Forecasting with micro panels: The case of health care costs

Denzil G. Fiebig*[a] and Meliyanni Johar[b]
[a]School of Economics, University of New South Wales, Sydney, Australia
[b]Economics Discipline Group, University of Technology Sydney, Sydney, Australia

Version: 24 January 2014

Abstract

Micro panels characterized by large numbers of individuals observed over a short time period provide a rich source of information but as yet there is only limited experience in using such data for forecasting. We exploit the linkage of a representative survey of more than 250,000 Australians aged 45 and over to several years of hospital, medical and pharmaceutical records. After calculating total health care cost for each survey respondent, we explore the relative forecasting performance of alternative models of health expenditures. In contrast to much of the risk adjustment literature, where such modelling is prevalent, the availability of panel data allows the use of a fixed effects approach in order to guard against possible omitted variable biases associated with unobservable individual specific effects. When parameter estimation is the main objective there is a well-established preference for using fixed effects and existing simulation evidence also supports the use of a fixed effects approach when forecasting is the main objective. We demonstrate in many forecasting tasks this is a fragile conclusion that is likely to be overturned, including in our models of health care costs. Simulation results add support and additional insights into the results obtained in the application.

Acknowledgements

This research uses data from the 45 and Up Study which is managed by the Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the Heart Foundation (NSW Division); NSW Ministry of Health; *beyondblue*; Ageing, Disability and Home Care, NSW Family and Community Services; Australian Red Cross Blood Service and UnitingCare Ageing. This project was undertaken by the University of Technology Sydney and utilized Medical Benefit Schedule (MBS) and Pharmaceutical Benefit Schedule (PBS) data supplied by the Department of Human Services. Data linkage for the project was undertaken by the Centre for Health Record Linkage. The 45 and Up Study has the approval of the University of NSW Health Research Ethics Committee; this project has ethics approval from the NSW Population and Health Services Research Ethics Committee and the Department of Health and Ageing Departmental Ethics Committee. The study's findings are those of the authors and do not necessarily represent the views of the Department of Health and Ageing, or the Department of Human Services. The project is funded by an ARC Discovery Project grant (DP110100729). We are grateful for the helpful comments made by Randy Ellis and seminar participants at the University of New South Wales, Monash and Wollongong.

*Correspondence to: E-mail: d.fiebig@unsw.edu.au

Keywords: Health expenditure, risk adjustment, panel data, fixed effects, forecasting
JEL Codes I10, C23, C53

1. Introduction

In a recent survey on forecasting with panel data, Baltagi (2008) contrasts the rich and extensive forecasting literature using time series data with the relatively modest literature on forecasting using panel data. Even within the area of panel forecasting, there is very little general research on micro panels where *N*, the number of individuals, is large relative to *T*, the number of time series observations. Baltagi (2008) references only one relevant paper by Baillie and Baltagi (1999) who consider forecasting in the context of the standard one-way error component model and compare a number of predictors including models estimated by OLS and fixed effects (FE) as well as variants of the best linear unbiased predictor (BLUP) assuming a random effects specification. Given the relatively good performance of the FE forecasts, even when the random effects specification is the true generating process, Baillie and Baltagi (1999) argue there is a clear preference for FE in order to guard against possible omitted variable biases associated with correlations between unobservable individual specific effects and covariates included in the models. This preference for FE mirrors the conventional wisdom when parameter estimation is the main objective.

Our primary aim is to provide further comparisons of alternative predictors in the context of the motivating example of risk prediction in health economics. In particular, we ask whether the superiority of models estimated by FE over those estimated by OLS is in fact clear cut in the context of micro panels such as the data we use. The key feature of such panels is the availability of a very large number of individuals (more than 250,000 in our data) but for a limited time dimension (4 years in our data). The survey of Baltagi (2008) highlights the existing focus on forecasting in the time dimension. In our risk adjustment application, this task is about forecasting the future health costs of a given pool of individuals where their past cost histories are available. With panel data there is a second type of forecast where the task is to predict outside the sample to a new group of individuals for whom there is no past data (e.g., potential customers). In our analysis we consider both forms of forecasting, and to distinguish them, we refer to forecasting in the cross section dimension as "out-of-sample" and forecasting in the time series dimension as "post-sample" forecasting.

There is no shortage of recent research dealing with various aspects of modelling individual health care treatment costs and expenditures. The primary motivation derives from the use of such models for risk adjustment. Broadly defined, risk adjustment in health means the use of patient-level information to explain variation in health care utilization, costs and health outcomes (Ellis, 2008). The application of risk adjustment is typically for payment purposes, such as payment to competitive health insurance plans or to health providers, and health insurance premium setting. A good risk adjustment prediction model is one that can forecast accurately the resource use of the individual.

Key features of health expenditure data that make modelling a challenge are the presence of a substantial proportion of zero observations (non-users of health services) and positive costs that are highly skewed to the right with long thick right-hand tails. Thus many papers compare alternative modelling approaches tailored to accommodate these features using cross sectional data; see for example Jones, Lomas and Rice (2013) who ignore the non-users in search for an accurate model of highly skewed data using over 6 million healthcare observations.

What are in relatively short supply are analyses involving panel data. Two recent survey papers on econometric modelling in health economics by the same author serves to illustrate this divide. Jones (2011) surveys the health care cost modelling literature as characterized above but references Jones (2009) for discussion of panel data methods. However, there is only a brief mention of health care cost modelling for risk adjustment in Jones (2009). Instead the applications focussed on policy evaluation. Now the choices made on coverage in these two survey papers could merely be driven by the need to narrow the focus but a close look at the literature does suggest there has been little work on modelling health care costs using panel data.

Studies where panel have been used to estimate health expenditure models include Seshmani and Gray (2004a,b), Stearns and Norton (2004), Albouy et al (2010) and Hill and Miller (2011).[1] In estimating the impact of age and time of death on hospital costs, Seshamani and Gray (2004a) find that ignoring individual fixed effects results in significant omitted variable bias. Their sample is large in both *N* and *T* dimensions: over 90,000 individuals with up to 24 years of observations (unbalanced panel). However, because the impact of time invariant factors such as sex is of key interest, they rely on a random effect model. Seshamani and Gray (2004b) use the same data set but have smaller *N* (9,371) and account for the panel nature of the data only through robust standard errors. Stearns and Norton (2004) estimate fixed effect models but their final predictions of future health costs are made based on a random effects specification. Albouy et al (2010) consider state dependence in panel data models. Their sample has moderate size *N* (about 7,000) and *T* (up to 6 years). Instead of fixed effects, the panel structure is accommodated via Wooldridge (2005) type corrections for initial conditions. Hill and Miller (2011) use the US Medical Expenditure Panel Survey (MEPS) data to compare the performance of various models of health expenditure. Although the MEPS data is panel in nature due to quarterly interviews, the study focuses on annual expenditure so the analysis is cross-sectional.

[1] There are also several descriptive studies using commercial MarketScan Databases in the US, which contains data from the employer and health plan sources concerning medical and drug data for several million commercially-insured individuals, including employees, their spouses, and dependents collected since 1995 (e.g. Aizcorbe, 2012) and MEPS data (e.g. Zuvekas and Olin, 2009; Bernard et al., 2011).

Thus there is little research that specifically addresses forecasting issues when modelling health costs and expenditures using panel data. Before proceeding to an extensive analysis of our particular micro panel of healthcare costs, a Monte Carlo study is conducted in order to illustrate the key issues and tradeoffs involved in selecting appropriate predictors.

2. Econometric framework

We abstract from econometric issues that have been the primary subject of many studies comparing approaches for modelling health costs such as transforming the dependent variable, which invites the problems of re-transformation as policymakers require forecasts in raw scale, and accounting for the presence of zero observations. In part this is to focus attention on issues arising when panel data are available but it is also a choice supported by past comparisons where simple linear models estimated by OLS do relatively well. As Jones (2011) concludes:

> "*It is notable that the simple linear model, estimated by OLS, performs quite well across all of the criteria, a finding that has been reinforced for larger datasets than the one used here.*"

This is in fact what we have found with preliminary analyses using our data on health expenditures; see Ellis et al (2013). Our expenditure data happen to have a small proportion of zero observations due to the setting of a universal public healthcare system. Thus the forecasting comparison here is kept narrowly focused on the differences between predictors using estimates produced by OLS and by variants of fixed effects models with a common set of regressors in the context of a basic linear panel model:

$$(1)\quad y_{it} = \alpha + x_{it}'\beta + z_i'\gamma + \mu_i + u_{it}; i = 1, \dots, N, N+1, \dots, N+n; t = 1, \dots, T, T+1, \dots, T+\tau.$$

In this setup the distinction between regressors that vary over both time and individuals (x_{it}) and those that are time invariant (z_i) has been made explicit and we have allowed for the potential presence of unobservable time-invariant factors, the μ_i. There has also been a distinction made between two types of forecasting exercises. In using (1) to extrapolate beyond the data used for estimation purposes, one could consider producing forecasts for $y_{i,T+1}$. In other words, one could forecast future costs for the sample of individuals used for estimation of the models parameters; forecasting in the time dimension will be termed **post-sample**. But it would also be relevant to consider a different sample of individuals not used in the estimation stage; forecasting $y_{N+1,t}$ in the cross sectional dimension will be termed **out-of-sample**. (Obviously there is also the case of post-out-of-sample forecasting $y_{N+1,T+1}$). Several alternative predictors are considered although not all will be available for these different forecasting tasks.

In their simulation experiments, Baillie and Baltagi (1999) specified the data generating process to be a classical one-way error component model where the μ_i in (1) are assumed random and independent of all regressors. In their comparison of predictors they emphasized the impact of accounting for individual effects on both estimation and prediction. But because of their choice of data generating process, the resulting estimation problem abstracts from possible biases in coefficient estimates and concentrates on relative efficiencies of alternative predictors. While this represents a reasonable base case, the alternative situation where the μ_i are potentially correlated with the regressors is an important and practically relevant extension. Here the generation of consistent parameter estimates becomes an issue in comparing alternative predictors. The Baillie and Baltagi (1999) framework is also extended by considering the two different types of prediction problems.

OLS applied to the pooled data, ignoring the individual specific effects, yields parameter estimates denoted by $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ and produces the OLS predictor (OLSP) given by:

$$\text{OLSP: } \tilde{y}_{it} = \tilde{\alpha} + x_{it}'\tilde{\beta} + z_i'\tilde{\gamma}.$$

For fixed effects estimation, the intercept is parameterized to be the mean of the individual specific effects implying that the estimated μ_i are restricted to have a zero mean. This provides options in defining predictors. If the estimated μ_i are not used in forming forecasts then we denote this predictor FE(xb). Alternatively, adding in the estimated fixed effects yields the predictor FE(xb+mu). Thus denoting the FE parameter estimates by $\hat{\alpha}, \hat{\beta}, \hat{\mu}_i$ these alternative predictors are defined as:

$$\text{FE(xb): } \hat{y}_{it} = \hat{\alpha} + x_{it}'\hat{\beta}$$

$$\text{FE(xb + mu): } \hat{y}_{it} = \hat{\alpha} + x_{it}'\hat{\beta} + \hat{\mu}_i.$$

In a post-sample forecasting task, both of these FE approaches are feasible predictors but only FE(xb) is available in the out-of-sample forecasting task.

Note that the time-invariant *z*'s will appear in the OLSP but not in either of the FE(.) predictors. From a forecasting perspective this may or may not be an important source of differentiation between predictors. In post-sample forecasting, FE(xb+mu) provides a very flexible alternative to allowing for the *z*'s in the predictor. However, in the out-of-sample forecasting task one might expect that OLSP potentially has an advantage over FE(xb).

In the case of out-of-sample forecasting an additional FE predictor is defined by generating estimates of γ from the following regression model:

$$(2) \quad \hat{\mu}_i = \theta + z_i'\gamma + \omega_i$$

where the estimated individual specific effects are regressed on the z's. These two-step estimates denoted by $\hat{\theta}, \hat{\gamma}$ are then used in conjunction with the first-step fixed effects estimates to form the predictor FE(xb+zg):

$$\Rightarrow \text{FE(xb + zg)}: \ \hat{y}_{it} = \hat{\alpha} + x_{it}'\hat{\beta} + \left(\hat{\theta} + z_i'\hat{\gamma}\right)$$
$$= (\hat{\alpha} + \hat{\theta}) + x_{it}'\hat{\beta} + z_i'\hat{\gamma}.$$

This estimator is discussed in Hsiao (1986) and is a Hausman and Taylor (1981) type estimator that would result under the assumption that any correlation between unobservables and regressors is confined to the time-varying regressors.

Using our data we compare the performance of alternative predictors, OLSP and all variants of FE(.), using concurrent and prospective specifications for both post-sample and out-of-sample forecasting tasks. Equation (1) depicts a concurrent specification where current period's expenditure is explained by current period covariates. Alternatively, the prospective model has next year's total health expenditure as the dependent variable and current period covariates. Performance is evaluated in terms of forecast mean squared errors (MSE) with the ranking based on minimizing MSE (MAPE was also calculated but leads to qualitatively the same results). We also compute the predictive ratio which is a group-level measure of predictive accuracy. It involves adding up the total predicted expenditure for a group of individuals and comparing that value to the actual expenditure for the same group. A predictive ratio that is closer to 1 indicates a better fit.

In the first instance, a Monte Carlo study is conducted in order to illustrate the key issues and tradeoffs involved in selecting the appropriate predictor. The Baillie and Baltagi (1999) experimental design is used as a base and extensions are restricted to cases where fixed effects produce consistent parameter estimates. Despite this restriction it is not a priori obvious that a fixed effects predictor is necessarily always superior to an OLS predictor. In Baillie and Baltagi (1999) the often substantial superiority of fixed effects over OLS derives from the individual heterogeneity in outcomes that is accommodated by estimating individual effects. When forecasting out-of-sample, this adjustment is not available and hence puts fixed effects predictors back on a more equal footing with the OLSP. Also what is required is a good approximation of the conditional mean function and this is not guaranteed by inserting consistent estimates of a subset of parameters, especially when these consistent estimators may have large variances. These arguments have prompted the inclusion of a simulation study in order to provide some guidance on what to expect when we undertake our extensive empirical analysis. The situation is further complicated when time-invariant variables are available. However, this additional issue is not addressed in the simulation study but is left to the substantive application that follows as is the comparison of concurrent and prospective specifications.

3. Monte Carlo experiment

3.1. Simulation design

The initial data generating process (DGP) to be considered is a variant of (1) where there are no time-invariant explanatory variables and a single time-varying covariate implying:

$$(3)\quad y_{it} = \alpha + x_{it}\beta + \mu_i + u_{it}$$

where $u_{it} \sim iid\ N(0, \sigma_u^2)$ and the DGP for x_{it} is given by

$$(4)\quad x_{it} = 0.1t + 0.5x_{i,t-1} + \varepsilon_i + \omega_{it}$$

where ω_{it} is uniformly distributed on the interval [-0.5, 0.5] and $x_{i0} = 5 + 10\omega_{i0}$ with the first 20 observations discarded. The unobserved individual effects are correlated with the explanatory variable through the following specification:

$$(5)\quad \mu_i = \varepsilon_i + \eta_i$$

with $\varepsilon_i \sim iid\ N(0, \sigma_\varepsilon^2)$ and $\eta_i \sim iid\ N(0, \sigma_\eta^2)$ which implies $\text{var}(\mu_i) \equiv \sigma_\mu^2 = \sigma_\varepsilon^2 + \sigma_\eta^2$ and $\text{cov}(x_{it}, \mu_i) = \sigma_\varepsilon^2$. This reduces to the Baillie and Baltagi (1999) experimental design when $\sigma_\varepsilon^2 = 0$. In order to facilitate comparisons, their results are reproduced in what is termed Experiment 1. This involves varying $\rho = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_u^2)$ as 0, 0.3, 0.6 or 0.9 with $\sigma_\mu^2 + \sigma_u^2$ fixed at 20 over 1,000 replications for each design point. In addition to one-period ahead post-sample forecasts, we also provide out-of-sample forecasts for a holdout sample of *n*=50 individuals. Experiment 2 then repeats the analysis adding correlation between the individual effects and the explanatory variable induced by setting $\sigma_\varepsilon^2 = 0.81$, a value that implies correlations between 0.10 and 0.25. All comparisons are done in terms of forecast mean squared errors (MSE).

Apart from introducing correlations between unobserved individual effects and the explanatory variable, the setup of Baillie and Baltagi (1999) needed to be extended in other respects to better reflect the type of applications we seek to explore. In their experiments they set *N*=50 or 500 and *T*=10 or 20 while in Experiments 1 and 2 we specify *N*=500 or 1,000 and *T*=3, 10 or 20. To get even closer to a more realistic situation, Experiment 3 is developed based on the actual data to be used in our application. Using total health expenditures (exp_{it}) as the dependent variable and number of standard (less than 20 minutes) visits to a general practitioner (GP) in a year (gp_{it}) as the sole explanatory variable, a fixed effects specification is estimated for *N*=1,000 and *T*=3 and the resulting parameter estimates are taken to be the "truth" to generate data for two panel configurations (*N*=500 or 1,000 and *T*=3). The resulting DGP is given by:

$$(6)\quad exp_{it} = 2.198 + 0.279gp_{it} + \mu_i + u_{it}$$

where $\sigma_\mu^2 = 46.06$ and $\sigma_u^2 = 41.65$ implying $\rho = 0.525$ and the gp_{it} values are fixed in this "real data" experimental design. Note that the estimated fixed effects are constrained to sum to zero implying that the intercept represents the overall mean of total health expenditures (in thousands). Further details on the data will be provided in Section 4.

The real data design is completed by specifying a source of omitted variable bias by generating the μ_i as follows:

$$(7) \quad \mu_i = \lambda sp_i + \eta_i$$

where sp_i is the mean number of initial specialist consultations in a year for each individual from our data normalized to have a unitary variance, again taken to be fixed over replications, and $\eta_i \sim iid\ N(0, \sigma_\eta^2)$. Setting $\lambda = 0$ and varying $\rho = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_u^2)$ while keeping fixed $\sigma_\mu^2 + \sigma_u^2 = 87.71$, mimics Experiment 1, but with a very different explanatory variable. Compared to the Baillie and Baltagi (1999) design, the contribution to total variation in the dependent variable due to the variation in unobservables is much smaller here as is the within relative to the between variation in the single explanatory variable. Because FE relies on the within variation we might expect a deterioration in the relative performance of the FE predictor in this alternative design.

Setting $\lambda = 1$ reproduces the sample correlation of 0.364 between sp_i and gp_{it} and provides a real data version of Experiment 2. While sp_i forms part of the DGP, it is taken to be unobservable by the researcher and thus is not part of the model specification estimated by OLS and FE. Because it is a time-invariant variable, this does not affect the FE estimation but will induce parameter biases when OLS is used and $\lambda \neq 0$.

3.2. Simulation results

Table 1 provides the MSE results for Experiment 1 when $\sigma_\varepsilon^2 = 0$. The post-sample forecasting analysis for *N*=500 and *T*=10 or 20 reproduces that part of Table 1 of Baillie and Baltagi (1999) corresponding to the predictors of interest here, namely OLSP and FE(xb+mu). The key feature of this subset of the results is that the FE predictor dominates OLSP except when $\rho = 0$ and the difference increases markedly as ρ increases. Baillie and Baltagi (1999) demonstrate that FE runs a close second to the operational optimal predictor (they term this the "ordinary predictor") based on the true random effects DGP and emphasize the importance of accounting for individual effects in estimation and prediction. Further they conclude that the FE predictor is recommended in practice because of this relatively good simulation performance and its robustness to correlation between random effects and regressors. Note that they do not extend their experimental design to provide simulation evidence in support of the robustness claim.

The remaining results in Table 1 provide some initial qualifications to the main findings of Baillie and Baltagi (1999). First, the post-sample forecasts when *T*=3 exhibit a similar pattern except there is a relative deterioration in FE(xb+mu) attributable to the reduced within variation in the regressor that is associated with the short time dimension. This means that for $\rho = 0,$ the OLSP dominates FE(xb+mu) by a greater margin than when *T* is larger. The second feature relates to the out-of-sample forecasting section of the results where the superiority of FE over OLS is eliminated and the performance of the two predictors, FE(xb) and OLSP is essentially the same for all *N* and *T* combinations. This serves to emphasize that the importance of accounting for individual effects relates to their use in prediction rather than estimation. In out-of-sample forecasting, FE(xb) accounts for individual effects in estimation but because forecasting relates to a new sample of individuals, only a single overall mean effect can be used for prediction.

Table 1: Post-sample and out-of-sample mean squared errors: Experiment 1

	Post-sample		Out-of-sample	
	OLSP	FE(xb+mu)	OLSP	FE(xb)
	MSE	**MSE**	**MSE**	**MSE**
***N*=500, *T*=10**				
ρ=0	20.045	22.056	20.027	20.027
ρ=0.3	20.004	15.450	19.969	19.969
ρ=0.6	19.971	8.828	19.924	19.922
ρ=0.9	19.947	2.207	19.884	19.882
***N*=500, *T*=20**				
ρ=0	19.995	20.994	20.028	20.028
ρ=0.3	19.998	14.661	20.061	20.061
ρ=0.6	20.001	8.425	19.994	19.994
ρ=0.9	20.001	2.097	20.065	20.065
***N*=500, *T*=3**				
ρ=0	20.030	26.657	20.064	20.077
ρ=0.3	20.012	18.675	20.064	20.067
ρ=0.6	20.016	10.671	20.106	20.102
ρ=0.9	20.023	2.668	20.108	20.096
***N*=1000, *T*=3**				
ρ=0	20.014	26.632	20.151	20.156
ρ=0.3	20.034	18.673	20.049	20.052
ρ=0.6	20.012	10.670	20.157	20.157
ρ=0.9	19.980	2.668	20.285	20.281

Notes: (i) The first two *N, T* pairs were chosen to be comparable with the design choices of Baillie and Baltagi (1999) while the second two *N, T* pairs were chosen to be more representative of micro panels.
(ii) OLSP is the OLS predictor; FE(xb) is the FE predictor without the estimated individual specific effects; FE(xb+mu) is the FE predictor with the estimated individual specific effects.

Table 2 provides the MSE results for Experiment 2 where OLS no longer produces consistent parameter estimates because of the non-zero correlation between the explanatory variable and the unobserved time-invariant effects. There are no results for $\rho = 0$ which implies no unobserved time-invariant effects. Setting $\sigma_\varepsilon^2 = 0.81$ implies modest levels of correlation (0.10 – 0.25) between the explanatory variable and the unobserved effects but the biases in the OLS parameter estimates are substantial. For example, with $N = 1{,}000,\ T = 3$ and $\rho = 0.9$, the mean correlation calculated over the 1,000 replications of the experiment was 0.143 while the means of the OLS estimates of $\alpha = 5$ and $\beta = 0.5$ were 1.452 and 1.347, respectively. Despite these large parameter biases the pattern in the post-sample MSEs was largely unchanged from Experiment 1: the FE predictor dominates OLSP and the difference increases as ρ increases.

Table 2: Post-sample and out-of-sample mean squared errors: Experiment 2

	Post-sample		Out-of-sample	
	OLSP	FE(xb+mu)	OLSP	FE(xb)
	MSE	**MSE**	**MSE**	**MSE**
***N*=500, *T*=10**				
ρ=0.3	19.863	15.447	19.514	20.051
ρ=0.6	19.762	8.780	19.542	20.074
ρ=0.9	19.890	2.205	19.622	20.052
***N*=500, *T*=20**				
ρ=0.3	20.092	14.708	19.714	19.999
ρ=0.6	20.121	8.411	19.718	19.984
ρ=0.9	20.161	2.094	19.755	20.019
***N*=500, *T*=3**				
ρ=0.3	19.386	18.660	19.471	20.241
ρ=0.6	19.471	10.721	19.352	20.175
ρ=0.9	19.316	2.670	19.474	20.202
***N*=1000, *T*=3**				
ρ=0.3	19.408	18.667	19.467	20.159
ρ=0.6	19.363	10.677	19.435	20.164
ρ=0.9	19.414	2.671	19.408	20.034

Notes: (i) The first two *N, T* pairs were chosen to be comparable with the design choices of Baillie and Baltagi (1999) while the second two *N, T* pairs were chosen to be more representative of micro panels.
(ii) OLSP is the OLS predictor; FE(xb) is the FE predictor without the estimated individual specific effects; FE(xb+mu) is the FE predictor with the estimated individual specific effects.

The pattern in the out-of-sample forecasting section of the results has changed though. In Experiment 1 there was essentially no difference between FE(xb) and OLSP. Now, while the differences are still modest, OLSP uniformly dominates FE(xb) despite the fact that OLS produces severely biased estimates of the true parameter values.

The out-of-sample superiority of OLSP over FE(xb) is an artefact of a more general phenomenon. Consider a classical linear regression variant of (1) where $\mu_i = 0$ implying that the optimal predictor, in a lowest MSE sense, would require replacing the unknown parameters with their OLS estimates. If instead z_i is not observed, giving rise to a standard omitted variable situation, then the "mis-specified" predictor is now only a function of x_{it}. Using OLS to estimate the short regression produces consistent estimates of population parameters that incorporate the biases due to the impact of omitting z_i and corresponds to the best linear predictor of y_{it} conditional on x_{it} alone. With panel data, one can obtain consistent estimates of α and β using FE but using these in the prediction equation that is solely a function of x_{it} means that there is no account being taken of z_i for the purposes of prediction and will produce poor forecasts in the case where omitted variable biases occur. It is essentially the same as taking the optimal predictor when both x_{it} and z_i are available in (1) but then specifying $\gamma = 0$.

Table 3: Post-sample and out-of-sample mean squared errors: Experiment 3

	Post-sample		Out-of-sample	
	OLSP	FE(xb+mu)	OLSP	FE(xb)
	MSE	**MSE**	**MSE**	**MSE**
Uncorrelated				
***N*=500, *T*=3**				
ρ=0	87.465	116.579	88.616	93.835
ρ=0.3	87.704	81.527	87.966	93.185
ρ=0.6	87.612	46.793	88.023	93.306
ρ=0.9	87.200	11.720	87.961	92.745
***N*=1000, *T*=3**				
ρ=0	87.696	116.856	88.083	93.327
ρ=0.3	87.712	81.908	87.564	92.785
ρ=0.6	87.820	46.859	88.384	93.230
ρ=0.9	87.733	11.684	87.940	92.957
Correlated				
***N*=500, *T*=3**				
ρ=0.3	87.253	81.898	87.597	92.102
ρ=0.6	87.550	46.775	88.255	92.698
ρ=0.9	87.389	11.683	89.578	93.200
***N*=1000, *T*=3**				
ρ=0.3	87.453	81.802	87.591	92.244
ρ=0.6	87.365	46.798	87.422	92.088
ρ=0.9	87.341	11.669	87.588	92.062

Notes: (i) Here only the *N, T* pairs more representative of micro panels are used but results are provided when the individual effects are uncorrelated with the explanatory variable (first two panels) and when there is correlation (second two panels).
(ii) OLSP is the OLS predictor; FE(xb) is the FE predictor without the estimated individual specific effects; FE(xb+mu) is the FE predictor with the estimated individual specific effects.

Table 3 provides the MSE results for Experiment 3 where the real data design is utilized. The table is partitioned into two parts; the first part mirrors Experiment 1 where the explanatory variable is uncorrelated with the unobserved time-invariant effects, and the second part, where such correlation is introduced, mirrors Experiment 2. Qualitatively the overall results are similar. Even with a very different explanatory variable, taken from our data, FE(xb+mu) does well in post-sample forecasting irrespective of whether OLS parameter estimates are impacted by omitted variable biases or not. The situation with out-of-sample results has changed somewhat in that OLSP now dominates FE(xb) in both the uncorrelated and correlated cases. Previously they were almost identical in the uncorrelated case but now the within variation is relatively small compared to the between variation and this adversely impacts the performance of the FE predictor.

These Monte Carlo results are not meant to represent an exhaustive investigation of issues relevant to choice of predictors with micro panel data. Rather they serve to highlight some important dimensions of the comparison between FE and OLS predictors that were not evident in the work of Baillie and Baltagi (1999), and they provide some indication of what to expect in the case study to follow. In particular, we expect the OLSP to perform relatively better in out-of sample forecasting task compared to the post-sample forecasting task.

4. Data

Our data are derived from the 45 and Up Study of 267,153 New South Wales (NSW) residents linked to several administrative data sources of health costs from 2006 to 2009: hospital inpatient data and emergency department (ED) data (linked by the NSW Centre for Health Record Linkage http://www.cherel.org.au/), Medical Benefits Schedule (MBS) data for medical services such as GP and specialist consultations and Pharmaceutical Benefits Scheme (PBS) data of prescription drugs for which a government subsidy was paid. The survey was collected only once during this period (45 and Up Study Collaborators, 2008), but the health records of the survey respondents are a panel. We exclude voluntary participants and those respondents with invalid age (0.1%), and those who died during the study period. The final sample is 1,056,096 person-years. The average age of the survey respondents is 63 years.

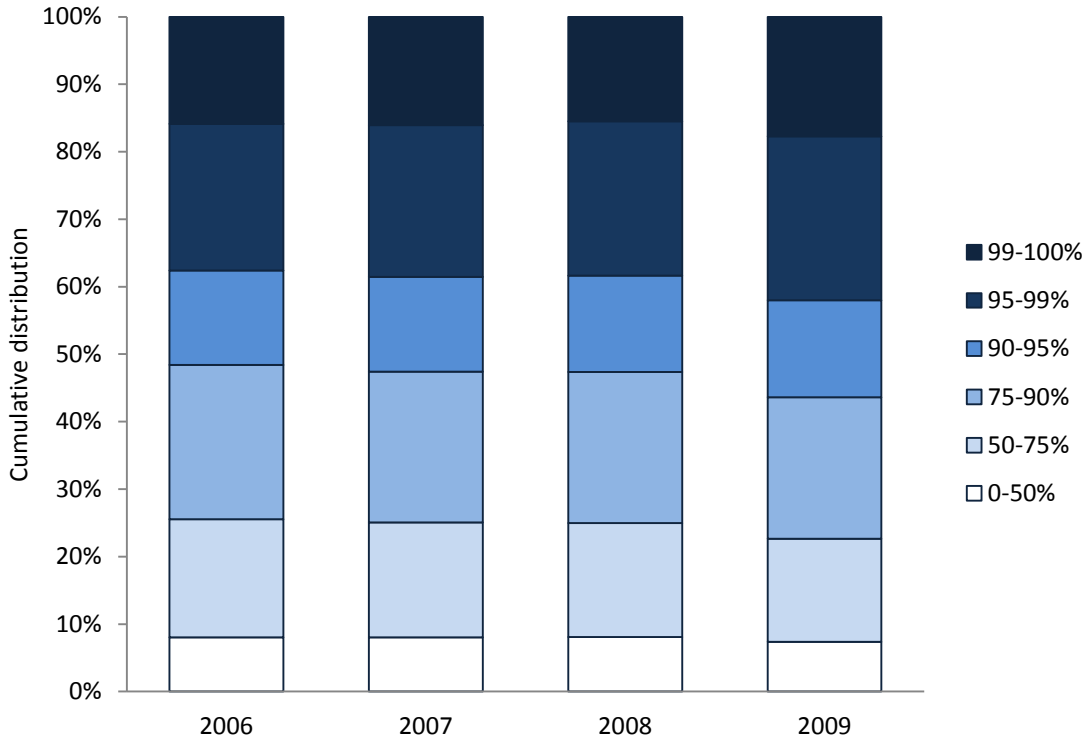
Individual annual total health expenditure is calculated as the sum of costs of hospital services, charges for MBS items and prices of PBS drugs in any given year. The cost of hospital services is imputed using the *NSW Costs of Care Standards 2009/10* guidelines released by NSW Department of Health. For hospitalization, it varies by diagnostic group, type of treating hospital, type of care (e.g., overnight or same day), length of stay, intensive care unit (ICU) hours and the use of ventilation

machine. For ED presentations, cost varies with hospital type, urgency category and whether or not the patient is subsequently admitted. All expenditures are annual and indexed to constant $2009.

A common feature of health expenditure data is positive skewness. Figure 1 illustrates, if the population is sorted by its health expenditure each year, the top 25% of the population accounts for almost 80% of total expenditure. The very highest health expenditure is concentrated in the top 1%. 50% of the population who incurs the least amount of spending accounts for less than 10% of total expenditure. This pattern is consistent across all years, suggesting the absence of any relevant structural break in demand within our study period that needs to be accounted for.

Another typical feature of health expenditure data is a large mass of zero expenditure. However in our data, less than 3% of individuals have zero expenditure in any given year. This could be explained by our older sample of individuals, whose demand for health services is relatively higher than the general population, and the Australian universal public health system which ensures access of health services by all. The mean expenditure was $3,450 in 2006, $4,050 in 2007, $4,670 in 2008 and $5,004 in 2009.

Figure 1: Health expenditure distribution by year



Note: For each year of the data the cumulative distribution of health expenditure is presented for the indicated percentages of the sample. For example, the white (lowest) bar indicates that 50% of people with the lowest health expenditure account for less than 10% of the total health expenditure and this is true in all years, 2006-2009. For a symmetric distribution the cumulative distribution should be 50% for this segment. The second lowest bar indicates that 75% of people with the lowest health expenditure account for less than 30% of the total health expenditure, and so on.

Time-varying regressors (x_{it}) are diagnoses during hospitalization and drug groups. To summarize this rich information into a manageable number of variables to be put into the regression model, we use a US-based risk adjustment tool called DxCG Risk Solutions developed by Verisk Health. This is a standard approach in the risk adjustment literature. The software, which extends the classification system used by the US Medicare program for paying competing health plans, organizes diagnose codes (International Statistical Classification of Diseases version 10) and drug codes (Anatomical Therapeutic Chemical Classification) into a large number of non-mutually exclusive categories, and imposes hierarchies on diseases and drugs so that more serious or expensive conditions take precedence over less serious or expensive conditions. The software also performs a number of data-cleaning steps to identify illegal (e.g. coding errors) or invalid (e.g. male pregnancies) diagnoses. It has been used by numerous academic papers in the US such as Ash et al. (2001), Einav et al. (2013) and Zhao et al. (2005). The result of the grouping is 110 non mutually exclusive dummy variables for diagnoses (Related Health Conditions, RCCs) and 123 non mutually exclusive dummy variables for drugs (RX groups). Age and its interaction with gender are also included as time-varying regressors.

All other variables from the 45 and Up Survey are time-invariant, some because they are only measured at one point in time. To explore the role of various observed individual specific factors in explaining variation in health expenditure, we define 3 sets of time-invariant variables. The first set $\mathrm{z}_{1\mathrm{i}}$ consists of basic demographic characteristics, such as sex, marital status, education, the possession of a health care concession card, region, foreign born and language status and skin colour to capture ethnicity. The second set $\mathrm{z}_{2\mathrm{i}}$, augments $\mathrm{z}_{1\mathrm{i}}$ by including self-reported health variables such as self-assessed general health and major chronic illnesses, such as diabetes, hypertension, cancers, heart disease, broken bone and asthma. The third set $\mathrm{z}_{3\mathrm{i}}$, augments $\mathrm{z}_{2\mathrm{i}}$ by including socioeconomic characteristics, such as income, employment and private health insurance status, and lifestyle variables such as smoking, obesity and alcohol consumption. The last set of additional variables is potentially endogenous.

5. Forecasting results

All of the analyses use data drawn from a balanced panel of 264,024 individuals observed for four years. For the prospective specification only three years of data are available. These data are then divided into estimation and holdout samples that vary depending on the type of forecasting analysis being undertaken.

In the first set of results provided in Table 4 we consider post-sample prediction. Here the holdout sample comprises one year of data for all individuals. Predictive MSEs are reported for both concurrent and prospective specifications as are comparable within sample measures as a baseline. Within sample, and for both concurrent and prospective specifications, the ranking of predictors is the same with FE(xb+mu) being better than OLSP which in turn is better than FE(xb). What is noteworthy is that the superiority of OLSP over FE(xb) is more pronounced in the prospective specification. What is also clear is the superiority of the concurrent specification over the prospective, but because the former captures contemporaneous associations, this is not unexpected.

Setting aside one or two years of data for post-sample prediction the superiority of FE(xb+mu) disappears. In all cases and for both the concurrent and prospective specifications, OLSP produces the lowest predictive MSE. For the concurrent specification the differences are relatively small across all three predictors but FE(xb+mu) is not necessarily superior to FE(xb). For the prospective case the post-sample ordering is clearer with the use of estimated individual fixed effects in FE(xb+mu) improving prediction performance relative to FE(xb) although still remaining inferior to the OLSP.

Table 4: Predictive mean squared error comparison of alternative predictors: Post-sample

	Concurrent MSE	**Prospective MSE**
Within sample fit		
OLSP	30.24	76.19
FE(xb)	31.04	104.7
FE(xb + mu)	18.78	40.58
Prediction, T = 3, τ=1		
OLSP	48.68	-
FE(xb)	49.81	-
FE(xb+mu)	49.65	-
Prediction T = 2, τ=1		
OLSP	30.94	106.1
FE(xb)	32.59	158.0
FE(xb+mu)	33.20	122.2
Prediction T = 2, τ=2		
OLSP	49.76	-
FE(xb)	51.30	-
FE(xb+mu)	54.60	-

Note: *N*=264,024 in all cases. For within sample fit *T*=3 for prospective and *T*=4 for concurrent. τ values indicate whether the prediction is one or two periods ahead.

Note that unlike the earlier simulation results, the predictive models include both time-varying and time-invariant predictors. These will remain in the OLSP and one would expect this feature to assist predictive performance relative to FE(xb). While the time-invariant predictors will also be absent from FE(xb+mu), one expects the use of estimated individual fixed effects which are so important in producing a good within sample fit, to compensate. For our application this proves not to be the case in the post-sample results presented in Table 4.

One might expect the role of time-invariant regressors to be even more important for the case of out-of-sample prediction because here FE(xb+mu) is not a feasible predictor. There are numerous time-invariant regressors available for use in the predictor specifications and many of them might be suspected of being endogenous. Recall that moving from z1 to z3, more time-invariant regressors are added but at the same time the exogeneity of these additional regressors becomes more problematic.

Table 5: Predictive mean squared error comparison of alternative predictors: Out-of-sample

	Concurrent MSE	**Prospective MSE**
Within sample fit		
OLSP(z1)	29.55	74.92
OLSP(z2)	29.53	74.36
OLSP(z3)	29.51	74.20
FE(xb)	30.26	102.6
FE(xb+mu)	18.19	39.19
FE(xb+zg1)	30.15	96.52
FE(xb+zg2)	30.11	90.66
FE(xb+zg2)	30.09	89.85
Prediction		
OLSP(z1)	33.37	85.23
OLSP(z2)	33.34	84.68
OLSP(z3)	33.33	84.54
FE(xb)	34.12	113.7
FE(xb+zg1)	34.02	107.4
FE(xb+zg2)	33.99	101.5
FE(xb+zg3)	33.97	100.7

Notes: (i) In all cases *T*=3 for the prospective and *T*=4 for the concurrent specifications. *N*=211,728 for the within sample fit then models are used to predict for a 20% holdout sample of *n*= 52,932.
(ii) Three variants of the OLS predictor (OLSP) and the FE predictor (FE(xb+zg) are defined depending on which of 3 sets of time invariant explanatory variables are used. Moving from z1 to z3, (and hence from OLSP(z1) and FE(xb+zg1) to OLSP(z3) and FE(xb+zg3)) more time-invariant regressors are added but at the same time the exogeneity of these additional regressors becomes more problematic.

Table 5 provides the analysis of out-of-sample predictions. Here the holdout sample refers to the 2009 expenditures for a randomly selected 20% sample of 52,932 individuals. In terms of both

within sample fit and out-of-sample predictions and in both concurrent and prospective specifications, it makes no substantive difference to the performance of OLSP which *z*'s are used. Within sample fit and out-of-sample predictive performance of FE(xb+zg) changes little with choice of *z*'s for the concurrent specification but does deliver modest improvement in the prospective specification as more *z*'s are added. However, OLSP is the best performing predictor although the differences are relatively small for the concurrent specification whereas they are substantial in the prospective specification.

Thus far, the predictive performances have been gauged on the basis of predictive MSEs but the relative performance of alternative predictors may very well be sensitive to choice of metric. As an alternative, selected comparisons are repeated using the predictive ratio that is commonly used in the risk adjustment literature. Unlike predictive MSE that measures model performance at the level of the individual, risk ratios measure performance at a group level. For each decile of actual expenditure, we calculate the ratio of the aggregate actual expenditure to the aggregate predicted expenditure. These ratios are represented in Figures 2 to 4 and predictor superiority is judged by how close these ratios are to unity.

Figure 2: Predictive risk ratios for post-sample prediction

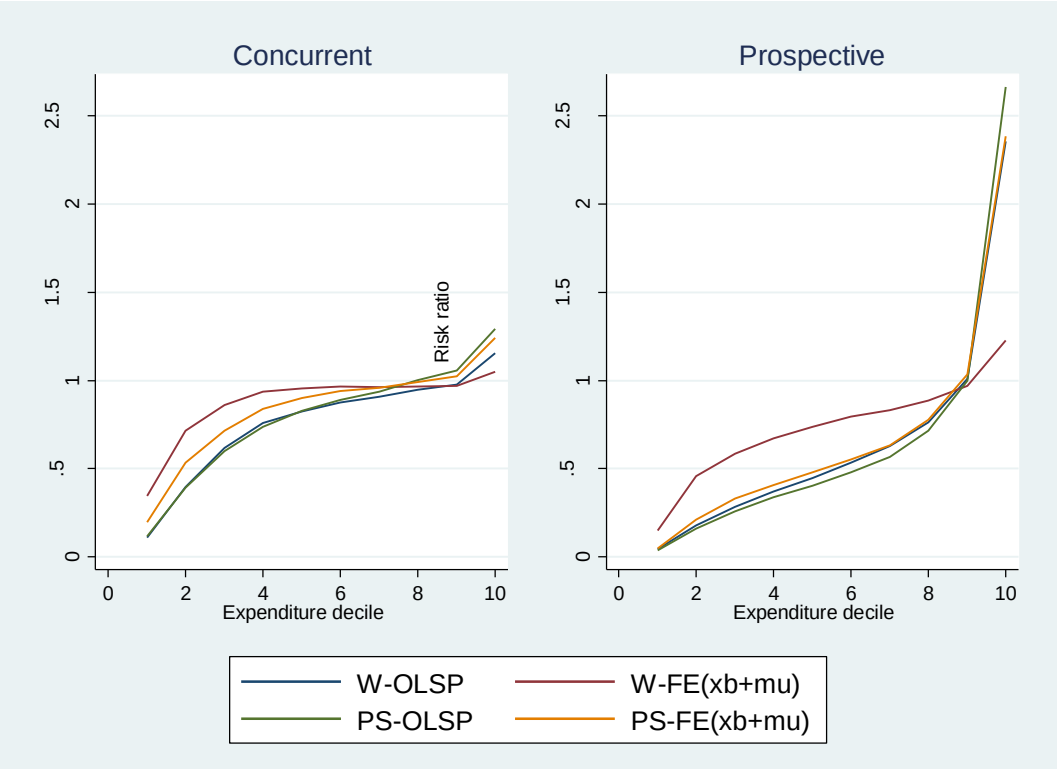


Figure 2 provides a comparison of predictive performance post sample for both concurrent and prospective specifications, together with within sample performance as a benchmark. In both concurrent and prospective specifications, overestimation of expenditures is the norm, risk ratios

are almost always less than unity for low risk (expenditure) deciles 1 through 9. The extent of overestimation is often extreme and this is more so for the prospective case. Expenditure in the highest decile is underestimated and again this is much more pronounced for the prospective specification. The ranking of predictors is stable across specifications. The fixed effect estimator, FE(xb+mu), dominates OLSP both within sample and post-sample.

Turning to the out-of-sample results, the concurrent and prospective specifications are provided separately in Figures 3 and 4. Because the choice of time-invariant regressors had limited impact on the performance of alternative predictors, all time-invariant regressors are included in OLSP and FE(xb+zg). For both the concurrent and prospective specifications, there is a recurrence of the pattern observed in Figure 2, where expenditure in all deciles except the top decile are typically overestimated and the expenditures in the largest decile are underestimated. Also, FE(xb+mu) dominates within sample but is not operational for out-of-sample forecasts.

The performance of the feasible concurrent predictors in Figure 3 is very similar, although the OLSP does better than either of the fixed effects predictors for most of the deciles. There is little degradation in performance out-of-sample relative to within sample for these predictors but this simply confirms the representativeness of the chosen holdout sample.

Figure 3: Predictive risk ratios for out-of-sample prediction with concurrent specifications

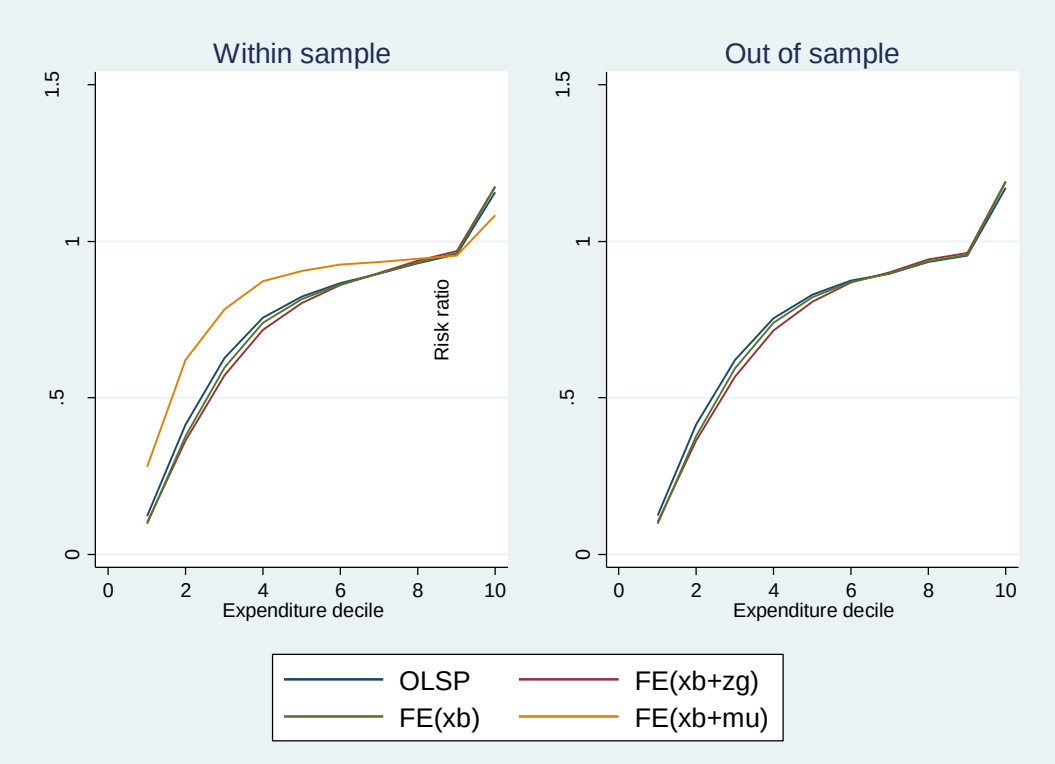
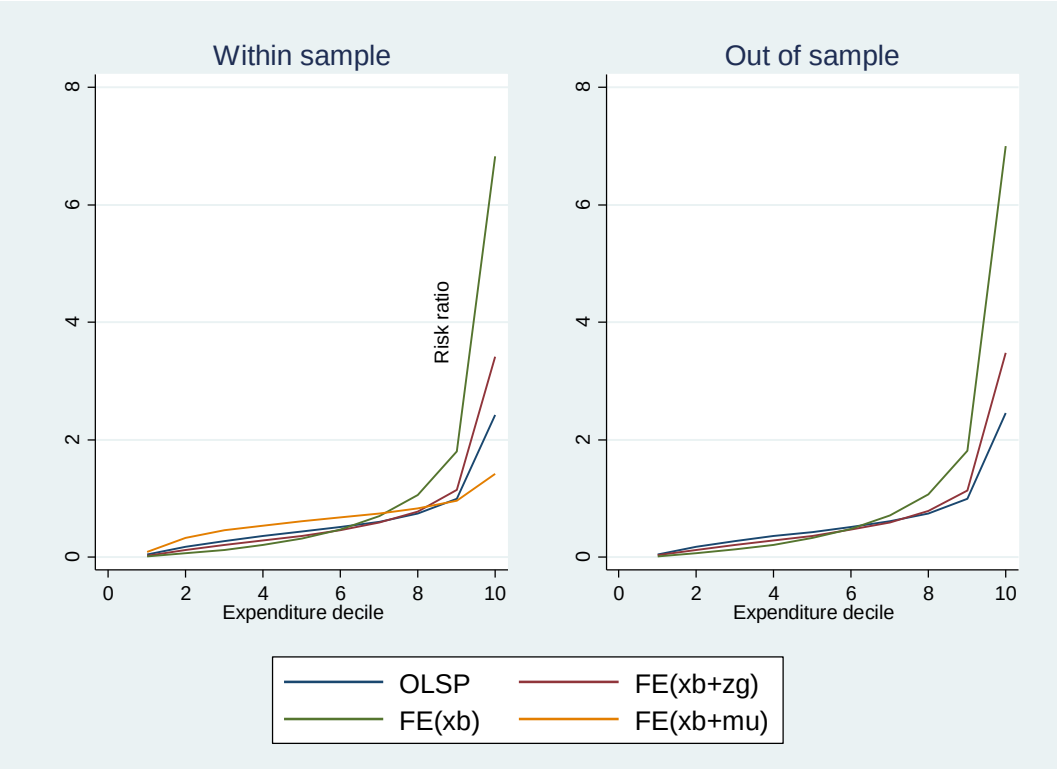


Figure 4 again highlights the poorer performance of the prospective model relative to that of the concurrent specification. Just as in the case of the concurrent specification, the OLSP tends to outperform the available variants of fixed effects for most deciles.

Figure 4: Predictive risk ratios for out-of-sample prediction with prospective specifications



6. Discussion

Overall, the results of the risk adjustment case study point to a clear preference for the OLSP over all of the variants of fixed effects predictors that were considered. There were some exceptions where fixed effect predictors proved superior but these were limited to some of the comparisons made on the basis of risk ratios (group-level prediction).

The superiority of the OLSP for out-of-sample prediction was anticipated from the simulation evidence that was presented. For fixed effects predictors to be competitive they need to include estimated individual effects otherwise the possibly better (less biased) estimates of the coefficients of the time-varying variables are detrimental to producing good forecasts. In contrast, the OLSP will be the best linear predictor conditional on the available regressors. One might anticipate that incorporating time-invariant regressors to allow for estimated individual specific effects might favourably impact on the relative performance of the fixed effects predictors. This conjecture was considered in our analysis of health expenditures. Using an extensive list of extra

time-invariant regressors to predict individual specific effects did indeed improve fixed effect predictors for use out of sample, but this improvement was not enough to overcome the superiority of OLSP.

Somewhat surprising, given the simulation evidence, was the relative performance of predictors in the case of post-sample prediction. Using the estimated individual specific effects in time series forecasting did not uniformly improve the forecasting performance of the fixed effect predictors and the OLSP was the best performing predictor in all cases when predictive MSE was the performance measure. This result proved sensitive to the performance measure and when predictive ratios were used to evaluate post-sample performance, the fixed effects predictor did outperform the OLSP. This was the only situation where this ranking occurred.

The relative performance of the OLSP and FE(xb+mu) in post-sample prediction is impacted by a number of factors. In the simulation evidence FE(xb+mu) did improve monotonically with ρ, the proportion of the variability in the unobservables attributable to the fixed effects. Holding constant the signal to noise ratio means the OLSP is unaffected, and so for small enough values of ρ, the OLSP is expected to dominate FE(xb+mu). In our analysis of health expenditures, the estimated ρ values were in the range 0.2 to 0.5 depending on the sample and whether the prospective or concurrent specification was used. So while in general this is a possible explanation of the relatively poor performance of FE(xb+mu), these estimates of ρ suggest this is not the explanation here.

Instead, two other contributing factors are likely to be the main reasons for the relatively poor performance of FE(xb+mu). The first of these is the extent of within variation, which if limited, as it was in the risk adjustment case study (and likely to be in most micro panel data), will result in deterioration in the relative performance of fixed effects predictors. The second and possibly most important factor that influences the relative performance of predictors is the quality of the estimates for the individual specific effects. The dominance of FE(xb+mu) within sample derives from the inclusion of these individual effects. But even though non-zero individual effects may exist, in cases such as ours with small *T* available to estimate these effects, they are likely to be estimated with considerable variability that contributes to higher predictive forecast variability. The consequence is that a simpler model, here OLSP, may prove superior in terms of predictive MSE.

This result, where the simpler OLSP performs well, is consistent with evidence drawn from the literature comparing homogenous and heterogeneous panels; i.e. whether coefficients on regressors are allowed to vary over individuals or not. Such comparisons require panels with at least modest *T* values and so are not strictly transferable to the micro panel case. But there is a common theme that emerges in this literature that Baltagi (2008) describes as:

"*...while the performance of various estimators and their corresponding forecasts may vary in ranking from one empirical example to another, the consistent finding in all studies is that homogenous panel data estimators perform well in forecast performance mostly due to their simplicity, their parsimonious representation and the stability of parameter estimates.*"

The other dimension of the comparisons made, is that between concurrent and prospective specifications. The superiority of a concurrent model in terms of lower forecast errors is known. Using German data, the MAPE from a prospective model is more than 60% greater than that from the concurrent model (Behrend et al. 2007). Using Taiwan data, the corresponding figure is about 25% (Chang and Weiner, 2010). The results presented here demonstrate similarly large differences in performance. This is primarily because the concurrent specification captures more of the costs of actual utilization during a year. However, from a post-sample forecasting perspective such a specification does not represent a truly operational predictive tool. On the other hand, the prospective specification relies on past factors and so predictions of future utilization are readily generated. While the concurrent model more accurately reflects actual spending, for payers using risk adjustment models, the prospective model gives advance indication of what their financial obligations will be. Thus for payment purposes, the prospective model is used more often than the concurrent model; concrete applications include social health insurance in Germany and Medicare Shared Savings program in the US.

To put the concurrent specification on an equal footing with the prospective specification requires the use of predicted covariates to make the associated predictors operational. This is not an approach implemented here and the question of which of these approaches is better is left for further research. What has been emphasized in this paper is the relative performance of alternative predictors when used in conjunction with different specifications and for use in post-sample and out-of-sample prediction.

7. Conclusion

A rich data set comprising the linkage of a large cohort-representative survey to several years of comprehensive health records provides a test bed to explore the relative forecasting performance of alternative models of health expenditures. In contrast to much of the risk adjustment literature, where such modelling is prevalent, our focus is on predictors that exploit the availability of panel data. We also stress the distinction between predictions made out-of-sample and post-sample that is possible when panel data are used. While it is unwise to draw strong conclusions on the basis of our single case study and a somewhat limited extension of the simulation

results of Baillie and Baltagi (1999), there are a number of general issues that our work highlights in terms of forecasting with micro panels that feature a large number of individuals but limited time periods.

First, the strong preference for fixed effects over OLS estimators when parameter estimation is the goal does not readily extend to situations where prediction is being undertaken. While existing simulation evidence also supports the use of a fixed effects approach when forecasting, we demonstrate this is a fragile conclusion that is likely to be overturned in many practical situations, including our models of health care costs. Simulation results add support and additional insights into the results obtained in the application. These results are supportive of the use of the OLSP in a wide range of circumstances.

While fixed effects predictors may prove useful in other applications, favorable circumstances would need to exist for this to happen. These would include a relatively high proportion of the variability in the unobservables attributable to individual specific effects and having regressors with considerable within variation. A third factor that would be helpful is to have a relatively large number of time series observations in order to better estimate individual specific effects. However, such a situation takes us out of the realm of micro panels into a situation where alternative approaches and model specifications might be entertained.

Despite the limited success of predictors that explicitly utilize the panel structure there are advantages of having panel data that have not been highlighted but nonetheless need to be recognized. First, the prospective model that arguably is the more useful specification for risk adjustment requires panel data. Second, there is likely to be gains in predictive performance that derives from simply having extra data. More specifically in the particular case of risk adjustment, the key set of predictors is associated with hospital diagnoses, which come sporadically. Panels provide richer data in this very important domain.

References

45 and Up Study Collaborators. Cohort profile: The 45 and Up Study. *International Journal of Epidemiology* 2008, 37(5): 941-947.

Aizcorbe A, Liebman E, Pack S, Cutler D, Chernew M, Rosen A. Measuring health care costs of individuals with employer-sponsored health insurance in the U.S.: A comparison of survey and claims data. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 2012, 28(1): 43-51.

Albouy V, Davezies L, Debrand T. Health expenditure models: A comparison using panel data. *Economic Modelling* 2010, 27(4): 791-803.

Ash AS, Zhao Y, Ellis RP, Kramer MS. Finding future high-cost cases: Comparing prior cost versus diagnosis-based methods. *Health Services Research* 2001,26(6) Part II December 194-206.

Baillie RT, Baltagi BH. Prediction from the regression model with one-way error components. In *Analysis of panels and limited dependent variable models*, Hsiao, C, Lee LF, Pesaran, H (eds.), Cambridge University Press, 1999, 225-267.

Baltagi BH. Forecasting with panel data. *Journal of Forecasting* 2008, 27(2): 153-173.

Behrend C, Buchner F, Happich M, Holle R, Reitmeir P, Wasem J. Risk-adjusted capitation payments: how well do principal inpatient diagnosis-based models work in the German situation? Results from a large data set. *European Journal of Health Economics* 2007, 8: 31-39.

Bernard D, Farr S, Fang Z. National estimates of out-of-pocket health care expenditure burdens among nonelderly adults with cancer: 2001 to 2008. *Journal of Clinical Oncology* 2011, 29: 2821-2826.

Chang HY, Weiner J. An in-depth assessment of a diagnosis-based risk adjustment model based on national health insurance claims: the application of the Johns Hopkins Adjusted Clinical Group case-mix system in Taiwan. *BMC Med* 2010, 8: 7.

Einav L, Finkelstein A, Ryan S, Schrimpf P, Cullen M. Selection on moral hazard in health insurance. *American Economic Review* 2013, 103(1): 178–219.

Ellis RP. Risk adjustment in health care markets: concepts and applications. In *Financing Health Care: New Ideas for a Changing Society*, Lu M and Johnson E (eds.) Wiley, 2008, 177-222.

Ellis RP, Fiebig DG, Johar M, Jones G, Savage E. Explaining health care expenditure: Large-sample evidence using linked survey and health administrative data. *Health Economics* 2013, 22: 1093-1110.

French E, Jones JB. On the distribution and dynamics of health care costs. *Journal of Applied Econometrics* 2004, 19(6): 705-721.

Hausman JA, Taylor WE. Panel data and unobservable individual effects. *Econometrica* 1981, 49(6): 1377-1398.

Hsiao C. *Analysis of panel data*. Cambridge University Press, 1986.

Hill SC, Miller GE. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics* 2011, 19: 608–627.

Jones A, Lomas J, Rice N. Applying Beta-type size distributions to healthcare cost regressions. forthcoming *Journal of Applied Econometrics* 2013.

Jones A. Models for health care. in *Oxford Handbook of Economic Forecasting* , Hendry, D and Clements, M (eds.), Oxford, Oxford University Press, 2011, 625-654.

Jones A. Panel data methods and applications to health economics. in *Palgrave Handbook of Econometrics Vol 2*, Mills TC and Patterson K (eds.), London Palgrave MacMillan, 2009, 557-631.

New South Wales Health. *Cost of Care Standards 2009/10*. New South Wales (AUST): NSW Health; 2011.

Seshamani M, Gray J. A longitudinal study of the effects of age and time to death in hospital costs. *Journal of Health Economics* 2004a, 23(2): 217-235.

Seshmani M, Gray J. Ageing and health-care expenditure: The red herring argument revisited. *Health Economics* 2004b, 13 (4): 303-314.

Stearns S, Norton E. Time to include time to death? The future of health care expenditure predictions. *Health Economics* 2004, 13(4): 315-327.

Wooldridge JM. Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 2005, 20(1): 39-54.

Zhao Y, Ash AS, Ellis RP, et al. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Medical Care* 2005, 43(1): 34-43.

Zuvekas S, Olin G. Accuracy of Medicare expenditures in the Medical Expenditure Panel Survey. *Inquiry* 2009, 46(1): 92-108.

Zweifel P, Felder S, Meier M. Ageing of population and health care expenditure: A red herring? *Health Economics* 1999, 8: 485-496.