# Random Utility Pseudo Panel Model and Application on Car Ownership Forecast

Huang, Biao

Dept. of Economics, Birkbeck College

2007

# Random Utility Pseudo Panel Model and Application on Car Ownership Forecast

Biao Huang[i]
Dept. of Economics, Birkbeck College

Car ownership forecast is an important tool in many policy areas and business sectors including transport, environment, energy and car manufacturing. It has been extensively researched and a large number of models have been developed. A review of international literature has identified two clear trends: from static to dynamic and from aggregate to disaggregate. Car ownership models were traditionally dominated by static approach, and static models still remain the most common type for forecasting purpose. However, dynamic models of car ownership have become a thriving area of research in the past two decades, with many classes of models utilizing a diverge range of theories and methodologies (e.g. dynamic car holding models such as time series model, equilibrium market model and panel data model; dynamic transaction models such as duration model, competing-risk-duration model and dynamic random utility model). Some early studies adopted aggregate trend extrapolation or time series methods, which have the drawback of not being able to include the important influences such as demographic factors. In more recent years, disaggregate models have become the dominant form of car ownership model, and this is the case for both static and dynamic models.

The trend towards dynamic and disaggregate models puts much heavier requirements on data. The panel data is the preferred form of longitudinal data, but they are difficult and expensive to collect so there are very few high quality panel data available. Furthermore, panel data suffer the problem of attrition, which can be very severe for long running surveys. One way to avoid the collection of expensive panel data is to use retrospective survey, where the respondents provide information on their vehicle holding and transactions in the past years. This is a common approach used in many dynamic transaction models. However, retrospective survey has a major shortcoming that it can at best collect limited past information of household characteristics and other relevant variables, so most dynamic transaction models have no or very few time-varying covariates (explanatory variables). Another approach to estimate dynamic disaggregate models without the need for panel data is to construct pseudo panel from the rich sources of repeated cross sectional surveys. This is the method adopted in few previous studies and is the main focus of the current project.

The pseudo-panel approach is a relatively new econometric approach to estimate dynamic demand models. A pseudo-panel is an artificial panel based on (cohort) averages of repeated cross-sections. The cohorts are defined based on time-invariant characteristics of the households and extra restrictions should be imposed on pseudo-panel data before one can treat it as genuine panel data. Using the cohort data over a number of periods, one could distinguish long run and short run effects while allowing for

heterogeneity between the cohorts. In this way, one is able to overcome the deficiencies in both the static models and aggregate time series.

Since the pioneering work of Deaton (1985), the pseudo panel approach has been widely applied in microeconomics research and in many areas of social science research. In transport studies, Dargay and Vythoulkas (1999), Dargay (2001) and Dargay (2002) are notable examples of dynamic car ownership model using pseudo panel approach. One common feature of the pseudo panel studies in the literature is the assumption of linear economic relationship between parameters (although not necessarily in data). For durable goods such as cars, such assumption can at best be regarded as approximation of the aggregate demand function.

In the current study, we apply the microeconomic theory of utility maximization for individual decision makers and combine the pseudo panel model with the random utility model. The proposed random utility pseudo panel models (RUPPM) can take the functional form of Logit, Probit or other forms of discrete choice model, while the underlying data are average characteristics of cohort sample. As a result, it becomes possible to investigate the impact of dynamics and saturation within a single modelling framework. Such models can be regarded as a "third way" of analysing repeated cross sections data, presenting advantages over the linear pseudo panel models and the cross sectional discrete choice models.

This paper is organised as follows: Section One evaluates of the pros and cons of the random utility pseudo panel model (RUPPM) by comparing it to the linear pseudo panel model and cross sectional model. Section Two discusses the formulation of RUPPM and the derivation of its utility function. Section Three applies RUPPM in a hierarchical car ownership model, which is subsequently extended to include saturation. Section four discusses model estimation and presents the econometric models with the best goodness of fit. Section Five reports the results of car ownership forecasts in Great Britain to year 2021. Section Six is a brief conclusion.

## 1. Evaluation of Random Utility Pseudo Panel Models

Random utility pseudo panel model can be regarded as "third way" in choice modelling using repeated cross sectional data. It has certain advantages compared to linear pseudo panel model and static discrete choice models of cross sectional data; however, it also has its drawback and limitations. Understanding its pros and cons will ensure such model is applied in the most appropriate context.

### *1.1 Comparison with Linear Pseudo Panel Models*

In the literature of durable goods such as cars, saturation is an important concept. It is a limit on the choices faced by decision maker, which may be reached by not exceeded. In the linear pseudo panel model, saturation can only be implicitly handled by choosing appropriate functional form, e.g. in semi-log models elasticity declines with the rise of income. On the other hand,

the impact of saturation can be explicitly considered and estimated, and its statistical significance can also be examined in nonlinear pseudo panel models. By specifying car ownership models with an S-shape functional form and a saturation level, forecasts of vehicle ownership will be curtailed as saturation is approached. Although probably not being significant in developing countries, this feature would be highly significant to forecasts in more mature passenger car markets such as Great Britain (Whelan et al, 2000).

Another advantage of using random utility pseudo panel model in the study of durable goods is the consistency with the micro-economic theory. As will be shown in the following section, the utility function of car ownership can be specified as a deterministic term based on the mean sample characteristics of households in each cohort plus various random components. In this way, the model would be based on economic theory rather than aggregate empirical functions, as is the case of linear pseudo panel models.

Nonlinear pseudo panel model also has its shortcomings. Firstly, it can be problematic to consistently estimate such model with unobserved heterogeneity. Similar to the genuine panel data model of discrete choice, the fixed effect models suffer the "incidental parameter problem" while the random effect models face the difficulty of tackling "initial conditions problems" (Heckman, 1981). For pseudo panel, these problems are further complicated by the measurement errors, making it difficult to establish the consistency conditions under various asymptotics. Secondly, for empirical work, only some basic models can be estimated using commercial econometric packages, while more advanced models such as random parameter logit models (also called mixed logit models) are beyond the reach of readily available software. The current study is the first attempt to address some of these issues, although further research into various theoretical and empirical aspects of nonlinear pseudo panel models is likely to produce more fruitful results.

### *1.2 Comparison with Cross Sectional Models*

One important motivation behind the development of random utility pseudo panel is to specify dynamic models using repeated cross sectional data. The static cross sectional models rely on the assumption of equilibrium, which in practice is exception rather than norm. The disequilibrium status might be revealed by the instability of cross sections, i.e. different parameter estimates are obtained using cross sectional data in different years. If we believe that each cross sectional sample is representative of the population and long run equilibrium exists, then the cross sectional instability can be explained by the divergence of each cross section from such equilibrium. As the divergence depends on the determining factors in the current and previous periods, the degree of disequilibrium will vary between years, so will the parameter estimates (Dargay and Vythoulkas, 1999).

When the cross sectional data are not in equilibrium, it can no longer be assumed that such data capture the long run relationship, and the model based on them will produce biased estimate of long run parameters. When

these biased parameters are used for policy analysis, it can lead to wrong conclusion; when they are used in forecasts, it can lead to biased results. However, all these problems could be tackled in the pseudo panel setting, where the dynamic effects can be explicitly quantified and analyzed. In a dynamic model, it is possible to examine the significance of the lagged effects, the speed of adjustment, the extent of asymmetry[ii], etc. Long-run as well as short run elasticity, which has significant policy and practical importance, can be obtained from dynamic models. The unbiased parameters derived from the dynamic models could in theory improve the performance of the forecasting models as well.

Another potential advantage of random utility pseudo panel model relates to the choice of aggregate and disaggregate model. In many practical applications, including car ownership forecasts in the current study, the subject of interest is the aggregate statistics. Traditionally, there are two approaches to obtain aggregate measures such as market shares from data at individual level, i.e. aggregating individual data either before or after model estimation. Various classical aggregate models belong to the first approach, which are subject to various criticisms including inefficiency in the use of data, not accounting for full data variability and the risk of statistical distortion such as ecological fallacy (Ortuzar and Willumsen, 2001). Although the second approach addresses most of these criticisms, the difficult question of how to perform aggregation based on micro relations remains.

The simplest method, naïve aggregation of the discrete choice model, use average characteristics of the individual (household) to forecast the aggregate choice probability or market share. It is well known that such approach give biased results, and consequently, a number of alternative approaches have been proposed (Ben-Akiva and Lerman, 1985; Ortruza and Williamsum, 2001; Whelan, 2003). Among them, the most robust approach is sample enumeration, where the choice probability of each individual is averaged over all observations within the sample. For long term forecasts, Daly and Gunn (1985) proposed a method called prototypical sample enumeration, which involves creating an artificial sample with the same aggregate characteristics of those forecasted by planners (e.g. age and sex distribution of the populations). Another method of aggregation is classification approach, i.e. classifying the sample into homogeneous group and using group average characteristics as input to the discrete choice model. Nevertheless, this method still involves using average characteristics as input to the disaggregate model and some degree of aggregation is inevitable.

The nonlinear pseudo panel model is the “third way” in obtaining aggregate statistics from data at individual level. Individual data are aggregated into homogenous group (cohorts), and the models are estimated using average characteristics of cohort sample. As a result, the empirical model describes the economic relationship between the observed share of choices and explanatory variables at the cohort level. The probability of (cohort) decision-makers making a certain choice, when estimated based on such model, would give unbiased estimate of the market share for that choice[iii]. Moreover, the explanatory variables are cohort average characteristics that could be directly

derived from published planning statistics, thus avoiding the need for more complicated procedures such as prototypical sample enumeration at the forecasting stage. This feature would make nonlinear pseudo panel particularly attractive for long term forecasting based on cross sectional data.

Nevertheless, the aggregation of cross sectional data into cohort thus reduces the variability of the data, a similar criticism suffered by the aggregate models. Furthermore, the information on individual decision makers will be lost after aggregation, and only the average characteristics of cohort sample remain observable. The discrete choice pseudo panel model would have a composite error term, which makes the model parameters estimated on pseudo panel data not directly comparable to those estimated on individual data due to the different scale. This issue will become apparent on the discussion of the random utility model in the next section. Generally speaking, one should probably be more cautious in using nonlinear pseudo panel model as analytical tool, as whether the various disadvantages are outweighed by the inclusion of dynamics remains to be seen.

We summarise the above discussions in Table 1, highlighting the advantage and disadvantage of nonlinear pseudo panel models compared with the linear models and cross sectional models.

**Table 1 Pros and cons of random utility pseudo panel model**

| | *Vs.* Linear Pseudo Panel Model | *Vs.* Cross Sectional Model |
|---|---|---|
| Advantage | • Explicitly modelling and estimating saturation level;<br>• Consistent with theory of utility maximization; | • Consideration of dynamic in modelling;<br>• Effective tackling of aggregation bias problem; |
| Disadvantage | • Bias in the Fixed Effect Estimator;<br>• Lack of ready-made software for advanced models; | • Reduction of data variability;<br>• Loss of information on individual decision makers; |

## 2. Formulation of Random Utility Pseudo Panel Model

The random utility model makes precise distinction between the behaviour of the decision maker and the analysis of the researchers. It assumes that the decision-makers have a perfect discrimination capability; however, the researcher does not have complete information about all the elements considered by the individual making a choice. Therefore, the utility $U_{a,it}$, which individual *i* associates with alternative *a* in year *t*, can be decomposed into two parts[iv]:

$$U_{a,it} = V_{a,it} + \varepsilon_{a,it} \quad (1)$$

where $V_{a,it}$ is the deterministic and observable part, which is a function of the measured attributes; and $\varepsilon_{a,it}$ is the stochastic part, capturing the uncertainty, which reflects unobserved alternative attributes, unobserved taste variation and measurement errors made by the researcher.

In the pseudo panel setting, the deterministic utility term ($V_{a,it}$) can be further decomposed into three components, among which only the first one is observable. After the individuals are aggregated into cohorts, the researcher can only observe the average deterministic utility of all sampled individuals in cohort *c* in year *t* ($\overline{V}_{a,ct}$); on the other hand, measurement error of true mean utility for the cohort ($\eta_{a,ct}$), and the deviation from the cohort mean utility ($\theta_{a,it}$) are unobservable to the researcher. As a result, expression (1), the utility of individual *i* in cohort *c* year *t* choosing alternative *a,* can then be rewritten as:

$$U_{a,it} = \overline{V}_{a,ct} + \eta_{a,ct} + \theta_{a,it} + \varepsilon_{a,it} \qquad (2)$$

More specifically,

$\overline{V}_{a,ct} = \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} V_{a,it}, i \in c$ , which is the sample mean observable utility of alternative *a* for cohort *c* in year *t*. Note that $n_{ct}$ is the sample size of cohort *c* in year *t*;

$\eta_{a,ct} = \overline{V}^{*}_{a,ct} - \overline{V}_{a,ct}$ , representing the measurement error. It is the difference between the sample mean utility and the (unobservable) true mean utility of alternative *a* for cohort *c* in year *t* ($\overline{V}^{*}_{a,ct} = \frac{1}{N_c} \sum_{i=1}^{N_c} V_{a,it}, i \in c$)[v];

$\theta_{a,it}$ represents the unobserved utility of alternative *a* for individual *i* in year *t*, which is the deviation from the mean utility for the cohort. Ignoring measurement errors then $\theta_{a,it}$ is observable to researchers in the cross-sectional models and is "lost" in the aggregation process of pseudo panel.

The derivation of choice probability for the discrete choice model of pseudo panel will be similar to the standard random utility model. The probability of individual *i* choosing alternative *a* is equivalent to the probability that the utility of alternative *a* is higher than that of any other alternatives:

$$P_{a,it} = \Pr ob(U_{a,it} > U_{b,it}), \forall b \in A, b \neq a \qquad (3)$$

Substituting (2) into (3), we have:

$$P_{a,it} = \Pr ob(\overline{V}_{a,ct} + \eta_{a,ct} + \theta_{a,it} + \varepsilon_{a,it} > \overline{V}_{b,ct} + \eta_{b,ct} + \theta_{b,it} + \varepsilon_{b,it}), \forall b \in A, b \neq a \qquad (4)$$

In the discussion above, the utility function is developed without specific consideration of dynamics. Dynamic random utility model can be specified in different forms such as state dependence model, propensity dependence model and dynamic optimization model. In this paper, we consider a simple form of state dependence model, i.e. first order Markov model. In this case, the utility function of (1) can be written as:

$$U_{i(t),t} = V_{i(t),t} + \varepsilon_{i(t),t} = \beta' x_{i(t),t} + \alpha \cdot y_{i(t),t-1} + \varepsilon_{i(t),t} \qquad (5)$$

For repeated cross sectional data, different individuals are sampled in different years, and the notation in (5) makes it specific by adding the time

dimension to the person identifier, i.e. person $i(t)$ is different from person $i(t-1)$. Also note that the subscript identifying the alternative in (1) is dropped for simplicity.

For repeated cross section data, household's choice in the previous period, $y_{i(t),t-1}$, is unobservable; instead, we only have information of $y_{i(t-1),t-1}$. In order to investigate the choice dynamics, repeated cross sectional data have to be aggregated into pseudo panel[vi]. Assuming no birth or death in the total population and defining cohorts based on time-invariant variables, the cohort population remains fixed over time. As a result, we can write the deterministic part of the utility function in (5) as true cohort population mean plus deviation from such mean ($\theta_{i(t),t}$) for individual $i$ in year $t$:

$$\beta' x_{i(t),t} + \alpha \cdot y_{i(t),t-1} = \frac{1}{N_c}\sum_{i=1}^{N_c}(\beta' x_{it} + \alpha \cdot y_{i,t-1}) + \theta_{i(t),t}, \forall i \in c \qquad (6)$$

However, the true cohort population mean of the deterministic utility components are unobservable; instead, we only have cohort sample mean calculated from two consecutive years. Note the total measurement errors in these two periods as $(\eta_{ct} + \eta_{c,t-1})$, we have:

$$\frac{1}{N_c}\sum_{i=1}^{N_c}(\beta' x_{it} + \alpha \cdot y_{i,t-1}) = \frac{1}{n_{ct}}\sum_{i(t)=1}^{n_{ct}}(\beta' x_{i(t),t}) + \frac{1}{n_{c,t-1}}\sum_{i(t-1)=1}^{n_{c,t-1}}(\alpha \cdot y_{i(t-1),t-1}) + \eta_{ct} + \eta_{c,t-1} \qquad (7)$$

Since it is important to distinguish true state dependence and the so-called "spurious state dependence"[vii], where the dynamic effect is caused by unobserved heterogeneity, we also assume a "components of variance" structure of the error term:

$$\varepsilon_{i(t),t} = \lambda_c + \varepsilon'_{i(t),t} \qquad (8)$$

where $\lambda_c$ is the (time-invariant) unobserved heterogeneity, which includes alternative specific constants and cohort fixed (random) effects. It is assumed to be distributed independently of the residual error $\varepsilon'_{i(t),t}$;

$\varepsilon'_{i(t),t}$ accounts for the randomness besides heterogeneity, which we assume to be independently identically distributed with mean zero and variance $\sigma^2$.

Substituting (6), (7) and (8) into equation (5), and noting the cohort sample mean of the deterministic utility component as $\overline{V}_{ct}$, we have:

$$U_{i(t),t} = \overline{V}_{ct} + \eta_{ct} + \eta_{c,t-1} + \theta_{i(t),t} + \lambda_c + \varepsilon'_{i(t),t} \qquad (9)$$

where $\overline{V}_{ct} = \frac{1}{n_{ct}}\sum_{i(t)=1}^{n_{ct}}(\beta' x_{i(t),t}) + \frac{1}{n_{c,t-1}}\sum_{i(t-1)=1}^{n_{c,t-1}}(\alpha \cdot y_{i(t-1),t-1})$.

Although direct estimation of model (9) involves multiple integrals, the use of simulation methods should make such task feasible. Alternatively, further assumptions can be made to simplify the model. This is the approach taken here and will be discussed in the next section.

## 3. A Simple Model of Car Ownership

For a car ownership model, the complete choice set is the number of car owned: 0 car, 1 car… *n* Cars. Due to smaller sample size for households with 3 or more cars, we limit the choice set of our car ownership model to 0 car, 1 car and 2+ cars. In the current study, multiple car ownership is modeled based on a hierarchical structure, which involves two binary choice models: the first is the choice between zero car and one plus cars (noted as Model 1+ hereafter); then conditional on owning at least one car, choice between owning exactly one car and two plus car (noted as Model 2+|1+ hereafter). The advantage of such specification is that for each binary choice model, it does not require the IIA assumption and the assumption on the random term can be general. It also has the advantage of choice probability (of higher car ownership) increases monotonically with income. As a result, the hierarchical model structure is adopted for the current project, similar to other car ownership models such as NRTF (1997), Whelan (2001) and RAC (2002b).

Normalizing the utility of owning no car to zero ($U_0 = 0$), the utility of owning at least one car for household *i* in cohort *c* can be presented by equation (9). To turn the model into a readily tractable form, two assumptions on the random error components have been made. Firstly, under the asymptotic of $n_{ct} \to \infty, \forall t$, the measurement errors converge in probability to zero:

$$\text{plim}(\eta_{ct} + \eta_{c,t-1}) = 0 \qquad (10)$$

In another word, when the cohort sample size is sufficiently large, which is the case for the current project, the measurement errors can be ignored.

Secondly, it is assumed that both the residual error $\varepsilon'_{i(t),t}$ and deviation from the true cohort mean (deterministic) utility $\theta_{i(t),t}$ are homoskedastic and they can be aggregated into a composite error term with a certain probability distribution[viii]:

$$e_{i(t),t} = \theta_{i(t),t} + \varepsilon'_{i(t),t} \qquad (11)$$

In this case, for households with one plus cars the utility function of (9) can be simplified into the following:

$$U_{1+} = \overline{V}_{ct} + \lambda_c + e_{it} \qquad (12)$$

While all the households in cohort *c* have the same mean deterministic utility ($\overline{V}_{ct}$) and unobserved cohort heterogeneity ($\lambda_c$), they have different composite error term $e_{it}$. This reflects the essence of the Random Utility Model: given the same observed deterministic utility, decision makers behave differently due to the unobserved random error. In the current study, this is manifested in the fact that only a proportion of household in a cohort choose to own car(s). Note the household in cohort *c* owning at least one car in year *t* is noted as $y_{ct}^{1+} = 1$, then:

$$y_{ct}^{1+} = 1 \Leftrightarrow U_{1+} > U_0 \Leftrightarrow \overline{V}_{ct} + \lambda_c + e_{it} > 0 \qquad (13)$$

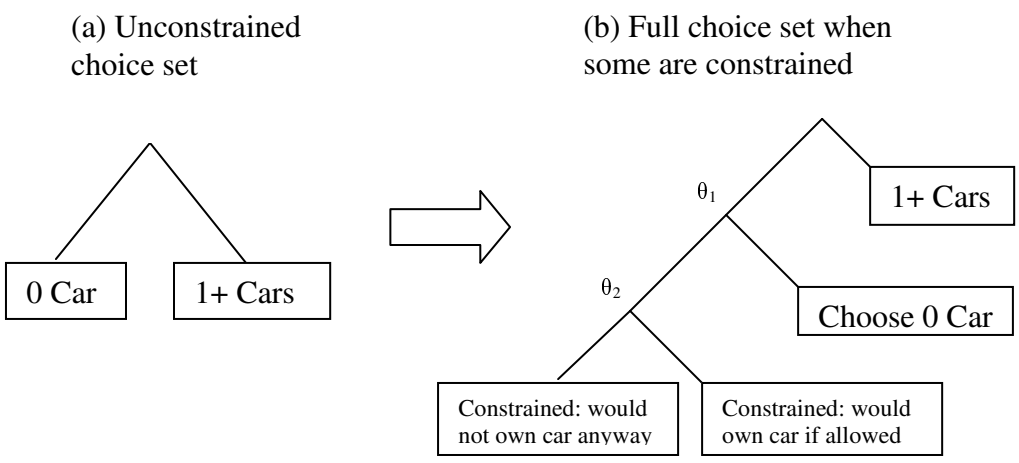
Assuming the distribution of $e_{it}$ is IID logistic, the probability of household *i* in cohort *c* owning at least one car is that of a familiar logit model:

$$P_{1+} = \Pr ob(y_{ct}^{1+} = 1 \mid \bar{x}'_{ct}, \lambda_c) = \frac{\exp(\bar{x}'_{ct}\beta + \lambda_c)}{1 + \exp(\bar{x}'_{ct}\beta + \lambda_c)} \qquad (14)$$

Model 2+|1+ would be estimated using a reduced pseudo panel dataset of car owning households, while having the identical formulation of Model 1+. The utility of owning exactly one car will be normalized to zero, and the utility of owning two or more cars will be defined in a similar fashion. The choice probability of household owning two or more car conditional on owning at least one car ($P_{2+|1+}$) is similar to (14) based on the IID logistic assumption of the composite random term.

In car ownership forecast model, saturation is an important concept, so it is important that the Random Utility Pseudo Panel Model can be formulated to account for the effect of saturation. This is achieved by specifying the RUPPM with a "Dogit" structure (Gaudry and Dagenais, 1979)[ix]. Saturation implies that some household are constrained not to own a car (captive to the alternative of zero cars), so reflected in the random utility model, the choice set faced by the decision maker[x] would have to be expanded to include new alternatives of "constrained choice". The difference in utility between the constrained and uncontained choice can then be used to infer level of saturation. However, it remains an issue how to estimate the random utility model with constrained choice. As shown in Daly (1999), by notionally separating the constrained choice set further into "voluntarily constrained" and "forcibly constrained", such model can be estimated using the conventional maximum likelihood method. The resulting model has a "tree logit" structure illustrated by Graph (b) in Figure 1.

**Figure 1 Choice Set when some decision makers are constrained not to own a car**




While the observable utilities of those choosing to own zero or one plus car remain unchanged, the observable utilities of those constrained not to own a car would include an additional linear modifier: $S^*, (S^* \in R)$. As a result, the observable utilities for the four alternatives in Graph (b) are:

$$V_{1+Car} = \beta' x$$
$$V_{Choose0} = 0$$
$$V_{forcibly\_constrained} = \beta' x + S^*$$
$$V_{voluntarily\_constrained} = S^*$$

For identification, the scale parameters of the two lower nests ($\theta_1$ and $\theta_2$) have to be constrained to 1. In this case, the above nested logit model collapses into a multinomial logit model with four alternatives, and the probability of household choosing 1+ car can be expressed as:

$$P_{1+} = \frac{\exp(\beta' x)}{1 + \exp(\beta' x) + \exp(\beta' x + S^*) + \exp(S^*)} \qquad (15)$$

As researchers are not able to empirically distinguish household that choose not to own a car and those constrained not to own a car, the probability of household not owning a car has to be considered in aggregate:

$$P_0 = \frac{1 + \exp(\beta' x + S^*) + \exp(S^*)}{1 + \exp(\beta' x) + \exp(\beta' x + S^*) + \exp(S^*)} \qquad (16)$$

In (15) and (16), $S^*$ reflects the impact of constraints on the probability of car ownership, whose sign, magnitude and statistical significance remained to be estimated. When the impact of such constraints on car ownership is negligible, then $S^* \to -\infty$, and the probability of household owning zero car is the same as when no constraints exists: $\lim_{S^* \to -\infty} P_0 = \frac{1}{1 + \exp(\beta' x)}$. On the other hand, when the impact of such constraints on car ownership is extremely large, we have $S^* \to \infty$, and the probability of household owning zero car is close to one: $\lim_{S^* \to \infty} P_0 = 1$.

Model (16) is a special case of the Dogit Model proposed in Gaudry and Dagenais (1979)[xi]. On the other hand, it is easy to show that the probability model of (15) is mathematically equivalent to the empirical probability function commonly assumed for logit model with saturation:

$$P_{1+} = \frac{S \cdot \exp(x'\beta)}{1 + \exp(x'\beta)} \qquad (17)$$

The equivalence is established by rewriting equation (17) as:

$$P_{1+} = \frac{\exp(x'\beta)}{[1 + \exp(x'\beta)] \cdot (1 + \frac{1-S}{S})} = \frac{\exp(x'\beta)}{[1 + \exp(x'\beta)] \cdot [1 + \exp(\ln\frac{1-S}{S})]}$$

and by noting $\ln\frac{1-S}{S} = S^*$. Instead of directly estimating the non-linear term *S* in (17), we now estimate a linear term $S^*$ in the exponential function in (15). As $S = \frac{1}{1 + \exp(S^*)}$, it would satisfy $0 < S < 1$, representing the probability limit that can never be exceeded.

## 4. Estimating RUPPM of Car Ownership

The consistent estimation of the Random Utility Pseudo Panel model is influenced by the specification of the unobserved heterogeneity $\lambda_c$. The most common specification is the "fixed effect" model, where $\lambda_c$ are represented by cohort dummies. It should be noted that the fixed effect estimator of discrete choice model is only consistent when the number of time period *T* is large[xii]. Alternatively, one can assume $\lambda_c$ follow a certain distribution, which leads to the "random effect" specification. However, the consistent estimation of the random effect model relies on the orthogonality assumption between the unobserved effects and the explanatory variables, which would be difficult to defend with the presence of lagged dependant variable.

In the current study, we rely on a rather optimistic result of the fixed effect estimator, i.e. consistency under the asymptotic of large *T*. This action is partially justifiable on the grounds that we have relatively long sample periods of up to 19 years. Alternatively, we adopt the more flexible random parameter specification. Unlike the random effect model, where a constant term is assumed to capture the unobserved heterogeneity, the unobserved effects are captured by a randomly distributed parameter vector in a random parameter model. As a result, the orthogonality condition of the random effect model becomes moot, as the individual specific heterogeneity is embodied in the marginal responses (parameters) of the model (Greene, 2001).

Both the fixed effect models and the random parameter models are estimated using a Gauss routine adapted from the code of Revelt and Train (1998), which uses the method of maximum simulated likelihood. The original code is modified so that the dependent variable is the proportion of the decision makers choosing each option, and the choice probability is weighted by the cohort sample size. For car ownership model with saturation, the code is further modified to include the constrained choice[xiii].

The empirical model is estimated using the pseudo panel dataset constructed from the UK Family Expenditure Survey (FES) data. The main dataset contains 254 observations, covering 16 cohorts for the period of 1982-2000. The indirect utility function includes the explanatory variables commonly found in the literature of car ownership models. Table 2 presents the key explanatory variables included in the current study.

**Table 2 Explanatory Variables in car ownership model**

| Type | Name | Description |
|---|---|---|
| Income | LnInc | Log of real average household disposable income |
| Household Demographic Characteristics | Adult | Average number of adults per household |
| | Child | Average number of children per household |
| | Worker | Average number of person in work per household |
| | HH1 | % of HH as type 1 (One adult, in work ) |
| | HH2 | % of HH as type 2 (One adult, not in work) |

| | | |
|---|---|---|
| | HH3 | % of HH as type 3 (One adult, with children) |
| | HH4 | % of HH as type 4 (Two adults, neither in work) |
| | HH5 | % of HH as type 5 (Two adults, no children) |
| | HH6 | % of HH as type 6 (Two adults, with children) |
| | HH7 | % of HH as type 7 (Three + adults, no children) |
| | HH8 | % of HH as type 8 (Three + adults, with children) |
| | Age | Age of household head |
| | AgSq | Square of age of household head divided by 100 |
| Household Locations | Area1 | % of household living in Greater London Area |
| | Area2 | %of HH living in metropolitan districts and central Clydeside conurbation |
| | Area3 | % of HH living in areas with a population density of 7.9 or more persons per hectare |
| | Area4 | % of HH living in areas with a population density of 2.2 to 7.9 persons per hectare |
| | Area5 | % of HH living in areas with a population density of less than 2.2 persons per hectare |
| | MET | Combining Area1 and Area2 |
| | RURAL | Equivalent to Area5 |
| Motoring Costs | LnPrice | Log of real car purchase price index |
| | LnRunCst | Log of read car running cost index |

A number of models have been estimated and they are selected based on various criteria including the log likelihood, sign and statistical significance of the parameters and implied elasticity. For household with **one or more cars**, the model with the best fit is a fixed effect model, where the household characteristics are represented by the eight-way categorization of household types. The log likelihood of the fixed effect model is -65859, and the adjusted likelihood ratio index is 0.1889. Table 3 reports the results of the fixed effect model for household owning at least one car, which will be used for forecasting in the next chapter. Both the slope coefficient and marginal effects are shown, with the latter estimated at the weighted average of the explanatory variables.

**Table 3 Model of 1+ cars (t-statistic in parentheses)**

| | Slope Coefficient | | Marginal Effect | |
|---|---|---|---|---|
| LagY | 1.0786 | (4.45) | 0.2115 | *** |
| LnInc | 0.3945 | (2.55) | 0.0774 | ** |
| HH2 | -1.3671 | (-1.87) | -0.2681 | * |
| HH3 | -2.0141 | (-1.90) | -0.3950 | * |
| HH4 | 1.4763 | (2.46) | 0.2895 | ** |
| HH5 | 1.0815 | (1.44) | 0.2121 | ' |
| HH6 | 1.8374 | (2.77) | 0.3603 | *** |
| HH7 | 1.2892 | (1.58) | 0.2528 | ' |
| HH8 | 1.6144 | (1.83) | 0.3166 | * |
| Met | -0.9489 | (-2.25) | -0.1861 | ** |
| Rural | 0.6772 | (1.57) | 0.1328 | ' |

| | | | | |
|---|---|---|---|---|
| LnPrice | -0.5800 | (-3.46) | -0.1137 | *** |
| LnRunCst | -0.8184 | (-4.30) | -0.1605 | *** |
| Age | 0.0919 | (4.01) | 0.0180 | *** |
| AgSq | -0.0583 | (-3.16) | -0.0114 | *** |
| S* | -2.4582 | (-6.42) | - | - |
| C2 | 0.1040 | (0.92) | 0.0204 | ' |
| C3 | 0.1679 | (1.45) | 0.0329 | ' |
| C4 | 0.3347 | (2.44) | 0.0656 | ** |
| C5 | 0.4439 | (2.82) | 0.0870 | *** |
| C6 | 0.5997 | (3.24) | 0.1176 | *** |
| C7 | 0.8187 | (3.72) | 0.1605 | *** |
| C8 | 0.9730 | (3.85) | 0.1908 | *** |
| C9 | 1.1991 | (3.87) | 0.2351 | *** |
| C10 | 1.4448 | (3.90) | 0.2833 | *** |
| C11 | 1.5371 | (3.86) | 0.3014 | *** |
| C12 | 1.7391 | (3.94) | 0.3410 | *** |
| C13 | 1.9663 | (4.04) | 0.3856 | *** |
| C14 | 2.1536 | (4.10) | 0.4223 | *** |
| C15 | 2.2183 | (3.91) | 0.4350 | *** |
| C16 | 2.3610 | (3.73) | 0.4630 | *** |
| Log Like'd | -65859 | | | |
| Null LL | -81240 | | | |
| Adj. LRI | 0.1889 | | | |

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

The coefficient of the lagged dependent variable is significant and positive, indicating state dependence in cohort car ownership levels. The income and price coefficients are all significant and with expected sign. Table 4 reports the income and price elasticity for cohorts at various income levels. The coefficients for the average age of household head and age square are positive and negative respectively, indicating a peak of car ownership during the household life cycle. The coefficient for the linear modifier $S^*$ implies saturation level of 0.9212, which appears sensible for model of one plus cars.

**Table 4 Income and cost elasticity of Car 1+**

| | Short Run | | | Long Run | | |
|---|---|---|---|---|---|---|
| **Income** | Income Elasticity | Price Elasticity | Running Cost Elasticity | Income Elasticity | Price Elasticity | Running Cost Elasticity |
| Low | 0.198 | -0.291 | -0.411 | 0.238 | -0.351 | -0.495 |
| Middle | 0.082 | -0.121 | -0.171 | 0.111 | -0.164 | -0.231 |
| High | 0.065 | -0.096 | -0.136 | 0.069 | -0.102 | -0.144 |

Under the hierarchical structure, the models of **two or more cars** are specified to be conditional on household owning the first car. As a result, the empirical models are estimated on a second pseudo panel dataset that was constructed from the sub-sample of car owning households. Specification search shows that fixed effect models do not have better goodness of fit, and

in the mixed logit model the standard deviations of the random parameters were not significantly different from zero. Consequently, all parameters are treated as fixed in the final model.

While using the five-area household location split improves model fit, there is no significant loss of fit when using average demographic statistics rather than eight-way household type split if degree of freedom is taken into account. The average household size variable is not significant and subsequently dropped, so the household characteristics are described by the average number of children and people in work per household. This leads to the model of best fit reported in Table 5.

**Table 5 Model of Car 2+|1+ (t-stat in parenthesis)**

| | Slope Coeff | | Marginal Effect | |
|---|---|---|---|---|
| ONE | -10.4365 | (-3.08) | -2.1587 | *** |
| LagY | 2.3361 | (5.23) | 0.4832 | *** |
| LnInc | 1.0649 | (4.63) | 0.2203 | *** |
| Child | -0.1362 | (-3.67) | -0.0282 | *** |
| Worker | 0.1844 | (2.26) | 0.0381 | ** |
| AREA2 | 2.2701 | (2.77) | 0.4695 | *** |
| AREA3 | 1.1823 | (1.45) | 0.2446 | ' |
| AREA4 | 1.6324 | (2.11) | 0.3376 | ** |
| AREA5 | 1.0771 | (1.44) | 0.2228 | ' |
| LnPrice | -0.6078 | (-1.88) | -0.1257 | * |
| LnRunCst | 0.6191 | (2.42) | 0.1280 | ** |
| Age | 0.0769 | (5.15) | 0.0159 | *** |
| AgSq | -0.0840 | (-5.08) | -0.0174 | *** |
| S* | -0.7891 | (-3.07) | - | |
| Log Like'd | -47147 | | | |
| Null LL | -56288 | | | |
| Adj. LRI | 0.1621 | | | |

***: Significant at 1% level;
**: Significant at 5% level;
*: Significant at 10% level;
': Not statistically significant

Examining the estimation results in Table 5, there are two parameters with unexpected sign. The coefficient of average number of children in the household is negative and significant, which might be due to the correlation between that variable and the average number of people in work. Also, the coefficient for the log of real running costs is positive and significant, which might be caused by the concurrent substantial rise of car running costs and ownership of two plus cars in the second half of 1990s. In terms of household location, if the proportions of households living in metropolitan and rural areas (Area type 2 to Area type 4) increase at the expense of that in Greater London (the base case of Area type 1), the conditional probability of household owning two or more cars would also increase.

The estimated linear utility modifier $S^*$ is -0.7891, implying a saturation level of 0.6876. We have also calculated the short run and long run income and costs elasticity for cohorts with low, median and high income level, which is reported

in Table 6. The income and purchase price elasticity are higher than those for models of one plus car, which is as expected. On the other hand, the running cost elasticity is shown in italic due to its unexpected sign.

**Table 6 Income and cost elasticity of Car 2+|1+**

| | Short Run | | | Long Run | | |
|---|---|---|---|---|---|---|
| **Income** | Income Elasticity | Price Elasticity | Running Cost Elasticity | Income Elasticity | Price Elasticity | Running Cost Elasticity |
| Low | 0.95 | -0.54 | *0.55* | 1.23 | -0.70 | *0.72* |
| Middle | 0.78 | -0.45 | *0.45* | 0.94 | -0.53 | *0.54* |
| High | 0.62 | -0.35 | *0.36* | 0.68 | -0.39 | *0.39* |

## 5. Car Ownership Forecasts

Several empirical models are selected for car ownership forecasts. In the current study, the geographic area covered is limited to Great Britain (as opposed to the United Kingdom) to be consistent with the National Road Traffic Forecasts (NRTF) and the National Transport Model maintained by the UK Department for Transport. The forecasting period is between 2001 and 2021, since more detailed household projection data are only available up to year 2021.

As all the empirical models are estimated using pseudo panel data, which are average statistics of cohort sample, it is easier to obtain aggregate measures such as total car stock compared to cross sectional models. Unlike the latter, it is not necessary to use the more complicated techniques such as prototypical sample enumeration (Daly and Gunn, 1985; Whelan, 2003; Whelan, 2007). However, it is still a challenging task to derive the cohort based household characteristics in future years using the available planning data. More specifically, it is important to separate the age effects and time trend effects (similar to the 'life cycle effects' and 'generation effects' in Dargay and Vythoulkas, 1999) on income and other characteristics over the life cycle.

It is also necessary to establish the age profile of the existing and new cohorts over the forecasting period. We decide to drop data points when household head is aged over 100, which also leads to the exclusion of the oldest cohort in the dataset (born between 1901 and 1906) from the forecasting model. On the other hand, five new cohorts have been introduced over the period, with the youngest born between 2001 and 2006. As a result, the model involves 20 cohorts in total over the forecasting period.

### *5.1 Forecast Assumptions*

For each of the twenty cohorts in the period between 2001 and 2021, one has to make projections of explanatory variables in the econometric models and two other variables: number of households and 'multiple-car factor' for households with two or more cars. The explanatory variables include

household real disposable income, average household demographic statistics, split of households between the eight household types, split of households between location types and aggregate real purchase price and car running costs index. The input data in 2000 (Year 0) are estimated based on Family Expenditure Survey data or backcast of 2001 census data. The future year growth assumptions are derived from social economic forecasts published by various sources.

Regarding the number of households in the 20 cohorts, we make use of the census product from Office of National Statistics, "Focus on Family" (ONS, 2005), which contains data on the number of families based on the 14 age bands of family reference person in 2001. By further taking into account the number of one person household in different age groups, it is possible to derive the number of household for all cohorts in 2001. ODPM (1999)[xiv] and Scottish Executive (2002) provide projections of household in England and Scotland and are used to derive the growth rates of household number by different age bands of household representatives.

Regarding household real disposable income, the base year (Year 2000) data is obtained from Family Expenditure Survey. The income growth is assumed to be in line with the growth of Gross Domestic Product (GDP), which however has to be adjusted downward to account for the increasing number of household in each cohort. We use the observed real GDP growth between 2001 and 2006, which is obtained from Treasury Weekly Economic Indicators Databank (Treasury, 2007). From 2007 onwards, the GDP is assumed to grow at 2.25% per annum, the same rate used in Department for Transport's National Road Traffic Forecasts (1997), National Transport Model (Whelan, 2003; 2007) and 10 year plan (DETR, 2000).

The base year estimates of household size and average number of children and working person per household by cohorts are obtained from Family Expenditure Survey. The publication by Government Actuary's Department on projected populations by age (GAD, 2003) is used to calculate the growth rate of household size and average number of children per household (taken into account the change of household numbers). Regarding the average number of working persons per household, we assume a constant labour market participation rate of 74.6% of the adult population (Treasury, 2007), so the workforce growth is entirely driven by population change.

It is not possible to project the change of location split by cohorts. As a result, the base year location split estimated from Family Expenditure Survey is assumed to be unchanged over the forecasting period. Regarding the real car ownership costs index, we assume the car purchase price falls by 0.37% per annum and the car running costs remain constant. These assumptions are also consistent with those in the National Transport Model and the 10 year plan.

### *5.2 Generating Projections of Input Variables*

It is well known that for pseudo panel data, household characteristics such as income go through a "hump" shape life cycle peaking at the age of late 40s; furthermore, at a given age, households in younger cohorts tend to have higher income than those in older cohorts. To derive sensible projection of the input variables, one should separate the age effect and time trend effect. The current study develops a sub-model of input projection, which includes 81 overlapping age bands and explicitly separates these two effects. This sub-model is implemented in three steps:

1. Estimating the base year figures for the relevant variables for 81 overlapping age bands of household head, e.g. those aged 15-19, 16-20, 17-21…94-98, 95-99. The data sources include census and Family Expenditure Survey, and because the original data are for non-overlapping age groups (15-19, 20-24, 25-29…), method of interpolation is used to obtain estimates for all 81 age bands. This stage isolates the age effect cross cohorts.
2. For each of the 81 age band, forecast the future year figures based on standard growth assumption described in the previous sub-section. Different growth rates are applied to different cohorts whenever it is possible. This stage introduces the time trend effect.
3. The first two steps have produced a matrix of 21 rows by 81 columns (21 years for 81 age bands) for each input variable. Within each matrix, identify the twenty cohorts by the age of the household head. For example, in 2001, age band 16-20 corresponds to cohort whose head is born between 1981 and 1985 (Cohort ID F5); age band 21-25 is cohort born between 1976 and 1980 (ID F6). In 2002, it is age band 17-21 that refers to cohort F5 and age band 22-26 refers to cohort F6. Similarly, age band 36-40 refers to cohort F5 and age band 41-45 refers to cohort F6 in 2021. Extract the appropriate cells for each of the 20 cohorts from the 21x81 matrix and arrange them by cohort and year, we obtain the projection of input variables that can be used in the car ownership forecasting models.

The above method is used to generate projection for most of the input variables. However, an alternative approach has to be adopted regarding the split of eight household types, because it is not possible to obtain the appropriate growth rate required in the second step of the projection model and it is not satisfactory to assume that there is no change of household type split within cohort over time. The alternative approach involves assigning the observations in the original pseudo panel dataset into a 69 by 20 matrix. The 69 rows cover cohorts aged between 19 and 87, and the 20 columns correspond to the 20 cohorts. At a certain age (in one particular row), there are a number of pseudo panel observations belonging to different cohorts, which gives the growth rate between generations (younger and older cohorts). To dampen down noise, we actually use the average growth rates of cohorts with similar age (within 5 years difference) to project the future year values. These future year figures are contained in different cells of the 69x20 matrix and have to be extracted and re-arranged by cohorts and years. A final

adjustment is made to ensure that the proportions of the eight household types sum to 100%.

### *5.3 Forecasting and Result Evaluation*

After projecting the future values of the explanatory variables in a sub-model, we are ready to apply the econometric models to generate car ownership forecasts. In order to identify the impacts of model specification on the forecasting results, the following econometric models are used: linear pseudo panel model, static RUPPM, dynamic RUPPM, and dynamic RUPPM with saturation.

For linear model, the dependent variable is the average number of cars per household, so the total car stock can be easily obtained by multiplying the fitted dependent variable by the household numbers in each cohort and summing over all cohorts. For nonlinear model, we estimate the probability of household owning at least one car ($P_{1+}$) and owning two or more cars conditional on owning the first one ($P_{2+|1+}$). The unconditional probability of household owning two or more cars are the product of $P_{2+|1+}$* $P_{1+}$. When the discrete choice model is a multinomial logit model with a constant term, first order condition ensures that these probabilities are the unbiased estimates of proportions of households owning certain number of cars. It thus follows that the proportion of household owning *exactly* one car is ($P_{1+}$ - $P_{2+|1+}$* $P_{1+}$) and the proportion of household owning two plus cars is $P_{2+|1+}$* $P_{1+}$.

For those with two or more cars, one has to estimate the average number of cars in the household, or the so-called 'multiple-car factor' (noted as $F$, $F \geq 2$). The base year values of multiple-car factors ($F_{c0}$) are derived using the Family Expenditure Survey data. The long term growth rate of $F$ is assumed to be 0.10% per annum, calculated using FES data over a 10 year period. Table 7 shows the assumed average number of cars in multiple-car household for six age bands in five years[xv].

**Table 7 Multiple-car factor used in forecasting**

| | 16-19 | 20-24 | 25-44 | 45-64 | 65-74 | 75+ |
|---|---|---|---|---|---|---|
| 2001 | 2.002 | 2.022 | 2.156 | 2.298 | 2.065 | 2.002 |
| 2006 | 2.013 | 2.033 | 2.167 | 2.310 | 2.076 | 2.013 |
| 2011 | 2.023 | 2.043 | 2.178 | 2.322 | 2.086 | 2.023 |
| 2016 | 2.034 | 2.054 | 2.190 | 2.335 | 2.097 | 2.034 |
| 2021 | 2.044 | 2.065 | 2.201 | 2.347 | 2.108 | 2.044 |

For every year, the total number of cars is then calculated by multiplying the total number of households by the proportions of car owning households for each cohort, and summing over all cohorts:

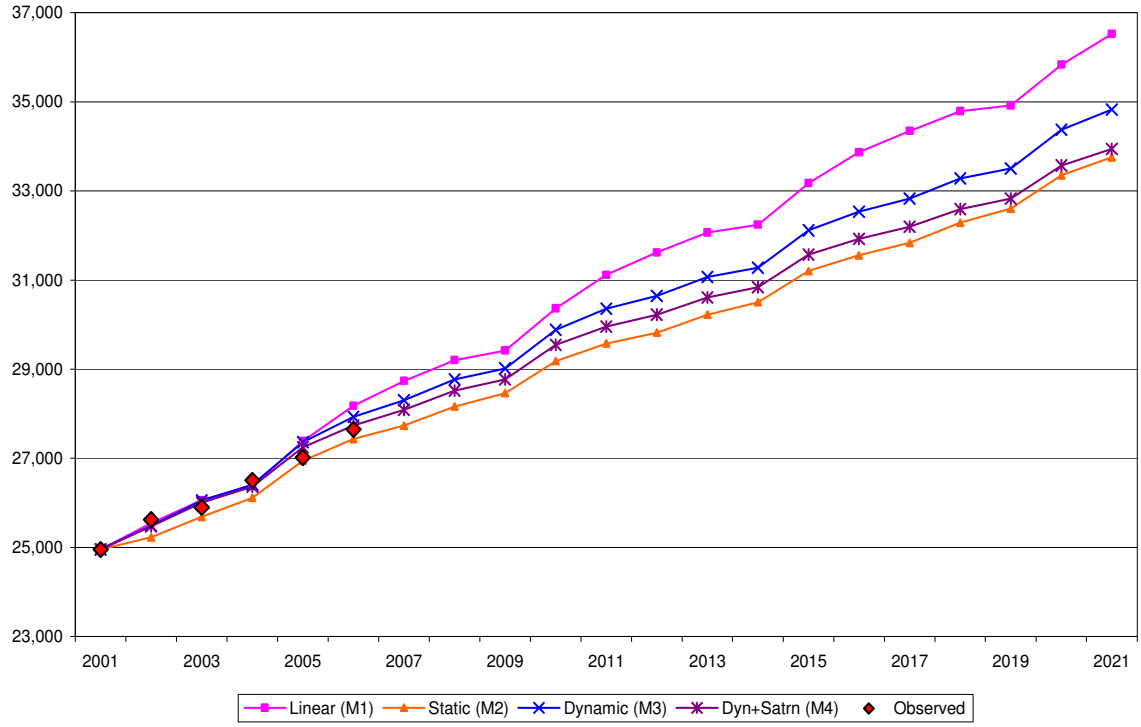
$$TA_t = \sum_c HH_{ct}[(P_{1+}^{ct} - P_{2+|1+}^{ct} \cdot P_{1+}^{ct}) + (P_{2+|1+}^{ct} \cdot P_{1+}^{ct}) \cdot F_{ct}]$$

$$= \sum_c [HH_{ct} \cdot P_{1+}^{ct} + HH_{ct} \cdot P_{2+|1+}^{ct} \cdot P_{1+}^{ct} \cdot (F_{ct} - 1)]$$

where, $TA_t$ = Total number of cars in year $t$;
$HH_{ct}$ = Total number of household for cohort $c$ in year $t$;
$F_{ct}$ = Average number of cars in multiple car household for cohort $c$ in year $t$.

The car ownership forecasts based on the four sets of econometric models are validated against the observed car stock in Great Britain in 2001. The forecasting period is up to year 2021, and between 2002 and 2006, the results can also be compared to the observed total. The observed stock are calculated from Transport Statistics Bulletin 'Vehicle Licensing Statistics' (DfT, 2005; 2006a) and 'Vehicle Exercise Duty Evasion' (DfT, 2006b), with the latter providing estimates of unlicensed car stock. The total number of cars includes private cars (whether owned by individuals or companies) but excludes "non-cars private light goods vehicle". Figure 2 shows the four sets of forecasting results based on linear dynamic pseudo panel model (M1), static RUPPM (M2), dynamic RUPPM (M3) and dynamic RUPPM with Saturation (M4). It also shows the observed car stock from 2001 to 2006, which are represented by "red diamond" points.

**Figure 2 Observed Total Car Stocks and Forecasts from 4 Models (000s)**




There are several points worth noting regarding the results reported in Figure 2. Firstly, the forecasts based on the dynamic RUPPM with saturation (M4, whose coefficients are reported in Table 3 and Table 5) match the observed car stock most closely. The forecast car ownership level based on M4 is 27.74 million in 2006, which is only 0.3% higher than the observed value. Secondly, ignoring saturation effects would lead to higher forecasts, as the result based

on M3 (dynamic RUPPM, saturation not modeled) is higher than that of M4 by 2.6% in 2021. Thirdly, ignoring dynamic effects would lead to lower forecasts, as the results based on the static model (M2) is 3.1% lower than that based on the dynamic model (M3). Fourthly, the forecasts based on the linear pseudo dynamic panel models (M1) appear to be too high: the 2006 forecast is 1.9% higher than the observed figure and the 2021 forecast is 7.6% higher that that based on M4.

The performances of the four forecasting models are also evaluated by comparing the results to other published studies, as no observed car ownership data are available beyond year 2006. The other studies used for comparison include National Road Traffic Forecasts (NRTF, 1997), car ownership model supporting the influential RAC report "Motoring Towards 2050" (RAC 2002a; 2002b), and car ownership sub-model in the UK Department for Transport's National Transport Model (Whelan, 2003 and Whelan, 2007). Table 8 compares the four sets of forecasts in the current studies with the above sources.

**Table 8 Forecasts Comparison: current studies *vs.* published studies (millions)**

| Year | Linear (M1) | Static (M2) | Dynamic (M3) | Dyn+Satrn (M4) | NRTF (1997) | RAC (2002b) | Whelan (2003)* | Whelan (2007)* |
|---|---|---|---|---|---|---|---|---|
| 2001 | 24.95 | 24.95 | 24.95 | 24.95 | 25.18 | 25.18 | 28.12 | 25.63 |
| 2006 | 28.18 | 27.43 | 27.93 | 27.74 | n.a. | n.a. | 30.28 | 28.59 |
| 2011 | 31.12 | 29.58 | 30.36 | 29.95 | 28.88 | 28.88 | 32.66 | 30.84 |
| 2016 | 33.87 | 31.56 | 32.54 | 31.92 | n.a. | n.a. | 34.48 | 32.71 |
| 2021 | 36.52 | 33.75 | 34.82 | 33.94 | 31.77 | 32.26 | 36.08 | 34.26 |

* National Transport Model

The forecasts in the early NRTF (1997) are the lowest, and all other studies predict higher car numbers in 2021. The early National Transport Model forecasts (Whelan, 2003) appear to be too high and have been subsequently revised down in Whelan (2007). In the current study, the forecasts based on RUPPM (M2 to M4) are broadly similar to the latter, which are the latest "official" figures. On the other hand, the forecasts based on the linear model (M1) are substantially higher than other studies (except Whelan, 2003).

Overall, the forecasts based on M4 (Dynamic RUPPM with saturation) appear most satisfactory and can be regarded as central estimate. While the forecasts based on the Static model of M2 are quite similar to those based on M4, it is because the two effects that are ignored in M2 (dynamics and saturation) have opposite effects on car ownership. Nevertheless, the result differences from the three sets of RUPPM are not very substantial. The forecasts based on the linear pseudo panel model are much higher. As the linear models can not explicitly account for saturation (although the semi-log *X* specification can be regarded as a form of approximation), it is likely that this would lead to over-prediction of car ownership level in the long term. As RUPPM has sigmoid probability functions, the effects of ignoring saturation

would be less significant, so this could be another advantage of using RUPPM rather than linear models in car ownership forecasts.

## 6. Conclusion

In this paper, we present a Random Utility Pseudo Panel Model (RUPPM) and argue for its potential as an effective "third way" in modelling and forecasting using repeated cross section data. More specifically, it has the distinctive advantages of allowing both dynamics and saturation without the need for expensive genuine panel data. However, some valuable information on individual decision makers would be lost during cohort aggregation. On balance, it appears that nonlinear pseudo panel model is most suitable for forecasting purpose, while the case is less clear for analytical purpose.

Under the framework of random utility model, it is shown that the utility function of the pseudo panel model is a direct transformation from that of cross-sectional model and both share similar probability model albeit with different scale. In a standard random utility model of cross sectional data, the utility function consists of a deterministic term and a random term. For pseudo panel model, the deterministic term can be further decomposed into three components including: sample mean observable utility, measurement error and individual decision maker's utility deviation from the cohort mean.

A simple RUPPM has been applied in car ownership modeling. The model has a hierarchical structure with 1+ cars and 2+|1+ cars treated separately. Three key assumptions have been made to make the model readily tractable. Firstly, it is assumed that the random utility term for individual households has a "components of variance" structure, which is the sum of cohort specific component representing unobserved heterogeneity and a temporally independently identically distributed (IID) residual error component. Secondly, it is assumed that the cohort samples are homoskedatic, so the component that represents utility deviation from cohort mean can be combined with the IID residual error component and the resulting composite error term has certain probability distribution. Thirdly, measurement errors can be ignored as they converge in probability to zero under the asymptotic of infinitive $n_{ct}$ (the sample size per cohort is sufficiently large in each year). These assumptions lead to models that have a similar probability function but with different scale compared to the cross-sectional model. Finally, due to the importance of saturation in car ownership model, the simple RUPPM has been formulated in a "Dogit" form so the saturation level and its statistical significance can be reliably estimated.

The econometric models have been used to forecast car ownership in Great Britain to 2021. The results based on linear pseudo panel models and RUPPM with different specifications are reported. The forecasts based on dynamic RUPPM with saturation closely match the observed car stock between 2001 and 2006 and can be regarded as the central estimates. Ignoring saturation would lead to higher forecasts and ignoring dynamic effect would lead to lower forecasts. Linear pseudo panel model, on the other hand, produces car ownership forecasts that are substantially higher than those

based on RUPPM and in other published studies. These results indicate the importance of selecting the most appropriate functional form and considering both dynamics and saturation in car ownership forecasts. Given the solid theoretical foundation and the satisfactory empirical results, it would be appropriate to recommend the use of RUPPM for car ownership forecasts in countries where long running repeated cross sectional survey data are available.

**Reference:**


Ben-Akiva, M.E. and Lerman, S.R. (1985), *Discrete Choice Analysis*, Cambridge, Massachusetts: MIT Press

Dargay, J. (2001), The Effect of Income on Car Ownership: Evidence of Asymmetry, *Transportation Research Part A*, Volume 35, pp807-821

Dargay, J. (2002), Determinants of car ownership in rural and urban areas: a pseudo-panel analysis, *Transportation Research Part E: Logistics and Transportation Review* Volume 38, Issue 5, September 2002, pp351-366

Dargay, J. and Vythoulkas, P. (1999), Estimation of a Dynamic Car Ownership Model, A Pseudo-Panel Approach, *Journal of Transport Economics and Policy*, Vol. 33, Part 3, Sept. 1999, pp 287-302

Daly, A. (1999), *How Much is Enough? Saturation Effects Using Choice Models*, Traffic Engineering and Control, Oct. 1999, pp 493-495

Daly, A.J. and Gunn, H.F. (1985), *Cost-Effective Methods for National-Level demand Forecastin*', IATBR Conference, Noordwijk

Deaton, A. (1985), "Panel Data from Time Series of Cross Sections", *Journal of Econometrics*, 30, pp109-26

DETR (2000), *Transport 10 Year Plan 2000,* http://www.dft.gov.uk/about/strategy/whitepapers/previous/transporttenyearplan2000

Department for Transport (2005), *Vehicle Licensing Statistics*, 2005, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/licensing/vehiclelicensingstatistics2005b

Department for Transport (2006a), *Vehicle Licensing Statistics Release*, 2006, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/licensing/vehiclelicensingstatrelease2006

Department for Transport (2006b), *Transport Statistics Bulletin, Vehicle Excise Duty Evasion*, 2006, http://www.dft.gov.uk/pgr/statistics/datatablespublications/vehicles/excisedutyevasion/vehicleexcisedutyevasion2006

Gaudry, M. and Dagenais, M. (1979), The Dogit Model, *Transportation Research* Part B, 13 (2), pp105-112

Government Actuary Office (2003), *Population Projections*, http://www.gad.gov.uk/Population/2003/gb/wgb035y.xls

Greene, W. (2001), *Fixed and Random Effects in Nonlinear Models*, Working paper, Department of Economics, New York University

Heckman, J. (1981a), Statistical Models for Discrete Panel Data, in *Structural Analysis of Discrete Data with Econometric Applications*, Manski, C. and McFadden, D. (eds.), Cambridge: MIT Press

Huang, B. (2007), *The Use of Pseudo Panel Data for Forecasting Car Ownership*, unpublished PhD thesis, Department of Economics, Birkbeck College

Moffitt, R. (1993), Identification and Estimation of Dynamic Models with Time Series of Repeated Cross Sections, *Journal of Econometrics* 59, pp 99-123

NRTF (1997), *National Road Traffic Forecasts (Great Britain) 1997, Working Paper No. 1, Car Ownership: Modelling and Forecasting,* Department of the Environment, Transport and the Regions

Office of Deputy Prime Minister (1999), *Projections of households in England 2021*, http://www.odpm.gov.uk/stellent/groups/odpm_housing/documents/page/odpm_house_604206.hcsp

Office of National Statistics (2005), *Focus on Family*, http://www.statistics.gov.uk/focuson/families/

Ortuzar, J. and Willumsen, L. (2001), *Modelling Transport*, 3rd Edition, London: John Wiley & Sons, Ltd

RAC (2002a), *Motoring towards 2050*, RAC Foundation, http://www.racfoundation.org/files/rac_foundation_2050.pdf

RAC (2002b), *Motoring towards 2050, Appendix 2: Technical Details of Forecasting Procedure*, RAC Foundation

Revelt, D. and Train, K. (1998), Mixed Logit with Repeated Choices, *Review of Economics and Statistics*, 80, pp647-657

Scottish Executive (2002), *Household Projections for Scotland: 2000-Based*, http://www.scotland.gov.uk/stats/bulletins/00179-00.asp

Treasury (2007), *HM Treasury Weekly Economic Indicator Data Bank*, 1st May 2007, http://www.hm-treasury.gov.uk/economic_data_and_tools/latest_economic_indicators/data_indic_index.cfm

Whelan, G. (2003), *Modelling Car Ownership in Great Britain*, unpublished PhD thesis, University of Leeds

Whelan, G., Wardman, M. and Daly, A. (2000), *Is There a Limit to Car Ownership Growth? An Exploration of Household Saturation Levels Using two Novel Approaches*, paper presented to European Transport Conference, PTRC, Cambridge

Whelan, G. (2007), Modelling car ownership in Great Britain, *Transportation Research* Part A 41 pp 205–219

[i] I would like to thanks Prof. Ron Smith and two PhD examiners for their constructive comments on my thesis. A version of this paper was presented to the 2007 European Transport Conference with the help of Dr. Gerard Whelan. Any errors and omissions remain my own.

[ii] A notable example on car ownership is Dargay (2001).

[iii] For multinomial logit model whose utility function includes a constant term, this result directly follows the first order condition of the log likelihood function.

[iv] We directly start from a panel data model and introduce a time dimension accordingly.

[v] Note that while cohort sample changes year by year, the cohort population remains fixed over time if cohorts are defined based on time-invariant variables and we assume total population is close, i.e. there is no birth or death and cohort size $N_c$ remains constant over time.

[vi] Another possibility is to use the first order Markov models proposed in Moffitt (1993), in particular the linear probability model for hazards. However, the data requirement for such model is very high, as it requires the previous values of the explanatory variables, or at minimum, the accurate backcast of such variables.

[vii] As pointed out by Heckman (1981a), the lagged dependent variable in the dynamic model might appear significant even if there is no true state dependence. In another word, inter-temporal correlation of the error term has to be accounted for before true state dependence can be revealed.

[viii] It should be noted that this assumption could be relaxed without rendering the model intractable. Based on cohort sample, we can estimate the standard deviation of $\theta_{i(t),t}$, which should in turn enable the model to be estimated using simulation method.

[ix] It is called Dogit Model because it dodges (avoids) the researcher's dilemma of choosing *a priori* between a format which commits to IIA restrictions and one which excludes them.

[x] The decision maker is an individual household in the micro survey. The probability model presented here refers to individual decision makers, and the corresponding pseudo panel model should be developed similarly as in Section 2. This is not shown in order to avoid duplication.

[xi] The Dogit Model has the probability distribution as: $p_i = \dfrac{\exp(V_i) + \theta_i \sum_j \exp(V_j)}{(1 + \sum_j \theta_j) \sum_j \exp(V_j)}$.

Model (24) is a binary Dogit with the following speficications: normalizing $V_1$ to zero, $\theta_1 = \exp(S^*)$ and $\theta_2 = 0$.

[xii] The biased estimation with small $T$ is commonly referred as "accidental parameter problem".

[xiii] Further technical details on estimation can be found in Huang (2007).

[xiv] While more recent household projection data have been published, they do not provide the growth rate by age of household representative so can not be used.

[xv] To obtain multiple-car factor for all cohorts, we follow a process similar to the sub-model of input projection, which involves calculating the future year factors in a 21 by 81 matrix.