) Do not create graphs that are so cluttered with written descriptions as to obfuscate the message of the data. The technology has also led to overkill. If, for example, we estimate some time-series pattern for each of a large number of industries, the reader need not be assaulted with pages of graphs depicting each set of results; a few typical ones, plus a footnote referring to the others, suffice. Similarly, repetition of graphics showing the results of simulations describing how one's results would be altered by a large variety of small changes in policy parameters has remarkably soporific effects that detract from the message of one's results.

Figure 1. Work Time Double Difference
1991 - 1973, Top-Bottom, Men



IX. 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 enable me to estimate the appropriate multinomial logit. In 1994 an article of mine was returned favorably from a major general journal, but with the admonition that I should replace the ordered-probit estimation of educational attainment (which the data classified into intervals) by least squares. I complied with the request.

In both cases I was wrong. Spending so much effort on what was for the 1970s sophisticated econometric wizardry was a poor way to allocate time for someone who is basically an economist. No reasonable person could have expected that using the more esoteric technique would produce much different results from those generated by simpler techniques, and my time would have been better spent trying to understand the economics of the behavior I was studying. Obversely, complying with the editor's request in 1994 represented a step backward from the frontier of knowledge and detracted from the story that the article had to tell. By the late 1990s the plethora of accessible statistical packages made it easy to use such techniques; and expunging results was foolish in the light of most readers' fluency with them.

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 general economic behavior. Sophisticated techniques can enhance description and shed additional light on behavior; but they should be pursued only to the point where their time cost does

not detract from improving the research along the margins of enhancing the quality of the underlying data and of thinking about the economic relationships that generate the outcomes. Even before we resort to wizardry we should assure ourselves that we are not confusing things by making mistakes with simpler techniques. And, since the sophisticated technique is too often the attraction in its own right, we should make very sure that it is apropos the research question we are studying.

REFERENCES

Agee, Mark, and Thomas Crocker. 1996. "Parental Altruism and Child Lead Exposure." Journal of Human Resources, Vol. 31, No. 3 (Summer), pp. 677-691.

Ahituv, Avner, Joseph Hotz, and Tomas Philipson. 1996. "The Responsiveness of the Demand for Condoms to the Local Prevalence of AIDS." Journal of Human Resources, Vol. 31, No. 4 (Fall), pp. 869-897.

Akerlof, George, and Janet Yellen. 1985. "Unemployment through the Filter of Memory." Quarterly Journal of Economics, Vol. 100, No. 3 (August), pp. 747-774.

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. 1015-1042.

----------, and Bruce Meyer. 1999. "Using a Natural Experiment to Estimate the Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims and Denials." Journal of Public Economics, forthcoming.

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. 338-375.

Biddle, Jeff, and Daniel Hamermesh. 1990. "Sleep and the Allocation of Time." Journal of Political Economy, Vol. 98, No. 5, pp. 922-943.

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. 409-423.

Bound, John. 1989. "The Health and Earnings of Rejected Disability Insurance Applicants." American Economic Review, Vol. 79, No. 3 (June), pp. 482-503.

----------, 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. 443-450.

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. 531-564.

Caballero, Ricardo, Eduardo Engel, and John Haltiwanger. 1997. "Aggregate Employment Dynamics: Building from Microeconomic Evidence." American Economic Review, Vol. 87, No. 1 (March), pp. 115-137.

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. 957-979.

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. 819-862.

Dusansky, Richard, and Clayton Vernon. 1998. "Rankings of U.S. Economics Departments." Journal of Economic Perspectives, Vol. 12, No. 1 (Winter), pp. 157-170.

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. 605-637.

Evans, David, and Linda Leighton. 1995. "Retrospective Bias in the Displaced Worker Surveys." Journal of Human Resources, Vol. 30, No. 2 (Spring), pp. 386-396.

Evans, William, and Edward Montgomery. 1994. "Education and Health: Where There's Smoke There's an Instrument." National Bureau of Economic Research, Working Paper No. 4949, December.

Genesove, David, and Christopher Mayer. 1997. "Equity and Time to Sale in the Real Estate Market." American Economic Review, Vol. 87, No. 3 (June), pp. 255-269.

Gronau, Reuben. 1974. "Wage Comparisons: A Selectivity Bias." Journal of Political Economy, Vol. 82, No. 6 (November-December), pp. 1119-1143.

Grubel, Herbert, and Dennis Maki. 1976. "The Effect of Unemployment Benefits on U.S. Unemployment Rates." Weltwirschaftliches Archiv, Vol. 112, No. 2, pp. 274-299.

Hamermesh, Daniel. 1982. "Social Insurance and Consumption: An Empirical Inquiry." American Economic Review, Vol. 72, No. 1 (March), pp. 101-113.

----------. 1983. "New Measures of Labor Costs." In Jack Triplett, ed., The Measurement of Labor Cost. Chicago: University of Chicago Press, pp. 287-308.

----------. 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.

----------. 1999. "The Timing of Work Time Over Time," Economic Journal, Vol. 109, No. 1 (January), pp. 1-30.

----------, and Jeff Biddle. 1994. "Beauty and the Labor Market." American Economic Review, Vol. 84, No. 5 (December), pp. 1174-1194.

----------, and Paul Menchik. 1987. "Planned and Unplanned Bequests." Economic Inquiry, Vol. 25, No 1 (January), pp. 55-66.

----------, and Stephen Trejo. 1997. "The Demand for Hours of Labor: Direct Evidence from California." National Bureau of Economic Research, Working Paper No. 5973, March.

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, pp. 475-492.

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. 556-572.

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.

Klerman, Jacob, and Arleen Leibowitz. 1997. "Labor Supply Effects of State Maternity Leave Legislation." In Francine Blau and Ronald Ehrenberg, eds., Gender and Family Issues in the Workplace. New York: Russell Sage Foundation, pp. 65-85.

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." National Bureau of Economic Research, Journal of Business and Economic Statistics, Vol 13, No. 2 (April), pp. 151-162.

Moffitt, Robert. 1997. "Comment on Stephen Dynarski and Jonathan Gruber, 'Can Families Smooth Variable Earnings?'" Brookings Papers on Economic Activity, pp. 285-292.

Mueller, Richard. 1997. Essays on the Canadian Labor Market. Ph.D. diss., University of Texas--Austin.

Parsons, Donald. 1980. "The Decline in Male Labor Force Participation." Journal of Political Economy, Vol. 88, No. 1 (February), pp. 117-134.

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. 107-148.

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. 667-683.

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

Voos, Paula, and Lawrence Mishel. 1986. "The Union Impact on Profits in the Supermarket Industry." Review of Economics and Statistics, Vol. 68, No. 3 (August), pp. 513-517.

Whittington, Leslie, and James Alm. 1997. "'Til Death or Taxes Do Us Part'." Journal of Human Resources, Vol. 32, No. 2 (Spring), pp. 388-412.

Table 1. LS and LAD Estimates from Four Data Sets

	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 -662 80512; 115021		-626 (315)
		(187)	
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)

Table 2. The Impact of Looks on Men's Earnings Moderated Through Job Tenure, QES 1977[a]

Variable:	**(4)**	**(5)**
Below-average looks	------	-.168
		(.079)
Below-average looks and:		
Tenure <1	-.159	.009
	(.101)	(.129)
Tenure 1-3	-.137	.030
	(.100)	(.127)
Tenure 3-10	-.193	-.026
	(.093)	(.122)
Above-average looks	------	-.001
		(.064)
Above-average looks and:		
Tenure <1	-.008	-.006
	(.074)	(.094)
Tenure 1-3	-.006	-.005
	(.074)	(.098)
Tenure 3-10	.069	.071
	(.063)	(.090)

[a]Each equation also contains the three main effects of the tenure indicators and a large number of other controls, including age-schooling-6 and its square.

Table 3. 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.

FOOTNOTES

[1] In the latter case successive moves from the simplest wage measure to closer approximations to the marginal labor cost affect the parameter estimates in sensible ways (Hamermesh 1983).

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

[3] Even checking the descriptive statistics of the variables of interest may not suffice: One must go behind them to consider other characteristics of the underlying observations. In some recent work I was constructing data on couples from CPS samples from various years in the 1970s. Although I wanted husband-wife couples only, an examination of the sex and family status of respondents showed that roughly 0.6 percent of the households consisted of one head and two spouses, or two heads and two spouses. An examination of state of residence showed that only 6 percent lived in states with a high concentration of Mormons; and it is hard to believe there was that much communal living, even in the 1970s. I deleted all these observations from the sample.

[4] 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 from Dusansky and Vernon (1998). A lower number denotes a higher ranking.

[5] A very careful comparison using all these approaches is provided by Anderson and Meyer (1999).

[6] This is not always done, e.g., Voos and Mishel (1986). In one equation in which the union status of a firm and the concentration ratio in its market are included as main effects, the former produces a negative impact on supermarkets' profits. The authors use a second equation from which the main effect of unionism has been deleted, but to which an interaction of union status and concentration has been added, to infer that the negative coefficient on the interaction implies that "unions' ... impact is greater when local markets are less competitive." This conclusion is probably incorrect, and certainly cannot be inferred from the regression on which it is based. This error is committed, although with much less serious consequences, in the lead article in a recent issue of the most widely read economics journal (Genesove and Mayer 1997, Tables 5 and 7).

[7] This too is not always done, e.g., Agee and Crocker (1996, Table 2, col. 4), although in that case a comparison of the interactions with the linear term when no quadratic main effects are included suggests the failure has only a minor impact.

[8] One might reasonably argue, however, that beauty is a proxy for such skills as the ability to inspire other workers and to induce cooperation, so that it is productive even for long-tenure workers.

[9] I did, however, find three instances just by perusing the most recent four years of issues of a leading labor journal.

[10] For example, in an otherwise clever and useful study Whittington and Alm (1997, Tables 2 and 5) find significant positive α_1 and negative α_2 relating the probability of divorce to education. They state, based on the estimate of α_1, that education raises the probability of divorce. In fact, at the mean

education the effect is in most cases small and negative, and it is probably (because one cannot calculate the standard error of $\partial Y/\partial X$ without $Cov(\alpha_1\alpha_2)$, which is not published) insignificant.

[11] 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).

[12] 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 that other things may very well change differentially. At the very worst, only in the highly unlikely event that one has chosen a control group perfectly and has reproduced laboratory conditions will conditioning on the X will be nugatory.

[13] One recent year's editions of the Journal of Labor Economics and the Journal of Human Resources contained seven articles that employed this technique.

[14] The use of individual-effects estimators is less frequent than 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 these techniques.

[15] The problem is insurmountable when the variable of interest in X takes a unique value for each unit (or group of units) i, for example, when one appends values for geographic units to observations on individuals and wishes to hold other geographic variation constant (e.g., Ahituv et al 1996).

[16] This is not a matter of measurement error. The compensation and demographic data are from administrative records, and the citations data are carefully collected from published volumes.

[17] The rate of nonparticipation among men 45-54 rose from 3.5 percent of the population in 1955 to 7.9 percent in 1975 (Parsons 1980, p. 132). The percentage on DI grew from 0 to 3.9 percent (Bound 1989, p. 483).

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

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

[20] Of empirical articles in the three leading American labor journals (Industrial and Labor Relations Review, Journal of Human Resources and Journal of Labor Economics) 15 percent of the 113 published in regular issues in 1988 and 1989 included graphs of results or data. Of the 188 published in 1995 and 1996 34 percent included graphs.