Heterogeneity and Microeconometrics Modelling

Browning, Martin; Carro, Jesus

Publication date:
2006

Document version
Publisher's PDF, also known as Version of record

CAM

Heterogeneity and Microeconometrics Modelling

Martin Browning and Jesus Carro

2006-03

Heterogeneity and microeconometrics modelling.[*]

Martin Browning

Center for Applied Microeconometrics (CAM)

Department of Economics, University of Copenhagen

Jesus Carro

Department of Economics, Carlos III, Madrid

January 2006

Abstract

Presented at the 2005 Econometric Society World Congress Plenary Session on "Modelling Heterogeneity". We survey the treatment of heterogeneity in applied microeconometrics analyses. There are three themes. First, there is usually much more heterogeneity than empirical researchers allow for. Second, the inappropriate treatment of heterogeneity can lead to serious error when estimating outcomes of interest. Finally, once we move away from the traditional linear model with a single 'fixed effect', it is very difficult to account for heterogeneity and fit the data and maintain coherence with theory structures. The latter task is one for economists: "heterogeneity is too important to be left to the statisticians". The paper concludes with a report of our own research on dynamic discrete choice models that allow for maximal heterogeneity.

1   Introduction.

There is general agreement that there is a good deal of heterogeneity in observed behaviour. Heckman in his Nobel lecture (Heckman (2001)) states: " the most important discovery [from the widespread use of micro-data is] the evidence on the pervasiveness of heterogeneity and diversity in economic life". This is true but to see it in print as a 'discovery' is a surprise since we have internalised it so thoroughly and it is now second nature for anyone working with micro data to consider heterogeneity.

We have been unable to find a consensus definition of heterogeneity. A definition we suggest (which derives from Cunha, Heckman and Navarro (2005)) is that *heterogeneity* is the dispersion in factors that are relevant *and known* to individual agents when making a particular decision. *Latent heterogeneity* would then be those relevant factors that are known to the agent but not to the researcher. The heterogeneity could be differences in tastes, beliefs, abilities, skills or constraints. Note that this definition does not impose that heterogeneity is constant over time for a given individual nor that because something is fixed and varies across the population that it is necessarily heterogeneous. Examples of the former would be changing information sets and an example of the latter would be, say, some genetic factor which impacts on outcomes but which is unobserved by any agent. Thus a 'fixed effect' in an econometric model may or may not be consistent with heterogeneity, as defined here.

Our definition of heterogeneity distinguishes it clearly from uncertainty, measurement error and model misspecification that are other candidates for the variation we see around the predictions of a given deterministic model.[1] The conceptual distinction between heterogeneity and measurement error and model misspecification is obvious (although it may be difficult to distinguish in empirical work), so we concentrate on uncertainty. To illustrate the issue we present an example about milk consumption. Most of the world's adult population cannot drink more than one cup of milk in day without feeling quite ill (see Patterson (2001) for details and references). This is because they lack the enzyme lactase that breaks down the sugar in milk (lactose) into usable sugars. The inability to digest milk is known as lactose intolerance but we should more properly speak of lactase persistence for those who can drink milk since this is a relatively late adaptation in our evolutionary history. The ability to drink milk as an adult seems to have arisen at least twice independently, both times amongst pastoralists. This happened very recently, perhaps as

---

[1] This definition has one major drawback which is that the vernacular term 'heterogeneity' does not coincide with the analytical definition suggested here. In some ways it would be best to have a wholly new term for heterogeneity as defined here, but that seems impossible at this late date.

late as 6, 000 years ago. This has lead to considerable variation across the world in lactase persistence; for example, the rate is 97% in Denmark, 47% in Greece, 29% in Sicily, 8% amongst Han Chinese and 5% for the San in South Africa. In early childhood no one knows their type (lactose intolerant or lactase persistent). If every child within a population has the same beliefs, then there is no heterogeneity and only uncertainty. In adulthood, everyone knows their type. Then the variation observed in the population, at age 20 for instance, is due to heterogeneity, but there is no uncertainty. If people were ignorant of their type in childhood but had different beliefs, then there would be uncertainty and heterogeneity (in beliefs, not in their ability to digest milk). Distinguishing between heterogeneity and uncertainty is an important but difficult task; see Cunha *et al* (2005) for an analysis of this in the context of schooling choice.

Before going on it is necessary to mention a contrary view on heterogeneity in tastes which derives from Stigler and Becker (1977). They took the position that "tastes neither change capriciously nor differ importantly between people". Tastes for Becker and Stigler take as their domain commodities that are produced from market goods and time use or even deeper structures. Consider, for example, food. Market goods constitute the highest level of observability. Obviously, tastes over different foods differ; for example, tastes for milk as discussed in the previous paragraph. At the next level, tastes are defined over the characteristics inherent in food market goods; for example, different kinds of fats, calcium, vitamins etc. It is by no means obvious that tastes over characteristics differ significantly over time or place. The interest in this intermediate level is that we may be able to recover the food/nutrient conversion mapping from independent sources. This then makes the consumption of characteristics observable. This would allow us to test for various levels of heterogeneity. For example, does everyone in a given population have the same tastes? A weaker and more interesting hypothesis is that there is heterogeneity but the dispersion of tastes is the same across populations, conditional on demographic factors such as the age distribution. But even moving to characteristics may not be a deep enough level to allow us to support the hypothesis that everyone has the same tastes. People with different metabolic rates will have different tastes over second level characteristics such as nutrients. Thus saturated fat is not valued because it is saturated fat but because it provides an energy source that allows us to function in a satisfactory way. Similarly, vitamin B12 (which is not found in any vegetables) is a vital ingredient for brain functioning but vegans might be happy to have some alternative that allowed their brain to continue working once they have depleted the five year store that the body usually keeps as a buffer stock. In this

view, we may have to go to the deeper level of capabilities for the domain of common preferences.

Explicit in the Becker-Stigler view is the contention that an appeal to undefined heterogeneity in tastes is too ready an admission of ignorance. Certainly, the immediate use of 'fixed effects' or other heterogeneity schemes to allow for observable differences in outcomes is an admission of failure in economics (even if it sometimes considered a triumph in econometrics). The Becker-Stigler approach leads naturally to a research program that attempts to rationalize all observable differences in behaviour with no differences in tastes and only taking account of prices and incomes (or potentially observable constraints) and heterogeneity in the mapping from market goods to the domain of preferences which can potentially be identified from observable background factors. However, to assert homogeneity of tastes at the level of capabilities (as suggested in the last paragraph) is much less attractive for modelling since it seems to substitute one ignorance (of the mapping from market goods to capabilities) for another (the undifferentiated distribution of tastes over market goods).

This paper attempts to make three main points. First, we claim that there is a lot more heterogeneity about than we usually allow for in our microeconometric modelling. This first theme is covered in section 2. The illustrations we present are from our own work but we believe that the point is a very general one. Second, in most contexts it makes a big difference for outcomes of interest whether and how we allow for heterogeneity. We shall illustrate this as we go along. Our third main point is that it is difficult to allow for heterogeneity in a general way. This is particularly the case if we want to fit the data and be consistent with economic theory. This is discussed in section 3 which provides some examples. Our main contention there is that most schemes currently employed in applied microeconometrics are chosen more for their statistical convenience than for their fit to the data or their congruence with economic theory. We believe that heterogeneity is too important to be left to the statisticians and that we may have to sacrifice some generality for a better fit and more readily interpretable empirical estimates. In stating this we certainly do not want to suggest that all microeconometric analyses suffer from this problem. The struggle to find heterogeneity structures that also provide interpretable estimates is an old one and one that continues to inform much structural estimation to this day. A classic example is Mundlak (1961) who allows for heterogeneity in managerial ability for farmers and more recent examples are McFadden and Train (2000) for discrete choice models, Heckman, Matzkin and Nesheim (2003) for hedonic models and Laroque (2005) for models of labour supply allowing for heterogeneity in the

taste for work that is correlated with productivity.

In section 4 we present some results from our own work on dynamic binary choice models that allow for maximal heterogeneity. Since not much is known about the properties of estimators in this context, even when we only allow for conventional 'fixed effects', we choose to study in depth the simplest model, a stationary first order Markov chain model. Working with such a simple model we can recover analytical results concerning bias, mean squared error, the power of tests etc.. As we shall show, this allows us to develop novel estimators that are designed with our choice criteria (bias or mean squared error) in mind.

2   There is a lot of heterogeneity about.

2.1   Earnings processes.

In many contexts the evidence points towards more heterogeneity than is usually allowed for. We hazard a conjecture that in a majority of published empirical papers there is 'significantly' more heterogeneity than is allowed for in the modelling. We shall illustrate this with two examples from our own work. The first example is for a linear dynamic model for earnings processes. A close to consensus model is that once we control for common time effects, log earnings are a unit root with homogeneous short run variances, an MA error term and no drift:

$$\Delta y_{it} = \varepsilon_{it} + \theta \varepsilon_{i,t-1} \text{ with } \varepsilon_{it} \sim iid \left(0, \sigma^2\right) \tag{1}$$

for agent i. This model has two parameters (or more if we allow an $MA\left(2\right)$ process or an error distribution with more than one parameter). This model seems to be popular because it is the reduced form for a model which has a 'permanent income' component and also because it is believed that it fits the data well. Although very popular, other processes have also been considered. For example, trend stationary models which allow for a negative correlation between starting values and the trend. This captures Mincer style on the job training in which some workers trade off initial earnings for higher earnings later on; see Rubinstein and Weiss (2005). The important point about all of these analyses is that they typically assume very little heterogeneity once we condition on the starting value. Figure 1 shows 10 paths for white, high school educated males from age 25 to 35, drawn from the PSID.[2] The

---

[2]In line with convention in the earnings literature, these are actually the residuals from a first round regression (on a larger sample) of log earnings on time and age dummies. The larger sample is identical to that taken in Meghir and Pistaferri (2004); it is an unbalanced panel that covers the years $1968 - 1993$ and includes workers aged between 25 and 55.

subsample displayed is chosen so that all have close to the mean earnings at age 25. This figure suggests that the consensus model may not be adequate to fit the data, even when we have the same initial observation. There seems to be clear evidence of differences in volatility and also some visual evidence of differences in trends. On the other hand, even practiced eyes may not be able to tell drift from the working out of a unit root with only 10 observations, so more formal evidence is required.

Alvarez, Browning and Ejrnæs (2002) (ABE) consider a model with lots more heterogeneity than is allowed for in previous empirical analyses. They start with a parametric model that has six parameters per worker and in which all of the parameters are heterogeneous. To overcome the curse of dimensionality for this model they use a simulated minimum distance (or indirect inference) estimation approach that requires that the final model fits *all* of the outcomes of interest that previous researchers have suggested are important (and some other data features that have not been considered before). They find that the following stable (but not stationary[3]) four parameter model gives a good fit to the data:

$$y_{it} = \delta_i \left(1 - \beta_i\right) + \beta_i y_{h,t-1} + \varepsilon_{it} + \theta_i \varepsilon_{i,t-1} \text{ with } \varepsilon_{it} \sim iiN\left(0, \sigma_i^2\right) \tag{2}$$

Interestingly, unit root models are decisively rejected (but models with a mixture of a unit root and a stable process do quite well). The consensus model fits very poorly, contradicting the widespread belief that it fits well. The important point in the current context is that all four parameters $(\delta, \beta, \theta, \sigma)$ are heterogeneous. ABE find that the joint distribution of these four parameters is described well by a three factor model. One of these factors is the starting value and the other two are latent factors.[4] The preferred version of the general model has heterogeneity in all of the four parameters above (plus a parameter for an ARCH scheme).

A finding that conventional empirical models do not make adequate allowance for heterogeneity in parameters does not necessarily mean that they are significantly wrong for all outcomes of interest. To illustrate, we consider two outcomes of interest for earnings processes. The first is actually a parameter: the short run standard deviation, σ_i. This is a crucial input in models of consumption which allow for precautionary saving. Typically, the level of precautionary saving is an increasing

---

[3]We follow the terminology of Arellano (2003a) and say that a first order dynamic process is *trend stable* if the *AR* parameter is less than unity and the initial values are unrestricted. If the initial values are restricted to be consistent with the long run distribution then the process is stationary.

[4]This modelling methodology is an extension of the scheme suggested in Chamberlain (1980) in which the distribution of the individual parameters is allowed to be conditional on the starting values. See Wooldridge (2005) for post-Chamberlain references and a strong defence of this mode of modelling.

Percentile	10%	25%	50%	75%	90%
Equation (1) with heterogeneous variances	0.07	0.09	0.13	0.19	0.25
Equation (2)	0.05	0.07	0.10	0.15	0.21

Table 1: The distribution of the standard deviation of income residuals

and strictly convex function of the standard deviation of income risk.[5] It will be clear that allowing for heterogeneity in this parameter will impact on estimates of precautionary saving. Agents who are identical in every respect except for the income risk they face will hold different levels of 'buffer stocks' and the mean of the homogeneous model may be very different to that from a model with heterogeneous variances. For the consensus model (1) (with allowance for ARCH and measurement error) the estimate of the standard deviation for the zero mean Normally distributed error is 0.142; this gives that the probability of a large drop in earnings (20% or more) between any two periods is about 8%. Table 1 presents the distribution of the standard deviation for (1) with allowance for a heterogeneous variance and for (2). For the consensus model with heterogeneous variances (row 1) there is a great deal of heterogeneity in variances (as we would expect from figure 1). Many workers face low risk and would have virtually no precautionary motive. On the other hand, about 10% of workers have a standard deviation of over 0.25 which implies a probability of a large drop of about 21%; for this group the precautionary motive would be very strong. When we move to the fully heterogeneous model (the second row of Table 1) the standard deviation distribution is lower (because the error terms in the consensus model with variance heterogeneity have to 'mop up' the heterogeneity in the other parameters) but there is still considerable dispersion in risk. The lesson we draw from this is that if the primary interest is in the variance, then a simple model with allowance for heterogeneity in the variances would probably suffice but the homogeneous variances model is way off.

Our second outcome of interest is the distribution of lifetime earnings. Cunha, Heckman and Navarro (2005) have a discussion of the formidable problems in using empirical distributions such as these in modelling schooling decisions. In particular, they treat carefully the distinction between heterogeneity (what subjective distributions do young people have over the parameters) and uncertainty (the residual uncertainty given a model with a set of parameters) that motivated the definition of heterogeneity given in the introduction. Unlike the distribution of the error variances, the moments or quantiles of the lifetime earnings distribution are highly

---

[5] The leap from the error variance in (2) to risk is a large one - there is measurement error and the changes may be anticipated or even have been chosen - but it is often made in consumption modelling.

nonlinear function of the model parameters and it is impossible, *a priori*, to judge whether allowing for heterogeneity will make much difference. To generate the distribution of lifetime earnings we simulate $25,000$ paths from age 25 to 55 using the model (2), add back in the age effects taken out in the first round regressions and discount earnings back to age 25 with a discount rate of 3%. In figure 2 we present estimates of the trade-off between median and interquartile range based on the consensus model with heterogeneous variances, (as in the first row of Table 1), and the preferred model, (2). As can be seen, the trade-off is very close to linear and increasing for the consensus model with heterogeneous variances but nonlinear for the preferred model. In particular, the trade-off between median and interquartile range is much steeper for those who expect a relatively low median lifetime income. Whether or not these significantly different outcomes would translate into different estimates of, say, schooling choices would depend on the exact details of how we use these estimates, but there is at least the potential for serious error if we use the consensus model rather than the preferred model which allows for significantly more heterogeneity.

2.2   Dynamic discrete choice.

Our second example of the ubiquity of heterogeneity is for a dynamic discrete choice model for the purchase of whole (full fat) milk; see Browning and Carro (2005). The data are drawn from a Danish consumer panel which is unusual in that the panel follows a large and representative group of households over a long period. Specifically we consider weekly purchases of different varieties of milk and we observe each household for at least 100 weeks (and some for 250 weeks). After some selection to meet various criteria (such as buying whole milk in at least 10% and at most 90% of the weeks we observe) we have a sample of 371 households. The availability of such a long panel enables us to explore with real data the effectiveness of different heterogeneity schemes suggested for small-T panels. In section 4 we shall return to a detailed study of estimators in this context but for now we simply want to show that there is more heterogeneity in this choice than we would usually allow for. To do this, we use a dynamic Probit for each household:

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right) = \Phi\left(\eta_i + \alpha_i y_{i,t-1} + x_{it}'\beta\right) \tag{3}$$

where y_{it} is a dummy for household i buying whole milk in week t and x_{it} is a vector of covariates such as seasonal dummies, a trend and family composition variables. In this analysis we impose that the parameters for the latter are homogeneous but we

8

could let them be idiosyncratic for each household in which case we would treat the data as a collection of 371 time series. We do, however, allow that the AR parameter may vary across households. The usual approach is to impose homogeneity on this parameter:

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right) = \Phi\left(\eta_i + \alpha y_{i,t-1} + x_{it}'\beta\right) \tag{4}$$

Our interest here is in whether the latter is a reasonable assumption.

Before presenting results it is worth considering the role of the homogeneous AR parameter assumption in dynamic models generally. If we had a linear model:

$$y_{it} = \eta_i + \alpha_i y_{i,t-1} + \varepsilon_{it} \tag{5}$$

then the most common restriction to impose is that the marginal dynamic effect, in this case α_i, is the same for everyone. This is usually assumed more for econometric convenience than for its plausibility but it has the virtue of imposing a restriction on an object of interest. When moving to a nonlinear model the letter of this restriction is usually retained but the spirit is lost. Consider, for example the dynamic discrete choice model:

$$pr\left(y_{it} = 1 \mid y_{it-1}\right) = F(\eta_i + \alpha_i y_{it-1}) \tag{6}$$

where $F\left(.\right)$ is some parametric cdf. The marginal dynamic effect is given by:

$$M_i = F(\eta_i + \alpha_i) - F(\eta_i)$$

Imposing that the AR parameter α_i is homogeneous does not imply a homogeneous marginal dynamic effect. Furthermore, this restriction is *parametric* and depends on the chosen cdf. Thus assuming that $\alpha_i = \alpha$ for all i for one choice of $F\left(.\right)$ implies that it is heterogeneous for all other choices, unless α is zero. This emphasizes the arbitrariness in the usual homogeneity assumption since there is no reason why the homogeneity of the state dependence parameter α should be linked to the distribution of $F\left(.\right)$. In contrast, the homogeneous marginal dynamic effect assumption, $M_i = M$, that is the correct analogue of the linear restriction gives:

$$\alpha_i = F^{-1}\left(M + F\left(\eta_i\right)\right) - \eta_i \tag{7}$$

for some constant M. Thus the homogeneous AR parameter model is conceptually at odds with the same assumption for linear models. Although we believe the (7) assumption to be more interesting, we shall continue the analysis of the restriction in (4) since it is the conventional approach.

9

Figure 3 shows the marginal distributions (top panel) and the joint distribution (bottom panel) for the parameters (η, α) in equation (3). The two panels show two important features. First, both parameters display a lot of variability. Moreover, the heterogeneity in α is 'significant'; the formal LR test statistic for a homogeneous AR parameter is $3,058$ with 370 degrees of freedom. This is very strong evidence that the AR parameter is heterogeneous as well as the 'fixed effect', η. The second important feature is that the joint distribution is far from being bivariate Normal. There is evidence of bimodality and fat tails relative to the Normal. This suggests that the first resort to modelling the data, a random effects model with a joint Normal distribution, will not suffice to adequately model the heterogeneity.

Once again, we make a distinction between heterogeneity in parameters and in of objects of interest. For a dynamic discrete choice model there are two natural objects of interest: the marginal dynamic effect:

$$M_i = pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right) - pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right) \tag{8}$$

and the long run probability of being unity which for a first order Markov chain is given by:

$$
\begin{aligned}
L_i &= \frac{pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right)}{\left(1 + pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right) - pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right)\right)} \\
&= \frac{pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right)}{\left(1 + M_i\right)}
\end{aligned} \tag{9}
$$

Figure 4 show the marginal densities for M and L, with and without heterogeneity in the slope parameter. As can be seen, allowing for 'full' heterogeneity makes a great deal of difference. Thus the heterogeneity in the AR parameter is not only statistically significant but it is also substantively significant.

3   Heterogeneity is difficult to model.

3.1   The need to allow for heterogeneity.

The concern to allow for heterogeneity arises from one of two considerations. First the heterogeneity may not be of interest in itself (it is a 'nuisance') but ignoring it would lead to faulty inference for objects of interest. The latter could be qualitative outcomes such as the presence of state dependence in some process (see Heckman (1981)) or the presence of duration dependence in duration models (Lancaster (1979)). Even if ignoring heterogeneity did not lead to errors regarding qualitative

outcomes, it might lead to inconsistency in the estimation of parameters of interest. This has been one of the traditional concerns in modelling heterogeneity. The prime example of a scheme to deal with heterogeneity in linear models with strictly exogenous covariates is to assume that only the intercept is heterogeneous so that we can first difference the heterogeneity away. This highlights the tension between the desirability for generality (we do not have to assume anything about the distribution of the fixed effect once we assume that only the slope parameter is homogeneous) and the need to fit the data (the slope parameters might also be heterogeneous).

Even in the linear case, heterogeneous panel models are difficult to deal with. In a large T framework a few solutions have been proposed, most of them in the macroeconomics literature; for example for endogenous growth models using cross-country panels with long time periods. There, allowing for country-specific coefficients due to heterogeneity can be important, but it is out of the scope of this paper that focus on microeconometrics. Pesaran and Smith (1995) consider this problem and discuss how to estimate one of the possible parameters of interest from dynamic heterogeneous panel in that context. Once we move away from linear models or focus on outcomes other than actual parameter values, then heterogeneity becomes of importance in its own right. As a well known example, suppose we are interested in the marginal effects of an exogenous variable in a nonlinear model; this will usually depend on the heterogeneity directly. Thus if we have $y_{it} = G\left(\eta_i + \beta x_{it}\right)$ then the marginal effect for of a change in x of Δ for any individual is given by:

$$G\left(\eta_i + \beta\left(x + \Delta\right)\right) - G\left(\eta_i + \beta x\right) \tag{10}$$

which obviously depends on the value of η_i.

It is usually impossible to model allowing for unrestricted heterogeneity 'everywhere' and we have to make *a priori* decisions about how to include allowance for heterogeneity in our empirical models. A major disappointment in panel data modeling is that simple first differencing schemes that work for linear models do not work in nonlinear models. The classic example is limited dependent variable models, but the point is more pervasive. The result has been that we have developed a series of 'tricks' which often have limited applicability. Almost always decisions on how to include allowance for heterogeneity are made using conventional schemes that have been designed by statisticians to put in the heterogeneity in such a way that we can immediately take it out again. As just stated, the leading example of this is the use of a 'fixed effect' in linear panel data models. More generally, various likelihood factoring schemes have been suggested for nonlinear models; see, for example, Lindsey (2001), chapter 6. The most widely used of these is for the panel data discrete

11

choice model, due to Andersen (1970). Our main contention is that decisions about how to incorporate heterogeneity are too important to be left to the statisticians for two major reasons. First, these schemes may not fit the data. Indeed it is an extraordinary fact that the great majority of empirical analyses choose the heterogeneity scheme without first looking at the data. Second, conventional schemes, which are often presented as 'reduced forms', implicitly impose assumptions about the structural model. Usually it is impossible to make these assumptions explicit, so that estimated effects are effectively uninterpretable. One feature of this that will emerge in the examples below is issues of fitting the data and theory congruence arise whether or not the source of the heterogeneity is observed by the researcher.

The ideal would be to develop economic models in which the heterogeneity emerges naturally from the theory model. An alternative is look to other disciplines for structural suggestions; psychology ('personality types'); social psychology (for the effects of family background) or genetics. For example, psychologists suggest that personalities can be usefully characterized by a small number of factors. As an example, for economists looking at intertemporal allocation the relevant parameters are risk aversion, discount factor and prudence. We might want particular dependence between all three. For example, more risk averse people may be likely to also be more prudent.[6] Information on the nature of this dependence might be found in psychological studies. To date, such attempts have not been very encouraging. Heterogeneity that arises from genetic variations is of particular interest since this is probably as deep as we wish to go (as economists) in 'explaining' observed differences in behaviour. We already have mentioned in the introduction one important commodity, milk, for which genetic variation explains why most of the world's adult population do not consume it. This is an important factor if we are modelling the demand for different foods, even if we use a characteristics framework with preferences defined on calcium, different fats, different vitamins etc.. The mapping from market goods to characteristics depends on the lactase gene. It is now known exactly where the gene for lactase persistence resides (it is on chromosome 2) and with a DNA sample we could determine exactly whether any given individual was lactase persistent. This may be considered fanciful, but increasingly DNA samples will be collected in social surveys; see National Research Council (2001) for details on the feasibility, practicalities, possibilities and ethics of this.

---

[6]In conventional schemes that use a simple felicity function (such as quadratic or iso-elastic utility functions) risk aversion and prudence are often deterministically dependent, but they need not be. Marshall, for example, seemed to believe that 'labourers' were risk averse but imprudent whereas people like himself were both risk averse and prudent.

3.2   Examples from economics.

3.2.1   Empirical demand analysis.

Demand theory presents many good examples of the interactions between the specification of heterogeneity, fitting the data and coherence with the theory. Here the theory is in its purest form either as the Slutsky conditions or revealed preference conditions. There is general agreement that extended versions of AI demand system are needed to fit data reasonably well (at least for the Engel curves) but for illustrative purposes, it is enough to consider the basic form. The AI functional form for the budget share for good i, w_i, given prices $(p_1, p_2, ..., p_n)$ and total outlay x is given by:

$$w_i = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \ln p_j + \beta_i \ln \left(\frac{x}{a(\mathbf{p})} \right)$$

where $a(\mathbf{p})$ is a linear homogeneous price index that depends on all the α and γ parameters. If we wish to estimate with data from many households, we have to allow for heterogeneity. The simplest approach is to assume that all of the parameters are heterogeneous with a joint distribution that is independent of the prices and total expenditure (a 'random effects' approach). If we do this then we run into problems when we impose the Slutsky conditions. The homogeneity and Slutsky symmetry are OK:

$$\sum_{j=1}^{n} \gamma_{ij} = 0, \forall i$$
$$\gamma_{ij} = \gamma_{ji}, \forall i, j$$

since the restrictions are independent of the data. However, the Slutsky negativity condition does depend on the data. For example, the condition that the own price compensated effect should be non-positive is given by:

$$\gamma_{ii} + (\beta_i)^2 \ln \left(\frac{x}{a(\mathbf{p})} \right) + w_i (w_i - 1) \leq 0$$

which clearly depends on the data. Thus the parameters and data are not variation independent[7] which is a necessary condition for stochastic independence. At present it is an open question as to the class of preferences that admit of a random effects formulation. The Cobb-Douglas restriction on the AI system ($\beta_i = \gamma_{ij} = 0$ for all i, j) shows that the class is not empty. On the other hand, although the Cobb-

---

[7]If we have two sets of random variables $\theta \in \Theta$ and $\gamma \in \Gamma$ they are *variation independent* if their joint space is the cross product of the two spaces: $\Theta \times \Gamma$. Thus the support for one set of the random variables does not depend on the realizations of the other.

Douglas does well in terms of consistency with theory it does spectacularly poorly in fitting the data.

3.2.2 Duration models.

Our second example of the difficulty of finding heterogeneity that fit the data and are consistent with theory models is from duration modelling; our discussion here relies heavily on van den Berg (2005). The Mixed proportional hazard (MPH) model is very widely used in duration modelling; this is given by:

$$\theta\left(t \mid x, \upsilon\right) = \psi\left(t\right)\theta_0\left(x\right)\upsilon \tag{11}$$

where $\theta\left(.\right)$ is the hazard given observables x and unobservable υ, $\psi\left(.\right)$ is a baseline hazard that is assumed common to all agents, $\theta_0\left(x\right)$ is the 'systematic' component and υ captures unobserved heterogeneity. This scheme, which was initially devised by statisticians and has been refined by econometricians, has the twin virtues of being easy to estimate and of treating latent and observed heterogeneity in the same way (as multiplicative factors). There are, however, problems with both fit and theory. First, as a matter of fact, stratifying often indicates a significantly different baseline hazard for different strata. Although this can be overcome in the obvious way if we have a lot of data and we observe the variables that define the strata, it is worrying that we just happen to assume the same baseline hazard for stratification that is not observed. The second major problem with the MPH scheme is that it is often presented as reduced form analysis but it is never very clear exactly which structural models are thus ruled out. van den Berg (2005) presents an insightful discussion of this which shows that the class of structural models which have the MPH as a reduced form is relatively uninteresting and many interesting structural models do not have the MPH as a reduced form.

3.2.3 Dynamic structural models.

Our next example is taken from Carro and Mira (2005). They propose and estimate a dynamic stochastic model of sterilization and contraception use. Couples choose between using reversible contraceptive methods, not contracepting and sterilizing. These contraceptive plans are chosen to maximize the intertemporal utility function subject to the laws of motion of the state and, in particular, to birth control 'technology', $\{F_{jt}\}$, for the probability of a birth in period t given contraceptive option j. A homogenous model could not fit the data nor give sensible estimates of the parameters of the model. Two sources of heterogeneity have to be introduced. First,

heterogeneity in the value of children and, second, heterogeneity in the probabilities of a birth (the ability to conceive). Heterogeneity only in preferences did not solve the problem, because in the data there were groups of people not contracepting in almost any period and also having children with much lower probability than other couples who were not contracepting. That is, some couples have lower fecundity, not just different preferences over number of children. Without unobserved heterogeneity in the probability of having a birth (ability to conceive) the model explained the data by saying that the utility cost of contracepting was not a cost; that is, it was positive and significant; and the estimated model did not fit the patterns of contraceptive use across number of children and age. Simple forms of permanent unobserved heterogeneity across couples, using mixing distributions with a small number of types, capture these features of the data. Estimating a structural model allows us to introduce separately heterogeneity in both the probability of having a birth and the value of children. It turns out that both are significant and they are stochastically dependent. A reduced form equation could not separate both sources of unobserved heterogeneity, and using only a fixed effect in a reduced form model will probably not be able to capture the complex effects of both kinds of heterogeneity over couples' choices in the life cycle.

Adding unobserved heterogeneity in $\{F_{jt}\}$ complicated the estimation procedure, since we could no longer write separate likelihoods for choices and conditional probabilities of a birth. Furthermore, with unobserved heterogeneity the dynamic structural model implied by the forward looking behaviour has to be solved for each unobserved type, significantly increasing the computational costs. This is why in this literature only a small number of unobserved types are consider as forms of permanent unobserved heterogeneity, in contrast to the more general specifications considered in reduced form models.

Another example of this is Keane and Wolpin (1997). They estimate a dynamic structural model of schooling, work, and occupational choice decisions. They allow for four unobserved types. Each type of individuals differs from the other types on the initial endowments of innate talents and human capital accumulated up to the age of 16, which is taken as the start of the process in this model. The endowment is known by the individual but unobserved by the researcher. A fundamental finding in Keane and Wolpin (1997) is "that inequality in skill endowments 'explains' the bulk of the variation in lifetime utility". According to their estimates, "unobserved endowment heterogeneity, as measured at age 16, accounts for 90 percent of the variance in lifetime utility". As they say, "it is specially troublesome, given this finding, that unobserved heterogeneity is usually left as a black box.". Nevertheless,

15

they have a clear interpretation for this unobserved heterogeneity coming from the structural model, they can determine some of the correlates of the heterogeneity and compute the conditional probability distribution of the endowment types using Bayes rule. This helps to understand the source of this important unobserved factor on the life-time well being, and to obtain some knowledge about how inequality could be altered by policy.

3.2.4 Returns to schooling.

The estimation of the returns to schooling is another good example where unobserved heterogeneity has played a major role, in both the theoretical and the empirical literature. During more than three decades a vast number of research papers have tried to address this issue in a convincing and theoretically coherent way. It has proven to be a difficult task. The classical Mincer equation used in many papers to estimate the returns to schooling in practice assumed homogenous returns to schooling. Nevertheless, a model with heterogeneous returns to schooling is an integral part of the human capital literature. A recent example of such a model where heterogeneity is allowed to affect both the intercept of the earnings equation and the slope of the earnings-schooling relation can be found in Card (1999 and 2001). These heterogeneous factors are in principle correlated with schooling, since they are taken into account in the schooling decisions of the individuals. A widely used solution to estimate this heterogeneous model is instrumental variables. The conditions under which this method identifies the return to schooling and the interpretation of this estimate are directly related with the treatment effects literature. The identification of the average return to schooling by IV is only possible under certain restrictive conditions. These are unlikely to be satisfied by many of the supply side instruments used since individual schooling decisions are taken depending also on the supply characteristics.[8] In the context of a dichotomous instrument, if those conditions are not satisfied, conventional IV estimates give the Local Average Treatment Effect, see Imbens and Angrist (1994). More generally, Heckman and Vytlacil (2005) show how the Marginal Treatment Effect can be used to construct and compare alternative measures or averages of the returns to schooling, including the Local Average Treatment Effect. As explained by Card (2001), the IV estimate of the returns to schooling on a heterogeneous earnings equation can be interpreted as a weighted average of the marginal returns to education in the population. The weight for each person is a function of the increment in their schooling induced by the

---

[8]See Card (1999 and 2001) for a discussion of this result and of the conditions needed for the IV to identify the average returns to schooling. See also Heckman and Vytlacil (1998).

instrument. Depending on the problem considered, this average return to schooling for those affected by the instrument could be the policy parameter of interest.

This literature has the virtue of providing a connection between the traditional IV estimator and the economic decision model. In the case we have considered here, the estimation of the classical earnings equation using as instrument a change in the supply conditions (for example, distance to the closest college) gives an average effect for a subgroup of the population in the context of an economic model of schooling decisions with heterogeneous returns to schooling. Nonetheless, even if that is the parameter of interest, some homogeneity on the schooling choice equation is needed: the so-called 'monotonicity assumption' in the treatment effects literature, see Heckman and Vytlacil (2005). In a situation where the monotonicity assumption is not satisfied, or in the case where we want to estimate the average return to schooling for the whole population we need to look for alternatives to identify and estimate the effect of interest. A possibility is estimating a structural model of earnings and schooling; Keane and Wolpin (1997) is a good example.

3.2.5 Dynamic discrete choice modelling.

Our final example concerns smoking; although quite specific we believe it illustrates an important general point. Vink, Willemsen and Boomsma (2003) (VWB) present results based on smoking histories for identical twins, nonidentical twins and siblings. They conclude that the starting conditions (in late childhood or early adulthood) are homogeneous (conditional on potentially observable factors) but that persistence, once started, is largely genetic. Although we might have specific objections to the VWB analysis, let us take it as our 'theory' model for now. Let y_{it} be indicator for i smoking in month t and assume a first order Markov model. In line with the VWB hypothesis, suppose there are two types: 'tough quitters' (A) and 'easy quitters' (B) with:

$$pr_A\left(y_t = 0 \mid y_{t-1} = 1\right) < pr_B\left(y_t = 0 \mid y_{t-1} = 1\right) \tag{12}$$

(where, for convenience, we have dropped other covariates). The starting condition, according to VWB, is homogeneous, conditional on the variables that the researchers observe on environmental factors in late childhood (for example, family background) denoted z_{i0}, so that:

$$pr_A\left(y_{i0} = 1 \mid z_{i0}\right) = pr_B\left(y_{i0} = 1 \mid z_{i0}\right) \tag{13}$$

What of the 'resuming' transition probability: $pr\left(y_t = 1 \mid y_{t-1} = 0\right)$? We could model this as homogeneous or as heterogeneous. An obvious assumption is that the

resuming probability is negatively correlated with the quitting probability so that:

$$pr_A \left(y_t = 1 \mid y_{t-1} = 0 \right) > pr_B \left(y_t = 1 \mid y_{t-1} = 0 \right) \tag{14}$$

For economists, however, an attractive alternative assumption is that since people are forward looking and know their own type[9], a type A who has stopped might be much more reluctant to start again. This assumption reverses the inequality in (14). Whatever the case, we would not want an initial specification for the two transition probabilities and the starting condition that would rule out these possibilities. Now consider a conventional specification that has only one 'fixed effect':

$$pr \left(y_{it} = 1 \mid y_{it-1} \right) = \mathbf{F} \left(\alpha_i + \beta y_{i,t-1} \right) \tag{15}$$

With two types, the easy/tough quitting structure (12) is:

$$1 - \mathbf{F} \left(\alpha_A + \beta \right) < 1 - \mathbf{F} \left(\alpha_B + \beta \right) \tag{16}$$

which implies $\mathbf{F} \left(\alpha_A \right) > \mathbf{F} \left(\alpha_B \right)$ which is (14). Thus the conventional formulation (15), which is often presented as an unrestricted 'reduced form', rules out the interesting structure that would occur to most economists.

This concludes our brief and highly selective discussion of the difficulties of allowing for heterogeneity in a flexible enough way to capture what is in the data and to allow for a wide range of structural models. We now present some of our own recent work on dynamic discrete choice modeling that was motivated by the empirical findings on the demand for whole fat milk presented in section 2 and by examples such as the smoking analysis presented here.

4   Dynamic discrete choice models.

4.1   A stationary Markov chain model.

This section presents and summarizes some results from Browning and Carro (2005) (BC) concerning heterogeneity in dynamic discrete choice models. Since very little is known about such models when we allow for lots of heterogeneity, we consider the simple model with no covariates. An additional and important advantage of considering the simple model is that we can derive exact analytical finite sample

---

[9]So that the type is heterogeneous and not uncertain by our definition of heterogeneity. If those who never smoked do not know their type then we have both heterogeneity and uncertainty in types.

properties. The restriction on theory that we impose is that the reduced form for the structural model is a stationary first order Markov chain. In many contexts the stationarity restriction may be untenable, but we have to start somewhere. In this simple model the approach is fully nonparametric, conditional only on that modelling choice. We focus directly on the two transition parameters:

$$
\begin{aligned}
G_i &= pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right) \\
H_i &= pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right)
\end{aligned}
\tag{17}
$$

where i is individual indicator, t is time indicator and $t = 0, 1, ..., T$. This is the maximal heterogeneity we can allow in this context. For instance, in the smoking example of subsection 3.2.5, the structure described by equations (12) and (14), correspond to $H_A > H_B$ and $G_A > G_B$ respectively. In this exposition we shall focus on the marginal dynamic effect:

$$
M_i = H_i - G_i
\tag{18}
$$

which gives the impact on the current probability of $y_{it} = 1$ from changing the lagged value from 0 to 1. A model with homogeneous dynamic marginal effects would impose:

$$
H_i = M + G_i \in [0, 1]
\tag{19}
$$

for some $M \in [-1, 1]$. A parametric model with a homogeneous persistence parameter would impose:

$$
\begin{aligned}
G_i &= F\left(\alpha_i\right) \\
H_i &= F\left(\alpha_i + \beta\right)
\end{aligned}
\tag{20}
$$

for some cdf $F\left(.\right)$.

4.2 Estimation with one sequence.

We begin by considering a single realization of a sequence for one person. There are 2^{T+1} possible sequences of 1's and 0's for a single chain of length $T + 1$ (or 2^T if we condition on the initial value). An estimator $\left(\hat{G}, \hat{H}\right)$ is a mapping from the 2^{T+1} realizations to sets in the unit square. If the mapping is single valued then the parameters are point identified by that estimator. The first result in BC is that

there is no unbiased estimator for G and H. With this result in mind we look for estimators that have low bias or low mean squared error. The maximum likelihood estimator (MLE) has a simple analytical closed form:

$$\hat{G}^{MLE} = \frac{n_{01}}{n_{00} + n_{01}} \tag{21}$$

$$\hat{H}^{MLE} = \frac{n_{11}}{n_{10} + n_{11}} \tag{22}$$

where n_{01} is the number of $0 \rightarrow 1$ transitions, etc.. Note that $\hat{G}^{MLE}$ is point identified iff we have a sequence that has at least one pair beginning with a zero, so point identification requires us to drop some possible observations. The second result in BC is that the MLE estimate of the marginal dynamic effect (if it exists) has a negative bias, that is:

$$E\left(\hat{H}^{MLE} - \hat{G}^{MLE}\right) < M \tag{23}$$

This result is the discrete choice analogue of the Nickell bias for linear dynamic models (see Arellano (2003a)) . The degree of bias depends on the parameter values and the length of the panel, T. As we would hope, the bias of the MLE estimator of the marginal dynamic effect diminishes as we increase the length of the panel, but even for $T = 16$ it can be high.

Based on the exact formulae for the bias of the MLE, BC construct a nonlinear bias corrected (NBC) estimator as a two step estimator with the MLE as the first step.[10] We find that this estimator does indeed reduce the bias for most cases (as compared to MLE). For all but extreme values of negative state dependence, the NBC estimator also has a negative bias for the marginal dynamic effect. The order of bias for the MLE is approximately $O\left(T^{-1}\right)$ (the exact order depends on the parameter values) whereas the order for the NBC estimator is approximately $O\left(T^{-2}\right)$ so that the small sample bias diminishes much faster for NBC. Despite these advantages on the bias, in mean squared error (mse) terms the NBC estimator is never much better than MLE and it is worse in some cases. A detailed examination of the MLE and NBC estimators suggested that neither can be preferred to the other.

Given the relatively poor performance of the MLE and NBC in terms of mse, BC construct an estimator that addresses the mse directly. The mean square error

---

[10]BC show analytically that the estimator which continues to apply bias corrections after the first does not necessarily converge, so that only the two step estimator is considered.

for an estimator $\widehat{M}$ of the marginal dynamic effect is given by:

$$\lambda\left(\hat{M}; G, H\right) = \sum_{j=1}^{J} p_j\left(G, H\right)\left(\hat{M}_j - (H - G)\right)^2 \qquad (24)$$

where j denotes a particular sequence, $J = 2^T$ (if we condition on the initial value) and p_j is the probability of sequence j given the values of G and H. Since there is no estimator that minimizes the mse for all values of (G, H), we look for the minimum for some choice of a prior distribution of (G, H), $f(G, H)$ so that the integrated mse is given by:

$$\int_0^1 \int_0^1 \lambda\left(\hat{M}; G, H\right) f(G, H)\, dG dH \qquad (25)$$

Given that we consider a general case in which we have no idea of the context, the obvious choice is the uniform distribution on $[0, 1]^2$, $f(G, H) = 1$. Minimizing gives the following minimum integrated mse (MIMSE) estimator:

$$\hat{M}_j^{MIMSE} = \frac{n_{11} + 1}{n_{10} + n_{11} + 2} - \frac{n_{01} + 1}{n_{00} + n_{01} + 2} \qquad (26)$$

This estimator is the mean of the posterior distribution assuming a uniform prior. The attractions of the MIMSE estimator are that it is very easy to compute, it is always identified and it converges to maximum likelihood as T becomes large so that it inherits all of the desirable large sample properties of MLE. Figure 5 shows the small sample bias for the three estimators. The rate of convergence of the bias to zero for the MIMSE estimator is approximately $O\left(T^{-0.6}\right)$. As can be seen, the NBC estimator starts off with a relatively small bias and converges more quickly to zero. Thus the NBC estimator unequivocally dominates the other two estimators in terms of bias. When we turn to the the mse, however, MIMSE is much better than either of the other two estimators, particularly when there is some positive state dependence ($M > 0$). This is shown in figure 6. As can be seen there, the MIMSE estimator starts not too far above the CR bound (for the MLE estimator) and converges relatively slowly to it. The other two estimators have a relatively high RMSE, particularly when we have short observation period.

4.3 Estimation with pooled data.

In the previous subsection we considered households in isolation but in most cases the interest is not in individual households, but in the population. Thus, it may be that the distribution of M in the population is of primary interest, rather than the values for particular households. Suppose that we observe many households. We first consider the nonparametric identification of the distribution of (G, H) with

fixed T. We shall assume that we are given population values for outcomes. In this case the relevant population values are the proportions of each of the 2^T possible cases. Denote the population values by π_j for $j = 1, 2...2^T$. Now suppose that (G, H) are distributed over $[0, 1]^2$ with a density $f(G, H)$. The population proportions are given by the integral equations:[11]

$$\pi_j = \int_0^1 \int_0^1 p_j(G, H) f(G, H) \, dGdH, j = 1, 2...2^T \tag{27}$$

where, the probabilities are given by:

$$p_j(G, H) = G^{n_{01}^j} (1 - G)^{n_{00}^j} H^{n_{11}^j} (1 - H)^{n_{10}^j} \tag{28}$$

Assuming that the π_j's satisfy the conditions imposed by the model, we check the following necessary condition for identification: is there only one density $f(G, H)$ which is consistent with the set of 2^T equations (27)? The answer is negative. To show this we impose some structure on the model and show that even with these additional constraints the structure is not identified. Consider the case with $T = 3$ and in which we restrict the distribution of the G's and H's to be discrete, each with three values: $\{G_1, G_2, G_3\}$ and $\{H_1, H_2, H_3\}$. Let the probabilities of each of the nine combinations (G_k, H_l) be given by the (9×1) vector θ with values that sum to unity. Define the (8×9) matrix A by:

$$A_{jm} = (G_m)^{n_{01}^j} (1 - G_m)^{n_{00}^j} (H_m)^{n_{11}^j} (1 - H_m)^{n_{10}^j} \tag{29}$$

Then the analogue to (27) is:

$$\pi = A\theta \tag{30}$$

where π is observed and the values of $\{G_1, G_2, G_3\}$, $\{H_1, H_2, H_3\}$ and θ are to be solved for. Clearly the latter are not uniquely determined by the former since we have 8 equations and 14 unknowns.[12] There are more than one distribution of (G, H) that generates the same observed π in (30). Thus the distribution is not nonparametrically identified. We need to either put on more structure such as a parametric model for heterogeneity, or estimate nonparametrically M_i for each unit separately and then use those estimates to define the empirical distribution of the parameters. In BC we explore the latter approach, that is, obtaining the empirical

---

[11]Note that we have made an analysis conditional on the initial observation y_{i0}, so $f(G, H)$ here it is the distribution given y_{i0}. A similar result could be get about the identification of the unconditional distribution.

[12]The number of equations, 8 is equal to 2^T. As matter of fact two of the 2^T cases give the same equation on (30), so there are only 7 different equations.

distribution from estimates for each unit. Simulations with $T = 9$ (that is, with 10 observations, including the initial observation) suggest that the MIMSE based estimator significantly outperforms the MLE and NBC estimators in recovering the distribution of the marginal dynamic effect. This can be seen, for instance in Figure 7, that presents cdf's of the three estimators from simulations using the empirical distribution for (G, H) from the estimates reported in section 2.

This analysis suggests that it is feasible to define reasonably well performed estimators with maximal heterogeneity for the dynamic discrete choice model. Clearly such an estimator will be consistent with any theory model that generates a stationary first order Markov chains and it will also fit any generating process for the heterogeneity. In some contexts such estimators will significantly outperform (in terms of fit and congruence with theory) some version of (20) estimated using a conventional 'fixed effect' scheme.

4.4 Relation to recent developments in estimation of non-linear panel data models.

In the recent years new methods of estimation for nonlinear panel data models have been developed. Arellano and Hahn (2005) present a review and explain and derive connections between the different solutions developed.[13] The central focus is on nonlinear models with fixed effects and the attempt to overcome the incidental parameters problem that arises from the estimation by standard MLE of common parameters in these models. Usually the specific constant intercept, the so-called fixed effect, is the only heterogeneous coefficient and the consequent incidental parameters problem may lead to severe bias in panels where T is not large. The new methods developed reduce the order of the magnitude in T of that bias, so that it may be negligible in finite samples used in practice. They remove the first order term on the expansion of the asymptotic bias of the MLE, and consider asymptotics with both N and T going to infinity.

In BC not only the intercept but also the slope are individual specific. Given this, we have a separate model for each individual in the panel. In contrast with the literature reviewed in Arellano and Hahn (2005), where the object of interest is a common parameter in the population, we first consider estimating each separate model with the T observations of each individual. In this regard our analysis is closer to the time series literature. There is no asymptotic in N given that we are estimating with one sequence. Another difference is that the nonlinear bias corrected

---

[13] The literature reviewed by Arellano and Hahn (2005) includes Arellano (2003b), Carro (2004), Fernandez-Val (2005), Hahn and Newey (2004), and Woutersen (2004) among others.

estimator (NBC) considered by BC is not based on a first order reduction on the asymptotic bias. It is based on the exact finite sample properties of the MLE. So the NBC estimator is based on the exact formulae of the finite sample bias that BC derive. These differences imply that some of our conclusions diverge from the recent literature. Theoretically, the first order reduction on the asymptotic bias reviewed in Arellano and Hahn (2005) does not increase the asymptotic variance as N and T go to infinity at the same rate. Furthermore, finite sample experiments, for example in Carro (2004) and Fernandez-Val (2005), provide some evidence that bias reduction can lead to a better estimator in terms of mean square error for a panel with a moderate number of periods.[14] However, derivations of the exact mean square error of the NBC show that it does not dominate the MLE in this criterion since it is never much better and it is worse in some cases. This means that while NBC significantly reduces the bias, it also significantly increases the variances of the estimator in finite samples. MIMSE is a different approach in the sense that it is not derived to reduce the bias but to minimize the mean square error in finite samples. Given this, it is not defined as a bias correction of the MLE, but as new estimator in accordance with the chosen criterion.

There are two possible motivations for considering estimation of each individual's model with one sequence. First, we may be interested in each individual, for example if we are analyzing the default risk of each credit applicant. Second, in the kind of models considered, having an unbiased estimator for each individual is sufficient to define a fixed-T consistent estimator of a parameter on the population of individuals. Even if is not possible to define an unbiased estimator, as shown in BC for a first order Markov chain, having an estimator for each individual model with improved properties in finite samples could lead to an estimator of a parameter defined over the population of individuals, with good finite sample properties when pooling many households. Any parameter of interest defined as a function of the model's parameters will benefit from this. In the analysis discussed in the previous three subsections there is no parameter of the model that is common to all the individuals and the marginal effect of a variable is heterogeneous. Nevertheless, in many cases our interest is in particular moments of the distribution of the marginal effect on the population; for example, the median marginal effect of a variable. BC consider estimating the whole distribution of the marginal effect in the population with pooled data. As described in the previous subsection, BC explore using the nonparametric estimators for each unit already considered (MLE, NBC, MIMSE)

---

[14]Of course, simulation experiments have been done only for some specific sample sizes and binary choice models. More finite sample experiments are needed to evalute each of those new methods.

and then obtaining the empirical distribution in the population from estimates of the individual marginal effects. Focusing on a moment of that distribution, in principle it could be possible to apply the ideas in Arellano and Hahn (2005), since this is a common parameter to be estimated with a not very large number of periods that suffers the incidental parameters problem. This correction, following ideas in section 8 of Arellano and Hahn (2005), would be specific to each parameter of interest one may want to consider, in contrast to the case where you have good estimates of the model's parameters. In any case, this possibility remains unexplored in practice for models where all the parameters are heterogeneous.

5   Conclusions.

In this paper we have presented a selective and idiosyncratic view of the current state of allowing for heterogeneity in microeconometric modelling. Our main theme has been that there is more heterogeneity than we usually allow for and it matters for outcomes of interest. Additionally, it is difficult to allow for heterogeneity but when considering how to do it we have to keep an eye on fitting the data and on the interpretability of the estimates. Thus how we introduce heterogeneity into our empirical analysis should depend on the data to hand, the questions deemed to be of interest and the economic models under consideration. The lesson from the last thirty years seems to be that this requires case by case specifications and eschewing the use of schemes whose only virtue is that they are statistically convenient.

References

[1] Alvarez, Javier, Martin Browning and Mette Ejrnaes (2002): "Modelling income processes with lots of heterogeneity", CAM WP-2002-01, University of Copenhagen.

[2] Andersen, Erling B. (1970): "Asymptotic properties of conditional maximum likelihood estimators", *Journal of the Royal Statistical Society*, Series B, 32, 283-301.

[3] Arellano, Manuel (2003a): *Panel Data Econometrics,* Oxford University Press.

[4] Arellano, Manuel (2003b): "Discrete Choice with Panel Data" *Investigaciones Económicas,* vol. XXVII (3), 423-458.

[5] Arellano, Manuel and Jinyong Hahn (2005): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments". *CEMFI Working Paper No. 0507*. Prepared for the Econometric Society World Congress, London, August 2005.

[6] Browning, Martin and Jesus Carro (2005): "Heterogeneity in dynamic discrete choice models". *unpublished manuscript.*

[7] Card, David (1999): "The Causal Effect of Education on Earnings", in *Handbook of Labor Economics*, volume 3A, ed. by Orley Ashenfelter and David Card. Amsterdam and New York: North Holland.

[8] Card, David (2001): "Estimating the return to schooling: progress on some persistent econometric problems", *Econometrica*, 69, 1127-1160

[9] Carro, Jesus (2004): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", unpublished manuscript.

[10] Carro, Jesus and Pedro Mira (2005): "A dynamic model of contraceptive choice of Spanish couples", *Journal of Applied Econometrics*, forthcoming.

[11] Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47, 225.238.

[12] Cunha, Flavio, James Heckman and Salvador Navarro (2005), "Separating uncertainty from heterogeneity in life cycle earnings", *Oxford Economic Papers*, 57, 191- 261.

[13] Fernandez-Val, Ivan (2005): "Estimation of Structural Parameters and Marginal Effects in Binary Choice Panel Data Models with Fixed Effects", unpublished manuscript.

[14] Hahn, Jinyong and Whitney Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica,* 72, 1295-1319.

[15] Heckman, James (1981): "Statistical Models for Discrete Panel Data" in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (eds.), MIT Press.

[16] Heckman, James (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture", *Journal of Political Economy*, 109, 673-748.

[17] Heckman, James, Rosa Matzkin and Lars Nesheim (2003): "Simulation and Estimation of Nonadditive Hedonic Models", *NBER Working Papers*: 9895.

[18] Heckman, James and Edward Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling", *Journal of Human Resources*, 33, 974-987.

[19] Heckman, James and Edward Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluations", *Econometrica*, 73, 669-738.

[20] Imbens, Guido and Joshua D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467-476.

[21] Keane Michael P. and Kenneth I. Wolpin (1997): "The Career Decisions of Young Men", *Journal of Political Economy*, 105, 473-521.

[22] Lancaster, Tony (1979): "Econometric Methods for the Duration of Unemployment", *Econometrica*, 47, 939-56

[23] Laroque, Guy (2005): "Income Maintenance and Labor Force Participation", *Econometrica*, 73,

[24] Lindsey, J. K. (2001): *Parametric Statistical Inference*, Oxford Statistical Science Series, Clarendon Press, Oxford.

[25] McFadden, Daniel and Kenneth Train (2000), "Mixed MNL models of discrete response", *Journal of Applied Econometrics*, 15, 447-470.

[26] Meghir, Costas and Luigi Pistaferri (2004): "Income Variance Dynamics and Heterogeneity", *Econometrica*, 72, 1-32.

[27] Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias", *Journal of Farm Economics*, 43, 44-56.

[28] National Research Council (2001), *Cells and Surveys*, National Academy Press, Washington, D.C.

[29] Patterson, K. David (200): "Lactose intolerance", chapter IV.E.6, volume 1, *The Cambridge World History of Food*, Kenneth Kiple and Kriemhild Conee Ornelas (eds), Cambridge University Press, Cambridge.

[30] Pesaran M. Hashem and Ron Smith (1995): "Estimating long-run relationships from dynamic heterogenous panels ", *Journal of Econometrics*, 68, 79-113.

[31] Rubinstein and Weiss (2005): "Post Schooling Wage Growth: Investment, Search and Learning, *mimeo,* Eithan Berglas School of Economics, Tel-Aviv University.

[32] Stigler, George J; Becker, Gary S (1977): "De Gustibus Non Est Disputandum", *American Economic Review*, 67, 76-90.

[33] Vink, Jacqueline, Gonneke Willemsen and Dorret Boomsma (2003): " The association of current smoking behavior with the smoking of parents, siblings, friends and spouses", *Addiction*, 98, 923-931.

[34] Wooldridge, Jeffrey (2005), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity", *Journal of Applied Econometrics*, 20, 39-54

[35] Woutersen, Tiemen (2004): "Robustness Against Incidental Parameters", unpublished manuscript.
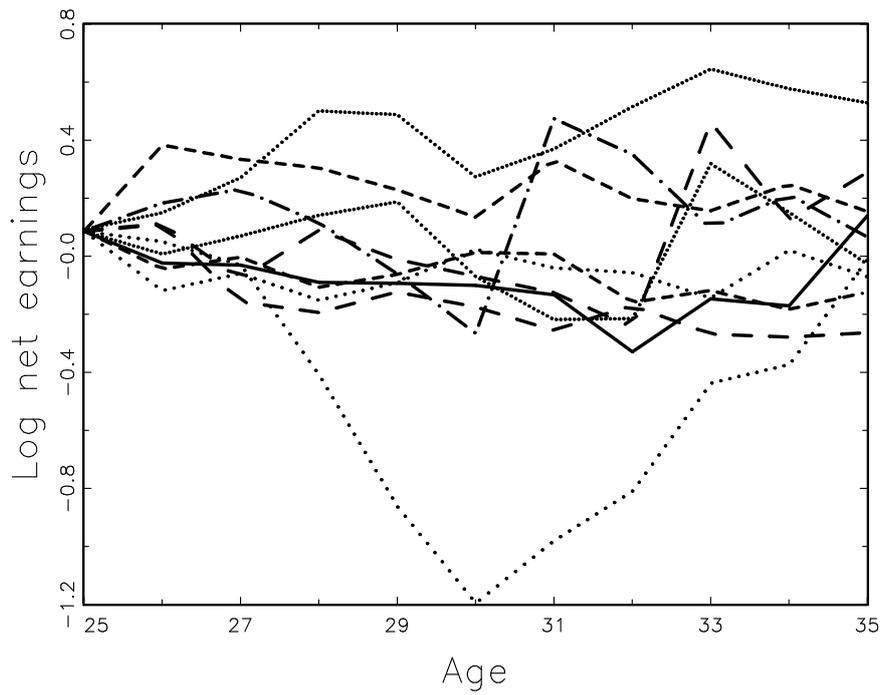
Figure 1: Ten earnings paths for US white male high school graduates.
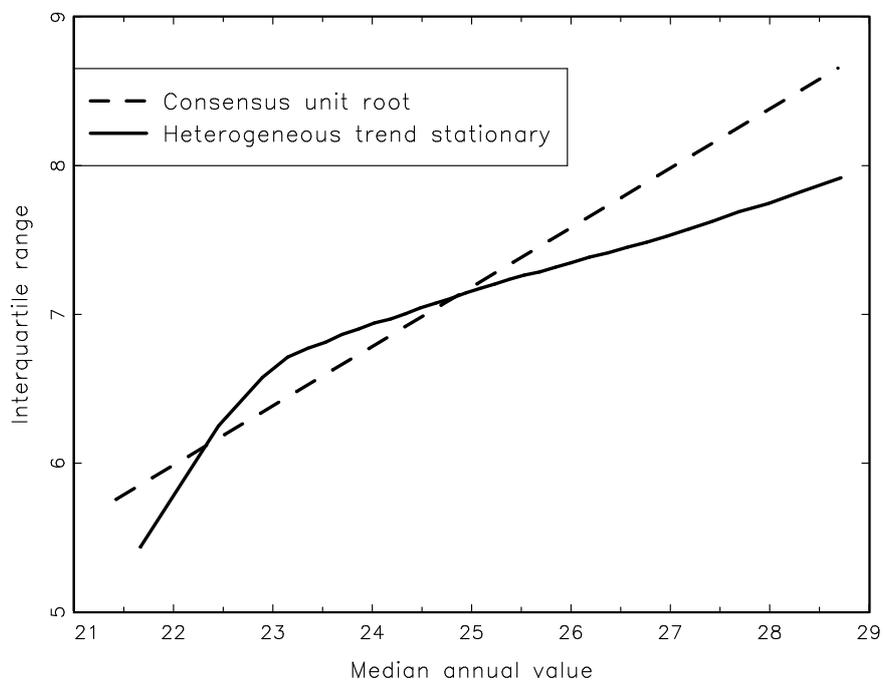
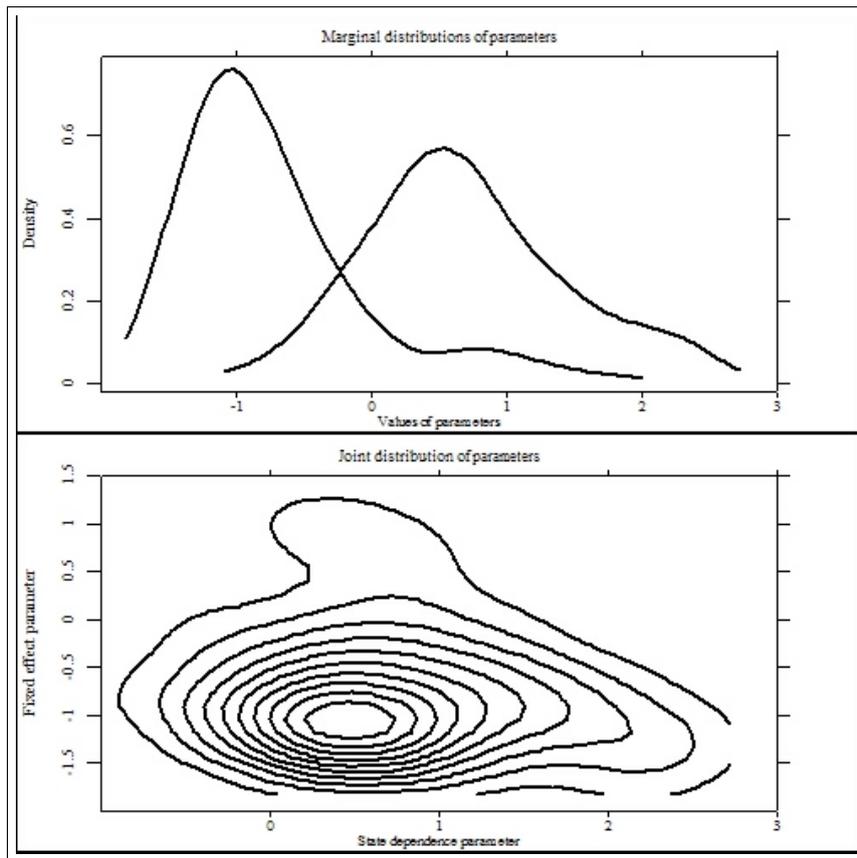

Figure 2: The trade-off for lifetime earnings.

Figure 3: Distribution of parameters for dynamic discrete choice model.
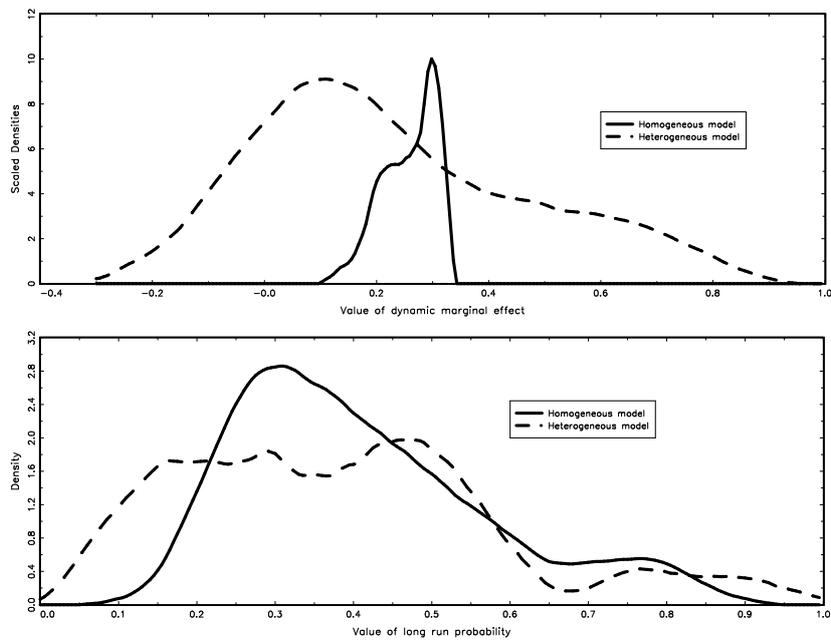

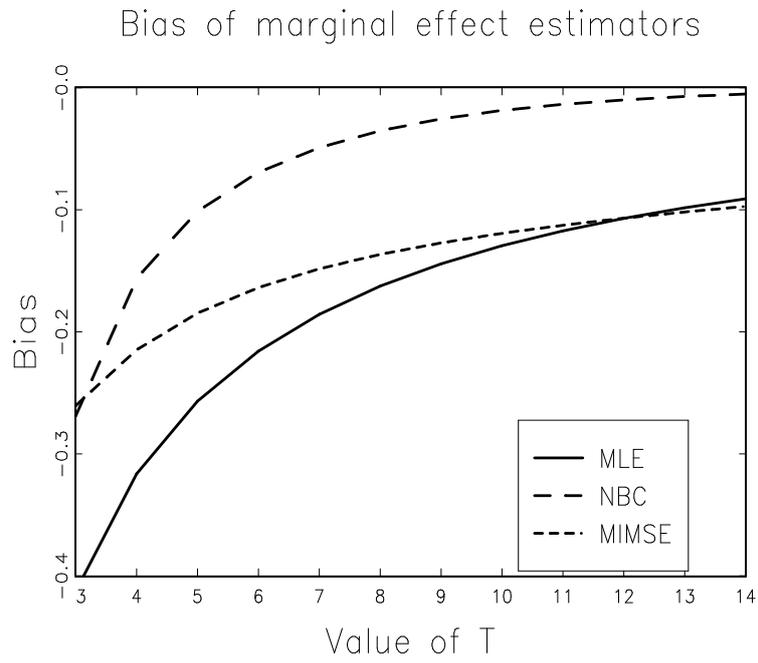Figure 4: Marginal distributions of parameters of interest.

Bias of marginal effect estimators

Figure 5: Bias of three estimators.
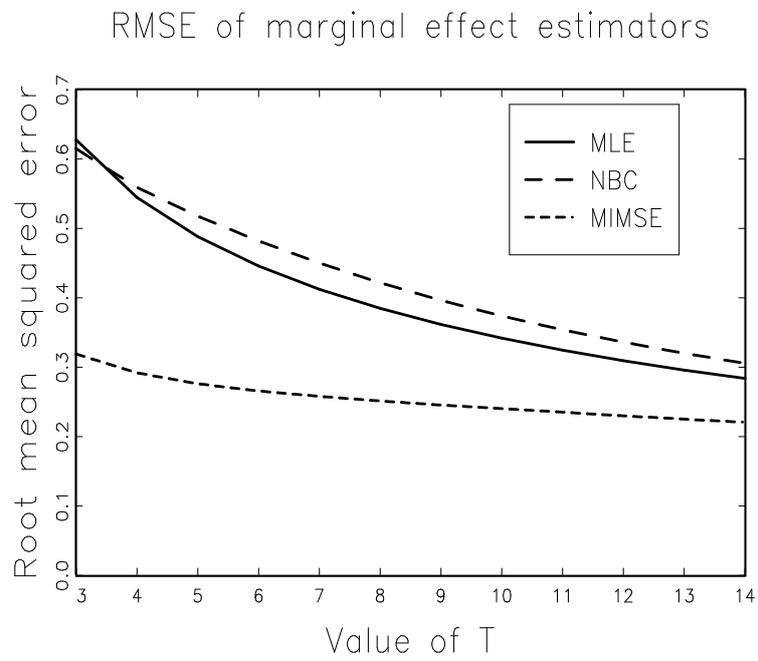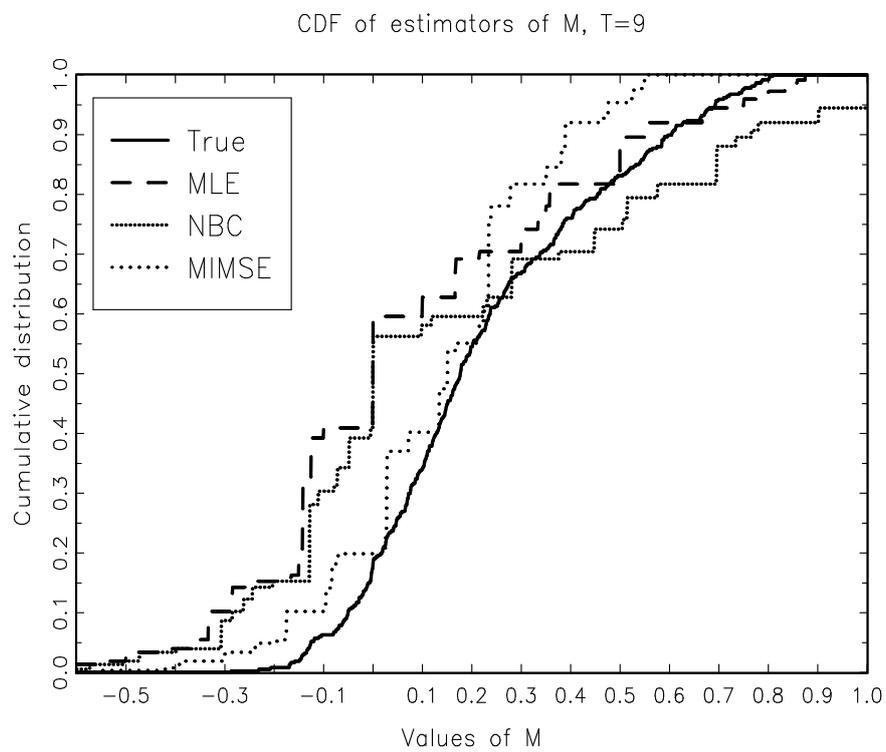
RMSE of marginal effect estimators

Figure 6: RMSE for three estimators.

31

Figure 7: Distributions of three estimators.