# Essays on the Econometric Analysis of Panel Data

A thesis presented

by

Keisuke Hirano

to

The Department of Economics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Economics

Harvard University
Cambridge, Massachusetts

July, 1998

# Acknowledgments

I will always be indebted to Gary Chamberlain, my thesis advisor, for his guidance, insight, and patience. Working with him has been a truly amazing intellectual experience. Guido Imbens encouraged me to pursue econometrics at the beginning of my graduate career, and has been a constant source of ideas and motivation. I am grateful to Caroline Hoxby for many helpful discussions, which often helped me see things from a different and more productive point of view. Donald Rubin taught me a lot about identification and inference, and I have benefited enormously from having the chance to work with him. Portions of this thesis were cowritten with Gary Chamberlain, Guido Imbens, Geert Ridder, and Donald Rubin; I thank them for permission to include this joint research as part of my dissertation.

Cheri Minton and Nancy Williamson helped me work with the Panel Study of Income Dynamics data set used in the first two chapters of the thesis. I also received helpful comments from Katherine Baicker, Nancy Beaulieu, Patrick Bajari, Siddhartha Chib, Alan Durrell, Lawrence Katz, Ellen Meara, Sendhil Mullainathan, Peter Müller, Daniele Paserman, Jack Porter, and Jonathan Wright, and participants at seminars at Harvard and at the Conference on Bayesian Nonparametrics in Belgirate, Italy. I also wish to thank the Traymore Institute, especially Rory MacFarquhar, Katharine Park, Peter Steinberg, and Nicholas Weiss.

The National Science Foundation, the Department of Economics at Harvard University, and the Harvard Graduate Student Council provided generous financial support while I

was working on these chapters. As a graduate student I was fortunate to be able to serve as a teaching fellow at Harvard University. It was an enjoyable experience, and helped me clarify my thinking about econometrics in many ways.

I would especially like to thank my parents, my brother Takuya, and Martha Few, for their support and forbearance. This dissertation is dedicated to my parents.

*To my parents*

# Abstract

This dissertation addresses econometric inference in panel data. The first essay, "Predictive Distributions based on Longitudinal Earnings Data," uses longitudinal data on individual labor earnings from the Panel Study of Income Dynamics (PSID) to generate predictive distributions for an individual's future earnings. We develop models for dynamic earnings data which allow the variance of earnings shocks to differ across individuals. There turns out to be considerable evidence supporting this form of heterogeneity, which has important implications for the shape of the predictive distributions. An individual with a very volatile earnings history has predictive distributions that are very disperse, in comparison to an individual whose earnings history has been relatively stable.

The essay "Semiparametric Bayesian Models for Dynamic Earnings Data" develops Bayesian semiparametric extensions of commonly used random effects autoregressive models and error components models, which allow model errors to have a flexible distribution. Using data from the PSID, the predictive distributions arising from the more general models are heavy-tailed and asymmetric, and respond quite differently to large shocks to earnings than predictive distributions from conventional models based on normality.

"Optimal Consumption with Estimated Earnings Processes" extends the results of the preceding essay on semiparametric earnings models, by examining the implications of the some of the new models for earnings for optimal consumption smoothing. It develops numerical solutions to stylized optimal consumption problems, and compares solutions under different specification for labor earnings. The semiparametric specification for earnings

yields noticeably different optimal consumption policies than a more conventional analysis based on normal errors.

"Combining Panel Data Sets with Attrition and Refreshment Samples" addresses the problem of attrition in longitudinal surveys, and shows how *refreshment samples*, additional units drawn to replace units who have dropped out of the survey, can improve the quality of inference. It is shown that more general models of selection are identified when the refreshment samples are used, and that even under conventional models of attrition, more precise inference can be obtained using the refreshment samples. These points are illustrated in an empirical application using data on transportation usage.

# Contents

# Introduction

Panel data is data with a hierarchical or grouping structure. A special case is longitudinal data, in which some units of observation, such as individuals, firms, or nations, are followed over a number of time periods. Panel data analysis plays an important role in modern econometric methodology, because it is often possible to take advantage of the grouping structure to address substantive economic questions more completely than is possible with simpler forms of data. In particular, the grouping structure can be used to estimate models with complicated forms of heterogeneity across units. An example is an earnings model in which individuals have different "permanent" levels of income, perhaps arising from unobserved differences in ability. At the same time, working with panel data requires care to ensure that the techniques used are appropriate to the problem at hand. With heterogeneity, an additional unit of data is at best only partially informative about the model parameters common to all units. Practical empirical analysis using panel data must balance heterogeneity with some statistically meaningful notion of commonality and information accumulation.

Recently, there has been renewed interest in likelihood methods for panel data, in part because new computational algorithms, such as Markov chain Monte Carlo methods,

make it possible to apply likelihood-based inference to a much wider range of models. However, since economic theory is rarely informative about the parametric form of the likelihood function, model specification becomes especially important. One approach is to take advantage of advances in numerical methods to work with increasingly rich, flexible parameterizations. This is the path taken in Chapters 1 and 2, which take advantage of new numerical methods to extend conventional likelihood models for earnings dynamics. In addition, these chapters emphasize the role of predictive distributions as the appropriate goal of econometric inference. Focusing on predictive distributions allows for comparisons between models on the basis of their implications for decision-making. Chapter 3 then reexamines some of the models developed in Chapter 2 from an explicitly decision-theoretic point of view, building on economic models of consumption smoothing.

Another problem that can arise in the analysis of panel data is that missing data problems can be severe, as units who initially responded to a longitudinal survey may drop out over time. These selection and survey design issues are the concern of the final chapter.

## Predictive Distributions based on Longitudinal Earnings Data

The first chapter, written jointly with Gary Chamberlain, uses longitudinal data on individual labor earnings from the Panel Study of Income Dynamics (PSID) to generate predictive distributions for an individual's future earnings.

To make the goals of the empirical analysis more precise, we suppose that the individual faces a stylized consumption problem, based on recent work by Skinner (1988), Caballero

(1990), Deaton (1991), Hubbard, Skinner, and Zeldes (1994, 1995), and Carroll (1997), among others. In every period the individual receives some labor earnings, and pays out-of-pocket medical expenses. Labor earnings less medical expenses are added to any accumulated financial wealth to form cash on hand. Decisions about how much of current cash on hand to consume, and how much to invest in various assets, are made using predictive distributions for future earnings, as well as predictive distributions for asset returns, and other expenses. The agent can contemplate various possible actions, and for each action calculate an expected utility based on the predictive distributions. We want to provide appropriate forecasts of future earnings that the individual, or a financial planner, could use to guide savings and other decisions. This requires us to decide how to combine data on other people with the individual's own earnings history to form predictive distributions for future earnings, by specifying a joint distribution for observed and future earnings.

We develop models for dynamic earnings data which have the novel feature that the variance of earnings shocks can differ across individuals. The model assumes that log income is equal to the sum of a permanent individual-specific intercept term, an autocorrelated term, and a white noise term. The variance of the white noise term is allowed to be heterogeneous across individuals, and its distribution is parametrically specified. In the PSID data, there turns out to be considerable evidence supporting this form of heterogeneity. The predictive distributions that result from our new models differ in interesting ways from the predictive distributions that arise from more conventional models. An individual with a very volatile earnings history has predictive distributions that are

very disperse, in comparison to an individual whose earnings history has been relatively stable. This result holds up in various extensions of the main model that are considered and implemented.

## Semiparametric Bayesian Models for Dynamic Earnings Data

The second chapter also develops new models for dynamic panel data and applies them to study longitudinal data on earnings from the Panel Study of Income Dynamics (PSID). In this chapter, Bayesian semiparametric versions of commonly used random effects autoregressive models and error components models are introduced and studied. Recent advances in the theory and computation of nonparametric Bayesian models using Dirichlet process priors are employed to model the unknown distributions without having to rely on strong parametric assumptions.

As in the first chapter, a key motivation for developing these earnings dynamics model is to use them to specify optimal consumption problems. The recent work on these problems, discussed above, has made strong parametric assumptions in order to obtain complete probability distributions for the earnings process. For example, with longitudinal data on earnings, classical semiparametric methods could be used to estimate parameters defined in terms of the second moments of the data-generating process for earnings (see e.g. MaCurdy (1982), Chamberlain (1984), Holtz-Eakin, Newey, and Rosen (1988)), but they do not provide estimates of the distribution of model errors. In practice, the recent work on optimal consumption has assumed log normality for any unknown distributions, largely erasing the potential benefits of semiparametric inference. There was evidence of

heavy tails relative to log normality in the results of Chapter 1, so we wish to explore this issue further.

Our proposed semiparametric Bayesian models deliver inference for the entire earnings process, while allowing us to remove the restriction that all components of the model belong to known parametric families. The approach developed in this chapter has a modular nature, and can be readily extended to even richer models, the primary cost being computational difficulty rather than analytic intractability or concerns about the adequacy of asymptotic approximations. There turns out to be strong evidence against normality in the earnings data, suggesting that consumption models which rely on this assumption should be viewed with caution. As in the preceding chapter, we emphasize the use of these new models to generate predictive distributions for an individual's future earnings. The predictive distributions arising from the more general models are heavy-tailed and asymmetric, and respond quite differently to large shocks to earnings.

Two key difficulties encountered in nonparametric and semiparametric Bayesian analysis are the need to specify prior distributions appropriately, and the complexity of some of the necessary calculations. In our application, we will see that the first issue corresponds very closely to the problem of choosing "smoothing parameters" in classical nonparametric and semiparametric methods. We regard the device of a prior distribution as a rich, flexible way to make such a priori judgments of smoothness, and will examine this issue in detail when we study the earnings data. While the computations remain difficult, they are now feasible by taking advantage of Markov chain Monte Carlo integration methods.

## Optimal Consumption with Estimated Earnings Processes

In Chapter 2, we found that relaxing the assumption of normality dramatically changes predictive distributions for future earnings. For example, using an AR(1) specification with random effects and general form for the error, the error distribution was estimated to be much more heavy-tailed than the normal distribution. In some cases, the semiparametric estimates of the autoregressive coefficients were much higher than the estimates under normality. In Chapter 3 we attempt to evaluate the consequences these differences could have for stylized models of optimal consumption behavior. The goal is to develop alternative, economically motivated measures for comparing statistical models for earnings. We develop solutions only in very simple cases; numerical solutions could be extended to more realistic situations, although the computations can quickly become very expensive.

We begin by setting up a dynamic consumption problem along the lines of the problem that motivated the previous two chapters. A consumer faces uncertainty about future earnings and uses a riskless asset to carry wealth into future periods and smooth fluctuations in income. Before attempting numerical solutions we look at some simple cases, for which closed-form solutions can be obtained. In the first case, utility has a logarithmic form and income can be insured against, so that the consumer consumes out of a predetermined endowment. The second case allows for uninsurable risky income, but assumes that income in each period is i.i.d., and the consumer has exponential utility. This allows an explicit expression to be derived for the value function in every period.

Next we bring in autocorrelated earnings, liquidity constraints and general isoelastic utility. Dynamic programming is used to obtain optimal consumption policies in a finite

horizon model, where the value function in the terminal period could be interpreted as giving continuation payoffs. We compare optima using the parametric and semiparametric correlated random effects models for college-educated male heads of household as the specification for the earnings process. The loss from using the "wrong" earnings dynamics model is calculated in certainty equivalent terms.

In the second set of numerical exercises, based on the work of Smith (1991), we consider a limited family of consumption policies, and try to find the best consumption policy within this restricted set. Looking only at a limited set of "rules of thumb" allows us to deal with longer horizons quite easily, and also to use predictive distributions that incorporate parameter uncertainty. In essence, we simply use the predictive draws from the previous chapter to perform Monte Carlo integration for expected utilities under a given decision rule.

The two models for earnings yield noticeably different optimal consumption policies. This seems to result mainly from the very different autoregressive coefficients estimated under the two assumptions about the model errors. That the estimated coefficients differ so greatly is in itself an empirical puzzle, so we reconsider this issue and try to provide an interpretation based on heuristic large-sample considerations.

## Combining Panel Data Sets with Attrition and Refreshment Samples

Although panel data can answer many substantive questions that cross-sectional data cannot, missing data problems can be more severe in panels. One important problem is that units which respond in initial waves of a longitudinal survey may drop out of the sample in subsequent waves, so that the subsample with complete data for all waves of the panel can be less representative of the population than the original sample (e.g., Hausman and Wise (1979), Ridder (1990), Verbeek and Nijman (1992), Abowd, Crepon, Kramarz, and Trognon (1995), and Vella (1998). An example in which using only the complete-data subsample can lead to very different substantive conclusions is given by Olley and Pakes (1996).

Sometimes, in the hope of mitigating the effects of attrition without losing the fundamental advantages of panel data over cross-sections, panel data sets are augmented by replacing units who have dropped out with new units randomly sampled from the original population. Ridder (1992) calls such additional samples *refreshment samples*. Data sets that include a refreshment component include the Dutch Transportation Panel analyzed in this paper, and the Health and Retirement Survey. In addition, *rotating panels* such as the Current Population Survey can be interpreted as including a refreshment sample every period.

The final chapter, written jointly with Guido Imbens, Geert Ridder, and Donald Rubin, explores the benefits of refreshment samples for inference in the presence of attrition.

The two themes of the chapter are, first, that refreshment samples can improve inference under conventional models of attrition in panel data, by providing additional sample information, and second, that refreshment samples allow for identification of more general models of attrition, without requiring auxiliary assumptions on distributions of the response variables. Thus, refreshment samples are potentially a relatively inexpensive way to improve the quality of longitudinal surveys.

We begin by setting up a framework to describe the data and the inferential problem. In Section 3 of the chapter, we describe two conventional models for attrition in panel data. The first model, essentially based on the *missing at random* assumption (MAR, Rubin (1976), Little and Rubin (1987)), allows the probability of attrition to depend on lagged but not on contemporaneous variables. This case is sometimes also referred to as *selection on observables* (Moffitt, Fitzgerald, and Gottschalk 1997). The second model allows the probability of attrition to depend on contemporaneous, but not on lagged, variables. We refer to this model as the HW model, because it originates in work by Hausman and Wise (1979). Related models have also been referred to as models with *selection on unobservables* (Moffitt, Fitzgerald, and Gottschalk 1997) because attrition partly depends on variables that are not observed when the unit drops out. In Sections 4 and 5 of the chapter we present the main theoretical contributions. We develop a model for attrition that includes the MAR and HW models as special cases. This Additive Nonignorable (AN) model is identified, with no testable implications, from a panel data set with a refreshment sample. We first discuss in Section 4 the identification issues in a simple two-period context with a single binary variable in both periods, and generalize the

model to allow for multi-valued variables as well as time-invariant covariates in Section 5.

Section 6 applies the ideas developed in earlier sections to a panel data set on travel behavior in the Netherlands, the Dutch Transportation Panel (DTP). This data set is based on a survey of Dutch households concerning their use of various modes of transportation. For a number of years households were asked to keep a detailed travel diary for an entire week each year. For every trip taken by a household member detailed information was gathered including destination, time, and mode of transportation. Attrition was severe, and because the considerable effort required to respond to the survey was directly related to the value of one of the variables (total number of trips taken by household members), it is plausible that among those who responded in the first wave the willingness to cooperate in the second wave depended on the value of these variables in either the first or second period. Because MAR essentially rules out dependence on second period variables, and the HW model rules out dependence on first period variables, this makes for a potentially interesting comparison of the performance of the MAR and HW models. In this application we implement the models by multiply imputing the missing data according to the various missing data models, and compare the results both in terms of estimates of the relation of the change in the number of trips to the change in income as well as in terms of estimates of the attrition process itself. Imputation has the advantage over joint estimation of the model for attrition and the substantive model of interest that it does not require the researcher to adapt the estimation procedures to allow for missing data. Given the complex nature of modern methods for estimating panel data models, including non-differentiable objective functions (e.g., Honore (1992)) and kernel methods

(e.g., Kyriazidou (1997)), this can very useful in practice.

## Conclusion

These essays represent initial steps along a set of interrelated lines of research: practically useful likelihood inference in large parameter spaces; model choice from a decision-theoretic point of view; the central role of predictive distributions in statistical inference; econometric identification without strong parametric assumptions; and survey design for economic data. Advances in econometric computing play a role, because they expand the options available for empirical practice. But to a considerable degree, the basic themes can be separated from these technical concerns, and there is every indication that they will continue to provide stimulating questions for future research.