



The
University
Of
Sheffield.

Department
Of
Economics.

Sheffield Economic Research Paper Series.

A New Formulation for Latent Class Models

Sarah Brown
William Greene
Mark N. Harris

ISSN 1749-8368

SERPS no. 2014006
April 2014

**Inst
EAD.** | Institute for
Economic Analysis
of Decision-making.

www.sheffield.ac.uk/economics

instead.group.shef.ac.uk

A New Formulation for Latent Class Models*

Sarah Brown^a William Greene^b Mark N. Harris^c

^aEconomics Department, University of Sheffield;

^bEconomics Department, Stern Business School,
New York University;

^cSchool of Economics and Finance, Curtin University

April 2014

Abstract

Latent class, or finite mixture, modelling has proved a very popular, and relatively easy, way of introducing much-needed heterogeneity into empirical models right across the social sciences. The technique involves (probabilistically) splitting the population into a finite number of (relatively homogeneous) classes, or types. Within each of these, typically, the same statistical model applies, although these are characterised by differing parameters of that distribution. In this way, the same explanatory variables can have differing effects across the classes, for example. *A priori*, nothing is known about the behaviours within each class; but *ex post*, researchers invariably label the classes according to expected values, however defined, within each class. Here we propose a simple, yet effective, way of parameterising both the class probabilities and the statistical representation of behaviours within each class, that simultaneously preserves the ranking of such according to class-specific expected values and which yields a parsimonious representation of the class probabilities.

JEL Classification: C3, D1, I1

Keywords: Latent class models, finite mixture models, ordered probability models, expected values, body mass index.

*We are grateful to the Data Archive, University of Essex, for supplying the British Household Panel Surveys, waves 14 and 16. We are also grateful to Daniel Gray for excellent research assistance and to seminar participants at the Institute of Population Health, University of Manchester, for excellent comments. We are also grateful to Anita Ratcliffe and Arne Risa Hole for excellent comments. Funding from the Australian Research Council is kindly acknowledged. The normal disclaimer applies.

1 Introduction and Background

Latent class, or finite mixture, modelling has been applied in a wide variety of areas of economics ranging from consumer behaviour (see, for example, Reboussin, Ip, and Wolfson 2008, Chung, Anthony, and Schafer 2011), to health economics (see, for example, Deb and Trivedi 2002, Bago D’Uva 2005b, Bago D’Uva 2005a) to transport mode choice (see, for example, Shen 2009) as a relatively straightforward way of introducing much-needed unobserved heterogeneity into empirical models right across the social sciences. For example, they typically represent a much more parsimonious representation of such heterogeneity than a standard random parameters approach, and moreover lend themselves to a much richer characterisation of the data under consideration by being able to group relatively homogeneous individuals into probabilistically defined, but unobserved, classes (or types, or clusters).

The technique involves (probabilistically) splitting the population into a finite number of (relatively homogeneous) classes, or types. Within each of these, typically, the same statistical model applies, although these are characterised by differing parameters of that particular distribution. In this way, the same explanatory variables can have differing effects across the classes.

Particularly with respect to examples of such empirical models in economics, several (related) estimation strategies are invariably employed. Firstly, although *a priori* nothing is known about the behaviours within each class, *ex post* researchers invariably label the classes according to expected values (*EVs*), however defined, within each class. Secondly, class probabilities do not respect the eventual labeling and ordering of the classes by *EV*. Instead, they are estimated using multinomial logit probabilities, which can become very heavily parameterised as the number of potential classes considered rises. And finally, the optimal number of unobserved classes is determined by a combination of model selection (or information) criteria (*IC*) and model (non-)convergence.¹ As *IC* metrics contain a penalty term for the

¹As an excellent example of an application of latent class modelling, and also as an example of the above points, see Bago d’Uva and Jones (2009).

number of parameters estimated, they will clearly be affected by a non-parsimonious representation of the class assignment probabilities.

Here we propose a simple, yet effective, way of parameterising both the class probabilities and the statistical representation of behaviours within each class, that simultaneously preserves their ranking according to class-specific expected values and which yields a parsimonious representation of the class probabilities. Explicitly we enforce ordering in the *EVs* across classes and suggest an ordered probabilistic specification for the class assignment probabilities, that is both consistent with the ordering in the *EVs* across classes and offers a much more parsimonious representation of the class assignment probabilities.

The model proposed here bears some superficial resemblance to that in Yang, O'Brien, and Dunson (2011), whose specification posits a latent class structure that governs the allocation of individuals to class specific subpopulations F_j . The (stochastic) ordering aspect of the model applies to F_j , not to the classes. One specification (discarded as insufficiently general) has F_j a normal population with ordering imposed on the means. The class allocation mechanism, their equation (4), is a multinomial (unordered) logit model extended to accommodate unobserved heterogeneity. In our specification, the ordering applies to the class allocation equation. The underlying implication being that class allocation itself is governed by positioning on the latent index. The class specific population is characterised by, in this case, a generic linear regression model.

Our approach is generally applicable to the analysis of any output variable which embodies a notion of ordering (either cardinal or ordinal). We illustrate our proposed technique with an example drawn from the existing health economics literature, which relates to modelling obesity levels as reflected by body mass index (*BMI*). The results show a clear preference for the suggested approach over standard ones. We also undertook a small Monte Carlo experiment which showed the fragileness of existing techniques and the robustness of the newly suggested approach. In short, regardless of the true data generating process considered for the class-assignment probabilities, the suggested technique works as well, or better, than standard ap-

proaches.

2 Econometric Framework

In a standard latent class model (*LCM*) the random variable of interest, is assumed to be drawn from a population of Q unknown and unobserved subpopulations, with corresponding mixing proportions π_q . Thus the overall density for individual i ($i = 1, \dots, N$), $f(y_i|x_i, \boldsymbol{\theta})$, is an additive mixture density of Q distinct densities weighted by their appropriate mixing probabilities π_q . The π_q are defined such that $\sum_{q=1}^Q \pi_q = 1$ and $\pi_q \geq 0 \forall q, q = 1, \dots, Q$. The outcome variable of interest is y_i , affected by the $(k_x \times 1)$ vector of covariates in the model, x_i , and where $\boldsymbol{\theta}$ denotes all of the parameters of the model.

We will assume the existence of Q latent *classes*, or *types*. These are heterogeneous across classes as to how they react to observed covariates, but homogeneous within each class. The corresponding mixture density is

$$f(y_i|x_i, \pi_1, \dots, \pi_Q; \theta_1, \dots, \theta_Q) = \sum_{q=1}^Q \pi_q \times f(y_i|x_i, \theta_q). \quad (1)$$

The usual approach to address estimation of π_q is to use a multinomial logit (*MNL*) form of the probabilities of these, given by

$$\pi_q = \frac{\exp(\gamma_q)}{\sum_{a=1}^Q \exp(\gamma_a)}, \quad (2)$$

where γ_q ($= 1, \dots, Q$) is a set of constants that are used to calculate class probabilities, and $\exp()$ is the exponential function, and where one of the γ_q is normalised to zero. However, the choice of functional form for this class assignment function is clearly inconsequential when class probabilities are treated as constants across individuals. However, this is not so when one considers an extension to this model that is increasingly used when the researcher has some prior reasoning as to the determinants of class membership; this involves an explicit parameterisation of the class assignment equation with respect to available covariates. Again, along the

lines of the *MNL* model this would become:

$$\pi_q = \frac{\exp(z_i' \gamma_q)}{\sum_{k=1}^Q \exp(z_i' \gamma_k)} \quad (3)$$

where z_i is a $(k_z \times 1)$ vector of explanatory variables that help allocate individuals to each of the unobserved classes. The *MNL* specification is evident in most (if not all) studies where class assignments are expressed as prior functions of covariates (“generalised”). Indeed, all modern econometric software estimates generalised latent class models in this manner.² Estimation can now be undertaken, using either the *EM* algorithm, or standard maximum likelihood techniques, based on equations (1) and (3).

After estimating potentially numerous variants of the *LCM* with regard to the possible number of (unknown) classes, the researcher will then clearly be faced with the choice of the most appropriate Q , Q^* . There are non-trivial issues here with regard to statistical testing across different values of $Q^* = 1, \dots$ *versus* any other potential value: for example, in testing the null of $Q^* = 1$ *versus* the alternative of $Q^* = 2$, then under the null γ (and therefore neither π_q) is (are) identified. Presumably for this reason, and moreover because the choice is essentially a model selection one, researchers invariably rely on the so-called information criteria (*IC*) metrics. These are a standard method of choosing across (potentially) non-nested models (although this is not a prerequisite). Specifically, with regard to choosing the optimal number of classes, the technique involves choosing Q^* such that

$$Q^* = \arg \min_Q IC(Q), \quad (4)$$

where:

$$IC(Q) = -2\hat{\ell}_Q + \lambda_N p_Q; \quad (5)$$

λ_N is a deterministic function of N ; and p_Q is the total number of parameters estimated in the Q class *LCM*. Some common choices of λ_N include the following:

$$\begin{array}{ll} \lambda_N = \ln N & BIC/SC \text{ (Schwarz 1978)} \\ \lambda_N = 2 & AIC \text{ (Akaike 1987)} \\ \lambda_N = 1 + \ln N & CAIC \text{ (Bozdogan 1987)} \\ \lambda_N = 2 \ln \ln N & HQIC \text{ (Hannan and Quinn 1979)}. \end{array} \quad (6)$$

²To the best of the authors’ knowledge. We use the term “generalised” here to denote the case where the class assignment probabilities are a function of covariates.

These criteria are derived from differing principles and as a result have differing properties; for example *AIC* has been shown to favour “large” models (see, for example, Hurvich and Tsai (1989)). There is no general agreement on the optimal criterion in the *LCM* setting, although there seems to be an empirical preference for *AIC* (despite - because of? - its preference for large models).

Post model estimation two estimates of the probability of class membership are available; *prior* probabilities are obtained by simply evaluating equations (2) or (3). However, more common, are the *posterior*, or *based on the data*, probabilities such that

$$\Pr(q_i | y_i) = \frac{\pi_q(z_i, \gamma_q) \times f(y_i | x_i, \theta_q)}{\sum_{k=1}^Q \pi_k(z_i, \gamma_k) \times f(y_i | x_i, \theta_k)}. \quad (7)$$

The posterior probabilities answer the question: *given that we observe y_i* what is the probability that the individual belongs to class q ?

Our point of departure concerns the specification of $f_q(y_i | x_i, \theta_q)$. Clearly the data at hand will dictate the functional form for the specification of this density: if y_i is a stochastic count, a Poisson or a Negative Binomial would be appropriate; an ordered discrete variable - an Ordered Probit/Logit; a censored continuous variable - a Tobit formulation; a continuous variable - a linear regression function; and so on. However, a defining feature of many empirical examples of *LCMs* is an *ex post* labelling of the Q classes based upon estimated *EVs* within each of the $q = 1, \dots, Q$ classes. Where ordinality exists, but there are no obvious *EVs* as such (for example in an Ordered Probit model) researchers might label classes according to the distribution of predicted probabilities at the “low” to the “higher” ends of the choice set.

It is useful here, to consider the determination of observed y_i *within* each $q = 1, \dots, Q$ class. We consider a latent index function of the form

$$y_{i,q}^* = x_i' \beta_q + \varepsilon_{i,q} \quad (8)$$

where β_q are the response parameters and $\varepsilon_{i,q}$ a disturbance term. For example, if there were no subpopulations we would have the set-up of

$$y_i^* = x_i' \beta + \varepsilon_i. \quad (9)$$

The y_q^* of equation (8) will be related to observations within group $y_{i,q}$ via a mapping dictated by $f(y_i|x_i, \theta_q)$. That is, in a linear regression model, $y_{i,q} = y_{i,q}^*$. In a Tobit setting, $y_{i,q} = \max(0, y_{i,q}^*)$. And so on. Regardless of the model, EV s (or probabilities for models such as the Ordered Probit) *on the assumption of underlying ordinality or cardinality of observed y_i* , are monotonically related to the index $x'_i\beta_q$: ensuring that $x'_i\beta_{q=1} \leq x'_i\beta_{q=2} \leq \dots \leq x'_i\beta_Q$ will therefore be a necessary and sufficient condition to ensure that $EV_{i,q=1} \leq EV_{i,q=2} \leq \dots \leq EV_{i,Q}$. For example, take the case of health-care utilisation addressed in Bago d’Uva and Jones (2009). The authors are interested, for example, in whether “low users” are more (or less) income elastic than “high users”. That is, they wish to firstly identify high and low use classes, and then to ascertain whether the drivers across these classes differ in magnitudes (and/or directions of effects). So, clearly the ranking of the classes (by expected values) is paramount in Bago d’Uva and Jones (2009), as well as in (nearly) all of the related literatures, as is the identification of these classes. Thus although in Bago d’Uva and Jones (2009) these classes are labelled *ex post*, below we suggest an easy way in which this can be enforced in estimation.

By defining generically $EV_{i,q}^*$ as the index $x'_i\beta_q$ (positively, and monotonically related to the true EV , $EV_{i,q}$), then this implicit ordering can be enforced via an estimation strategy as

$$\begin{aligned}
EV_{i,q=1}^* &= EV_{i,q=1} & (10) \\
EV_{i,q=2}^* &= EV_{i,q=1}^* + \exp(x'_i\beta_{q=2}) \\
EV_{i,q=3}^* &= EV_{i,q=2}^* + \exp(x'_i\beta_{q=3}) \\
&\vdots & \vdots
\end{aligned}$$

where the specification of $EV_{i,q=1}$ is likely to be model-specific. For example, in a linear regression $EV_{i,q=1} = x'_i\beta_q$; whilst in a count Poisson regression $EV_{i,q=1} = \exp(x'_i\beta_q)$; and so on. This is convenient in that ordering is ensured, it is applicable to a wide range of models, and has the added benefit that β_q , $q > 2$, can be directly interpreted as differential effects with respect to $EV_{i,q-1}^*$. Moreover, any variables that have no differential effects across neighbouring EV_q^* s are likely to manifest themselves by having large negative coefficients (due to the $\exp(\cdot)$ transformation).

Thus far we have shown how to enforce ordering with respect to expected values (broadly defined) in a latent class set-up. We now turn to the identification of the class probabilities. In the literature, there appear to be two major stands of how to address these. Firstly some authors do not wish to explain these class probabilities with respect to covariates, but then generally *ex post* attempt to explain the (now individual-varying) posterior probabilities of class membership by regressing them on a range of observed characteristics. However, if significant correlations are found in this second step, in some instances this could well cast doubt on the validity of the results from the first step. That is, model estimation and *ex post* estimates of quantities of interest, *including estimates of the posterior probabilities*, as this (these) would appear to be based on a mis-specified model. In general we would expect that the effect of omitted factors (variables) in a model is transmitted to the results estimated for the included factors (variables), in ways that are typically hard to predict, but regardless are likely to result in biased and mis-specified models.

If one has a notion that the classes are driven by observables, then clearly these should be allowed for in the modelling process. Indeed, as noted, it is possible that by ignoring these variables in estimation biased and mis-specified models will result. Researchers who take this approach (that is, entertain *generalised* latent class models) generally take the view that the classes, or types, are time-invariant (a kind of a “fixed effect”) and therefore best explained by any time-invariant variables available to the researcher.

Therefore, on the assumption that the researcher is parameterising the classes (presumably with respect to time-invariant variables), it is possible to reconsider the specification of the functional form for these probabilities. This are usually specified in the *MNL* form as per equation (3) above. This may be less than optimal for several important reasons. Firstly, it appears to represent a description of the probabilities that does not take advantage of the subsequent ordering applied to the estimated classes. Secondly, the *MNL* form embodies the undesirable *Independence from Irrelevant Alternatives* property.³ Here this would imply that the *odds ratio*

³See, for example, Fry and Harris (1996).

of the probability of any one class membership relative to any other, is independent of any additions to, or deletions from, the choice set. The probability of say, an individual being in class 1 (however labelled) relative to class 2, is independent of the possible existence of classes 3, 4 and 5. Clearly, as *a priori* the number of classes is unknown, this appears to be a somewhat untenable assumption. Indeed, in any other choice modelling situation, a *MNL* model approach would be unlikely to be considered.

The final reason why we believe that *MNL* class probabilities might not be ideal, relates to the number of estimated parameters such an approach entails: each additional class mandates an additional k_z parameters in the class assignment equations. This can have two adverse related consequences. Firstly, this can result in a very highly parameterised model for even small values of Q which will often cause numerical convergence problems. The true model with Q^* classes, where Q^* is a “large” number, might not even enter the researcher’s potential choice set due to numerical convergence issues. The second reason why a highly parameterised model, such as the *MNL*, might not be an ideal representation of the class assignment equation(s), relates to the *IC* metrics invariably used to determine the appropriate number of classes. As shown in equation(s) (3) above, all such metrics are “adversely” affected by the penalty term $(\lambda_N p_Q)$, which regardless of the metric, is an increasing function of p_Q . That is, the *IC*’s all depend on p_Q , and since the model size is not being determined by the likelihood statistic, but rather by the *IC*, there is a premium on parsimony: this puts the *MNL* form at a disadvantage compared to a more compact model. These two issues (of non-convergence and a larger *IC* penalty function) could jointly, or independently, result in a selected Q^* that is “too small”, and hence potentially bias any subsequent findings. Indeed, in the bulk of such empirical examples of *LCMs* we witness a preponderance of values of $Q^* \leq 3$. Consider, once more, the case of health-care utilisation considered by Bago d’Uva and Jones (2009). This research attempted to uncover the “true” number of underlying classes of individuals with respect to health-care utilisation, and moreover to ascertain any behavioural differences across the so identified classes (with a focus on income). Thus in using a more parsimonious form for the class probabilities, it may be that more than two

classes would have been identified, and that on this basis class-specific results might have been contaminated by a merging of heterogeneous classes.⁴

For all of the above reasons, we propose replacing the *MNL* probabilities with Ordered Probit (*OP*) ones (Greene and Hensher 2010). With the parameterisation suggested above of equation(s) (10) our class equation output variables ($y_{i,q}^*$) will necessarily be ordered by EV_q^* , such that such an *OP* formulation for the class probabilities will be internally, and explicitly, consistent with this ordering.

By defining an unobserved latent variable, q_i^* as a driver of the unobserved classes, itself a function of observed characteristics z_i with unknown weights γ and a (standard normally distributed) error term u_i , then *OP* probabilities of class membership can be derived along the following lines. Let q_i^* be of the form

$$q_i^* = z_i' \gamma + u_i, \quad (11)$$

where z_i has no constant term (for normalization). This latent variable translates to the (here, unobserved) ordered, discrete, class outcomes q_i ($q_i = 1, 2, \dots, Q$) via the mapping

$$q_i = q \cdot 1 \{ \mu_{q-1} < q_i^* \leq \mu_q \},$$

where $-\infty = \mu_0 < \mu_1 < \dots < \mu_{Q-1} < \mu_Q = \infty$ are the boundary parameters with $\mu = (\mu_1, \dots, \mu_{Q-1})'$ to be estimated in addition to γ . Under the assumption of normality, the *OP* probabilities are given by

$$f(q_i | z_i; \theta_{OP}) = \Phi(\mu_{q_i} - z_i' \gamma) - \Phi(\mu_{q_i-1} - z_i' \gamma),$$

where $\theta = (\gamma, \mu)$.

Moreover whereas the *MNL* approach is based upon an underlying system of $Q - 1$ latent utility equations (Greene 2012), and hence $Q - 1$ parameter vectors, the *OP* is based upon a single utility index with accordingly a single unique set of parameters per covariate. Increasing the number of classes in a *OP* set-up simply requires estimation of one additional boundary parameter, and not k_z as in the *MNL* one.

⁴The authors state that they only considered at most a 2-class model; although not stated this may well have been due to convergence issues for larger class variants.

3 Application: Body Mass Index

Given the serious health related issues associated with obesity, it is not surprising that modelling body mass index (*BMI*) and obesity rates are attracting increasing interest from both academics and policy-makers. Furthermore, as reported by the World Health Organisation (*WHO*) in 2011, since 1980, adult obesity rates have doubled worldwide. It is important to select an appropriate modelling approach in the context of such an important and highly relevant policy application. The determination of *BMI* levels have previously been addressed in the literature using a latent class framework (Greene, Harris, Hollingsworth, and Maitra 2014). The justification of such an approach here was based on medical evidence that an obesity predisposing genotype is present in 10% of individuals (Herbert, Gerry, and McQueen 2006): that is, it is (medically) very likely that individuals are genetically predisposed to being in different *BMI* classes.⁵

We analyse data drawn from the British Household Panel Survey (*BHPS*). The *BHPS* is a longitudinal survey of private households in Great Britain, 1991 to 2008, and was designed as an annual survey of each adult member of a nationally representative sample. The first wave in 1991 achieved a sample of some 5,500 households, covering approximately 10,300 adults from 250 areas of Great Britain. The same individuals are re-interviewed in successive waves and, if they split off from their original households are also re-interviewed along with all adult members of their new households. The *BHPS* is a rich source of information on labour market status, socio-demographic and health variables. In waves 14 (2004) and 16 (2006), information was collected on weight and height, which we use to calculate individuals' *BMI*. Accordingly our data set comprises of 22,430 observations covering individuals aged 16 and over. The average *BMI* in the sample is 27.06, with a standard deviation of 5.45 (Table 1), which lies in the lower end of the overweight *BMI* category suggested by the *WHO*.

⁵We note here that health professionals may not deem *BMI* an ideal measure of weight-related health status; however, it is still widely used for such and indeed is still collected in many major household surveys.

As justified above, we treat class membership as time-invariant and search for proxies for different genetic types to explain membership of these classes. Following the related literature, we include all available time invariant characteristics, such as birth cohort, gender and ethnicity.⁶ In the outcome equation, we again follow the received literature (see, for example, Cutler, Glaeser, and Shapiro 2003, Chou, Grossman, and Saffer 2004, Brown and Roberts 2013, Greene, Harris, Hollingsworth, and Maitra 2014) and control for age, number of children, marital status, household income, employment status, highest level of educational attainment, number of cigarettes smoked per day and a binary indicator for being active (specifically whether the individual walks, swims or plays sport at least once a week). Finally, we also include a set of eleven controls capturing a wide range of health problems, namely problems with: arms, legs, hands, *etc.*; sight; hearing; skin conditions/allergy; chest/breathing; heart/blood pressure; stomach or digestion; diabetes; anxiety, depression, *etc.*; migraine; and cancer.⁷ Descriptive statistics for the variables included in the empirical analysis are presented in Tables 1 and 2. In brief, the sample is evenly split by gender; just over half of the sample are married; and nearly 60% are in full-time employment. The majority of the sample is white, and having a vocational qualification is the most common highest educational attainment category.

INSERT TABLES 1 AND 2 ABOUT HERE

We firstly compare a range of different models using standard *IC* metrics in order to ascertain the preferred approach. We then present detailed estimation results based on our preferred specification. In order to evaluate the validity of the modelling approach suggested in Section 2 above, we compare the results from estimating numerous different models: we start with a one class linear regression model and then successively increase the number of latent classes within both a standard framework (unrestricted) and our new proposed framework (restricted⁸). We stopped searching

⁶Unfortunately there are no other, possibly more relevant, time-invariant variables available in the data.

⁷We consider the possible existence of reverse causation below.

⁸Note that here, and subsequently, we use the term *restricted*, not so much in the parametric sense, but more-so in that this joint approach (across class probabilities and expected values) will necessary enforce ordering in the final expected values.

for more potential classes when convergence problems were encountered within each framework. Convergence problems were encountered at $Q = 6$ for the restricted approach and 3 for the unrestricted approach. Therefore in total, we consider 6 potential models with regard to the standard *IC* metrics, which are presented in Table 3.

The *IC* metrics all support the three, four and five latent class restricted models (*i.e.*, using the newly suggested approach) over the two latent class unrestricted model (the current standard approach). This supports the modelling approach detailed in Section 2 above. The *AIC* and *HQIC* support the five latent class restricted model over the four latent class restricted model, whilst the *BIC* and *CAIC* do the opposite. In addition to the *IC* metrics we also look at simple correlations of the actual *versus* (prior probability weighted) predicted values for each model. This is labelled *Correlation* in Table 3. From these, we can see that the degree of correlation between the actual and the predicted values increases as we move from the one class to the five latent class restricted model, which further endorses the modelling approach detailed in Section 2. Based on a combination of the *IC* metrics and the correlation of actual *versus* predicted values, we take the restricted 5-class model as our preferred specification.

INSERT TABLE 3 ABOUT HERE

Table 4 presents the increasing pattern in the *EVs* from classes 1 to 5 for the restricted five latent class model and from classes 1 to 2 for the unrestricted two latent class model.⁹ The *EVs* for classes 4 and 5 lie in the obese range defined by the World Health Organisation (WHO) as a *BMI* of 30 and above. Indeed, worryingly, for class 4, the average posterior probability is “large” (at 0.22), suggesting that a large proportion of the population lie in this class. On the other hand, only 6% of the population is estimated to be in the top *BMI* range. The dispersion within these classes is relatively large, at 3.8 ($\sigma_4 = \exp(1.333)$) and 5.9 ($\sigma_5 = \exp(1.767)$)

⁹Note that these are evaluated at sample means of covariates, and for the overall value, weighted by prior probabilities. Averaged individual *EVs* gave very similar results.

respectively, for classes 4 and 5 (compared to the dispersion in the earlier classes). Only 6% of individuals are estimated to be in the lowest *BMI* class, with a tight distribution ($\sigma_1 = 1.9$); with an *EV* of just over 20, this would class these (according to the WHO) in the low end of *normal* weight. Nearly 30% of individuals are estimated to be in the upper end of *normal* (WHO classified as 18.5 – 24.9) with an *EV* of 23. However, the largest class (40%) would be WHO classified as overweight ($EV = 26$), with a variability larger than the lower classes, but smaller than the higher ones ($\sigma_3 = 2.5$).

These findings are quite distinct from the estimates from the traditional approach. With *EVs* at 25 and 32, these distributions are quite dispersed (with $\sigma_1 = 3.3$ and $\sigma_2 = 5.9$), which could be hiding the additional classes uncovered by the new approach. We re-visit this below, but one possible explanation is the *MNL* approach is suffering from convergence problems as a result of “over-fitting” such that it cannot entertain the “true” number of classes.

The class membership equation is reasonably well-specified (see Table 5), with the birth cohort controls generally driving the statistical significance.

INSERT TABLES 4 AND 5 ABOUT HERE

In Tables 6 and 7, we present the partial effects associated with our preferred five class restricted model (for demographic, and health-related variables, respectively). As would be expected, the partial effects differ quite dramatically across the five classes in terms of both size and statistical significance. In the case of age, the partial effects are positive and statistically significant in all five classes and increasing in magnitude from class 1 to class 5. The effect of being married follows a less distinct pattern, with the partial effects being positive and statistically significant in classes 1 to 4. A reduction in the magnitude of the effect is apparent from classes 1 to 3, then increasing in the case of class 4. Being employed has a significant positive effect in classes 1 and 2 only. Having a degree as the highest level of educational attainment is the only educational attainment variable to have a statistically significant effect,

with a statistically significant negative effect found for classes 2 to 5, which becomes more pronounced from classes 2 to 5. The number of cigarettes smoked is inversely associated with *BMI* in classes 1 to 5, with the largest inverse effect found in class 5.

Being active is positively associated with *BMI* in class 1 and inversely associated with *BMI* in classes 3 to 5. With respect to this variable as well as the health condition ones, we note the potential for reverse causation and that our findings relate to correlations rather than causal relationships. As expected due to the differences associated with the various health conditions, there is a wide degree of variability in terms of the partial effects with respect to sign, magnitude and statistical significance. For example, having a heart problem is positively associated with *BMI* across all five classes, with the largest effect observed in class 4, whereas the effects of mental health problems are generally statistically insignificant at the 1% level across the 5 classes.

INSERT TABLES 6 AND 7 ABOUT HERE

These results illustrate how such an approach (*LCM*), can highlight interesting differential partial effects across classes. However, the latent class approach may be simply used by some, as a tool to allow for more unobserved heterogeneity into the modelling exercise. Here one would assume that the researcher would be interested only in overall partial effects, and not those split by class. So, if the overall partials from both the 5-class restricted model and the 2-class unrestricted one were very similar, it could be argued that our suggested approach has very little benefit and/or effect in practice. Table 8 compares the overall (prior probability weighted) partial effects across the restricted 5-class model and the unrestricted 2-class model. Although the general pattern of results is broadly consistent across the two models, there are some substantive differences in terms of size and statistical significance for a number of explanatory variables (suggesting that using the unrestricted 2-class model may be yielding biased results). For example, age has a much larger effect in the restricted 5-class model, whereas the three labour market status variables are

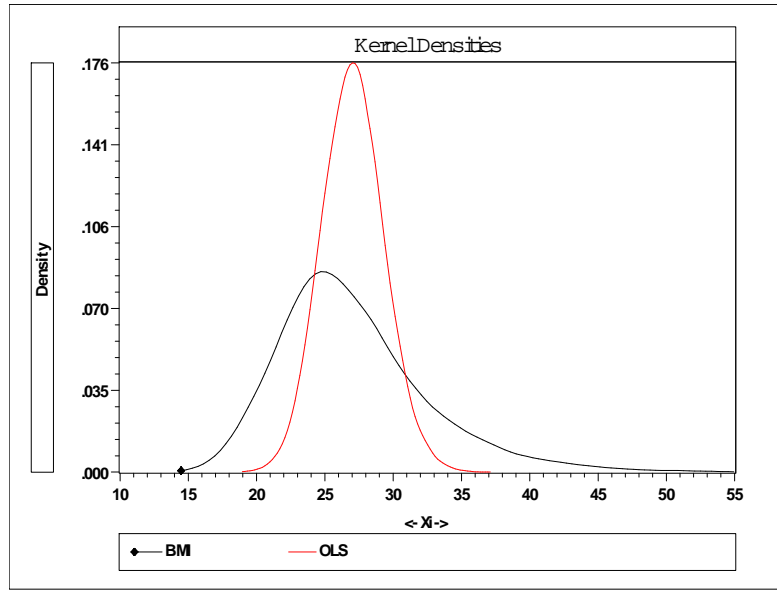


Figure 1: Kernel Densities of Observed BMI and OLS Predictions

statistically significant in the 2-class unrestricted model and statistically insignificant in the 5-class restricted model. The effects of education, number of cigarettes smoked and being active are more consistent across the two models although there are some differences in the magnitudes of the various effects. This is also the case for the health problems, where they are statistically significant at the 1% level. Such differences highlight the importance of selecting an appropriate modelling approach especially in the context of policy-relevant applications such as determining the influences on *BMI*.

INSERT TABLE 8 ABOUT HERE

To further explore behaviours within the estimated classes, and also to ascertain the overall appropriateness of our approach, we take a closer look at some estimated densities. Firstly, in Figure 1, we present kernel densities of both the raw *BMI* data, and that implied if simple regression techniques had been applied to model it. Clearly, the latter does a relatively poor of explaining the observed density.

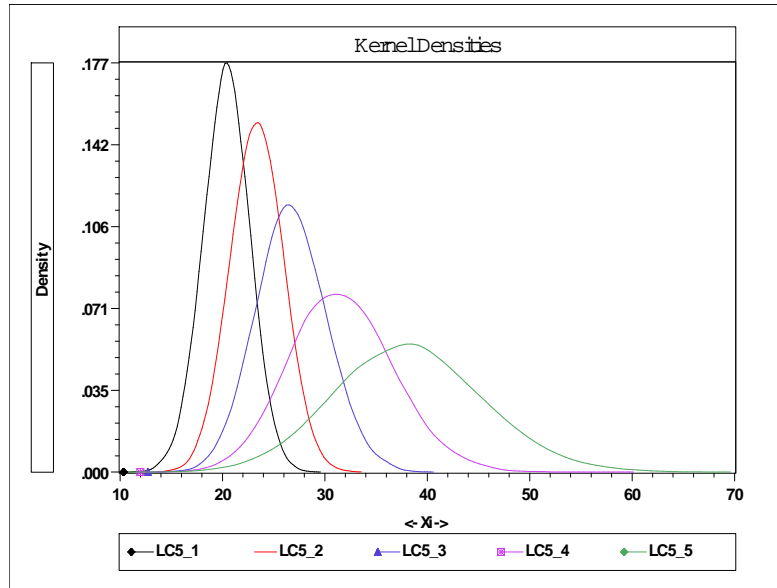


Figure 2: Kernel Density Estimates by Class

In Figure 2 we plot the implied estimated densities by class. The (enforced) ordering in these densities is evident, as their measures of central tendency (and dispersion) clearly increase over classes 1 to 5. Taken in consideration with their posterior probabilities, we can see that individuals have a very low chance of being in either the lowest or the highest *BMI* range classes. However, individuals in these are clearly likely to have very low and high, respectively, *BMI* levels with relatively low probabilities of having very high *BMI* values (for class 1) and very low levels (class 5). Interestingly, although freely estimated, the spread of these distributions clearly also increases with class. An implication of this is that although the highest *BMI* range class has a very high *EV*, it appears that behavioural choices, for example, can indeed help these individuals into more healthy *BMI* ranges. On the contrary, individuals in either of the lowest two classes, appear to be very likely closely bound to their class-specific *EV*s of low to mid 20s.

Finally, in Figure 3 we present (prior) probability weighted aggregate predicted and actual densities. As is clearly evident, our approach does an excellent job

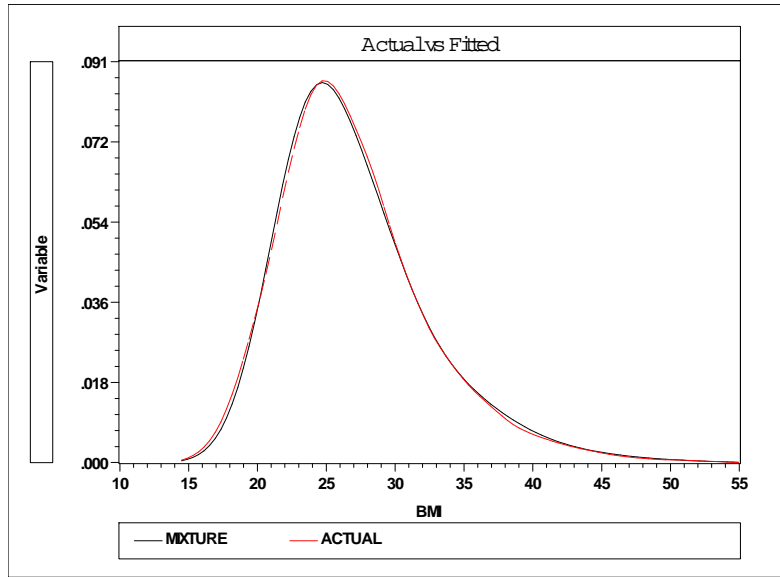


Figure 3: Kernel Density Estimates of Actual *BMI* and Overall Predicted *BMI*

in predicting the empirical density, especially as compared to a simple regression approach (Figure 1). Again, we would suggest that this is a further validation of the suggested approach.

3.1 Robustness Checks

An obvious robustness check against which to compare our model results, is to consider a constants-only variant, where following much of the *LCM* literature, the class-assignment prior probabilities are simply modelled as constants. To this end we re-estimate our model removing all covariates from the class equations. For reasons of space, we do not report the full set of results from this exercise.¹⁰ However, in Table 9 we present the model selection metrics from this exercise, along with the ones for our preferred model. Thus we can see that for all *IC* measures and for the values for *Correlation*, our approach consistently out-performs all possible contenders for the constants-only version (the overall preferred figures are presented

¹⁰These are available from the authors on request.

in **bold**, and those for the constants-only versions in *italics*). There is disagreement across the *IC metrics* as to the preferred number of classes, with *BIC* and *CAIC* favouring 3-classes, *AIC* 5-classes, and *HQIC* 4-classes. Taking these findings in conjunction with the fit of the model, as defined by the correlation values, we take the 3-class constants-only version as the preferred specification here.

The constants-only approach appears to favour a smaller number of classes, and in terms of the metrics considered, appears to perform worse than our preferred approach. However, again, if the researcher is predominantly interested in overall partial effects, then if the two approaches yield very similar results in this respect, one would presumably favour the less generalised approach. To address this, in Table 10 we compare (prior probability weighted) overall marginal effects from the 3-class constants-only approach with those previously presented from our preferred approach. In the final column we also present percentage differences in these. It is clear that the approach undertaken is consequential for these summary partial effects. For example, we find both large absolute and relative, changes in partials across the approaches, and moreover even changes in signs and significance levels of effects. For example, the (estimated) effect of age is almost halved; the number of children turns from insignificance to a significantly positive effect (as does “not in the labour force”); the effect of being employed increases over fourfold; and so on. However, we also note that large differences are not evident across-the-board: for example, the effect of the health problem (arms, legs, *etc.*) remains effectively unchanged (at 0.774 as compared to 0.732); as does the number of cigarettes smoked (-0.042 to -0.028); and being active (-0.530 compared to -0.479). We would surmise that the variables exhibiting the largest differences are probably those most severely affected by ignoring the omitted covariates in the class equation (that is, presumably the most highly correlated with the omitted drivers of the class equation); and those where the change is negligible, would be less affected (and presumably less strongly related to the omitted class covariates).

The second robustness check we consider, is that in our (*BMI*) output equation, as noted (as discussed above), we have several health indicators with the rationale that

BMI is affected by say, health problems related to breathing. However, clearly the strong possibility of reverse causation exists here, with the health condition not only causing the *BMI* level (in part), but also *BMI* levels (in part) contributing to the various health conditions. If we had appropriate identifying variables for these health conditions, that could be considered independent to *BMI*, we could apply the usual techniques for allowing for this endogeneity (for example, along the lines of Rivers and Vuong 1988, Terza, Basu, and Rathouz 2008). As always, such variables are hard to find and justify here, so instead we simply remove those likely offending variables (all of the health related ones and the activity one) and re-estimate the model. Reassuringly the results are effectively unchanged. Thus, we still find that the *MNL* can only entertain up to a 2-class model; whereas the *OP* approach can go as high as 5. The *IC* metrics similarly choose both *OP* 4 - and 5-class over the *MNL* 2-class one, and moreover are split between the choice between the two former (whilst the correlation measure again favours the *OP* 5-class model). Moreover, class-specific (and overall) *EVs*, partial effects and posterior probabilities are similarly extremely close to those models with the potentially endogenous variables included, overall leading us to the conclusion that the original results were not unduly affected by endogeneity.

4 Does the *MNL* approach tend to “over-fit”?

The empirical results presented above with regard to the new suggested approach, and the more traditional (*MNL*) one, tend to suggest that the latter might be subject to “over-fitting”, and that this could adversely affect the number of potential classes one could consider as appropriate. In essentially estimating a separate equation for each of the classes, it may well be that one, or more, covariates has no, or little, variation within a particular class, for example. Clearly issues such as this will adversely affect identification and model estimation.

To ascertain whether this is a likely finding for our application, we undertook a small Monte Carlo experiment. Here we used a 50% random sub-sample of obser-

vations used in the empirical example (which was held constant for the course of the experiment), and exactly the same model set-up as above with regard to the covariates in the model. For estimation purposes this left us with $N = 11,203$ observations. We explicitly generated the class probabilities via the *MNL* form, for a 4-class model, and then estimated the model via the *OP* and *MNL* procedures as described above. Coefficients for all parts of the model were simply generated as random numbers from a (standard) normal distribution, and then held fixed. The exceptions to this were the constants in the regression equations, which were set at 20, 30, 40 and 50, to ensure that the *EV* differed across the four classes, and the baseline *MNL* parameters which, as usual, were normalised to zero.

The results were really quite illuminating. What we found was that whilst the newly suggested (*OP*) procedure only encountered convergence problems in 15% of cases, the *MNL* (*the true data generating process*) did so in a remarkable 43% of cases.¹¹ Although it would be a stretch to generalise these findings to all such *MNL* class models, it will clearly be a potential problem in many instances, and one that the suggested procedure will predominantly circumvent.¹²

5 Conclusions

Building on the observation that in most empirical examples authors *ex post* rank and label their identified classes according to class-specific expected values, we extend the latent class methodology by proposing a procedure that allows for this ranking in estimation. We also develop a functional form for the class probabilities that is more parsimonious than the familiar multinomial logit model. There are numerous reasons why the *MNL* probabilities may not be an ideal choice in such a situation for the applied researcher (in addition to the fact that it does not take advantage of the ubiquitous ranking of classes post estimation). These

¹¹Full Monte Carlo results are available from the authors on request. Due to the length of time to estimate all models, the number of Monte Carlo repetitions was limited to 100.

¹²Indeed, we did not record in how many instances there was no variation in any variables within a particular class.

include both the unattractive *Independence from Irrelevant Alternatives* property (which appears particularly an issue for latent class modelling), as well as yielding a very heavily parameterised model. Indeed, it is our conjecture that researchers are quite often restricted in the number of classes they can estimate due to numerical convergence issues: a case of “over-fitting” (a finding confirmed by a small Monte Carlo experiment). Indeed, in our empirical example, only a two-class variant could be considered using traditional, *MNL*, class probabilities, whereas our suggested approach could estimate up to five.

The empirical example attempted to identify an unknown number of inherent classes with respect to peoples’ weight related health status, or *BMI*, levels. As noted, the traditional approach could only estimate a 2-class model, which would have been the preferred model in this case. However, our ordered approach could consider, and indeed favoured, a much more flexible mixing distribution, with up to 5-classes being supported.

The technique is widely applicable: wherever a latent class model is being applied to an output variable which embodies any ordinal, or cardinal, ordering. The suggested approach is only useful if covariates appear in the class probabilities. Otherwise the proposed model amounts only to a one-to-one reparameterisation. We would also suggest that, in general, explaining the classes with observed heterogeneity will be preferable, and will provide more reliable estimates of the posterior probabilities of class membership, and will be less likely to be adversely affected by any omitted variable bias. Indeed, in the empirical application, using a constants-only approach did appear to lead to biases in the summary partial effects measures.

References

- AKAIKE, H. (1987): “Information Measures and Model Selection,” *International Statistical Institute*, 44, 277–291.
- BAGO D’UVA, T. (2005a): “Latent Class Models for Utilisation of Health Care,” *Health Economics*, 15(4), 329–343.
- (2005b): “Latent Class Models for Utilisation of Primary Care: Evidence from a British Panel,” *Health Economics*, 14(9), 873–892.
- BAGO D’UVA, T., AND A. JONES (2009): “Health care utilisation in Europe: New evidence from the ECHP,” *Journal of Health Economics*, 28, 265–279.
- BOZDOGAN, H. (1987): “Model Selection and Akaike’s Information Criteria (AIC): The General Theory and its Analytical Extensions,” *Psychometrika*, 52, 345–370.
- BROWN, H., AND J. ROBERTS (2013): “Born to be wide? Exploring correlations in mother and adolescent body mass index,” *Economics Letters*, pp. 413–415.
- CHOU, S., M. GROSSMAN, AND H. SAFFER (2004): “An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System,” *Journal of Health Economics*, 23, 565–587.
- CHUNG, H., J. C. ANTHONY, AND J. L. SCHAFFER (2011): “Latent class profile analysis: an application to stage sequential processes in early onset drinking behaviours,” *Journal of the Royal Statistical Society: Series A*, 174, 689–712.
- CUTLER, D., E. GLAESER, AND J. SHAPIRO (2003): “Why have American become more obese?,” *Journal of Economic Perspectives*, 17(3), 93–118.
- DEB, P., AND P. TRIVEDI (2002): “The Structure of Demand for Health Care: Latent Class versus Two-Part Models,” *Journal of Health Economics*, 21(4), 601–625.

- FRY, T., AND M. HARRIS (1996): “A Monte Carlo Study of Tests for the Independence of Irrelevant Alternatives Property,” *Transportation Research - Part B*, 30B, 19–30.
- GREENE, W. (2012): *Econometric Analysis 7e*. Prentice Hall, New Jersey, USA, sixth edn.
- GREENE, W., M. HARRIS, B. HOLLINGSWORTH, AND P. MAITRA (2014): “A Latent Class Model for Obesity,” *Economics Letters*, 123, 1–5.
- GREENE, W., AND D. HENSHER (2010): *Modeling Ordered Choices*. Cambridge University Press.
- HANNAN, E., AND B. QUINN (1979): “The Determination of the Order of an Autoregression,” *Journal of the Royal Statistical Society, B*, 41, 190–195.
- HERBERT, A., N. GERRY, AND N. E. A. MCQUEEN (2006): “A Common Genetic Variant Is Associated with Adult and Childhood Obesity,” *Science*, 312, 279–283.
- HURVICH, C., AND C.-L. TSAI (1989): “Regression and time series model selection in small samples,” *Biometrika*, 76, 297–307.
- REBOUSSIN, B. A., E. IP, AND M. WOLFSON (2008): “Locally dependent latent class models with covariates: an application to under-age drinking in the USA,” *Journal of the Royal Statistical Society: Series A*, 171, 877–97.
- RIVERS, D., AND Q. VUONG (1988): “Limited information estimators and exogeneity tests for simultaneous probit models,” *Journal of Econometrics*, 39, 347–366.
- SCHWARZ, G. (1978): “Estimating the Dimensions of a Model,” *Annals of Statistics*, 6(2), 461–464.
- SHEN, J. (2009): “Latent class model or mixed logit model? A comparison by transport mode choice data,” *Applied Economics*, 41, 2915–24.
- TERZA, J. V., A. BASU, AND P. J. RATHOUZ (2008): “Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling,” *Journal of Health Economics*, 27(3), 531 – 543.

YANG, H., S. O'BRIEN, AND D. B. DUNSON (2011): "Nonparametric Bayes Stochastically Ordered Latent Class Models," *Journal of the American Statistical Association.*, 106, 807–817.

Table 1: Descriptive statistics; demographics

Variable	Mean	Standard Deviation
<i>BMI</i>	27.059	(5.45)
Birth cohort 1940	0.158	(0.36)
Birth cohort 1950	0.168	(0.37)
Birth cohort 1960	0.201	(0.40)
Birth cohort 1970	0.160	(0.37)
Birth cohort 1980	0.115	(0.32)
Birth cohort 1990	0.004	(0.06)
Female	0.504	(0.50)
White	0.976	(0.15)
(Log of) age	3.784	(0.42)
Married	0.560	(0.50)
Number of children	0.575	(0.96)
(Log of) household income	10.190	(0.74)
Employed	0.583	(0.49)
Not in the labour force	0.159	(0.37)
Unemployed	0.028	(0.16)
Degree	0.143	(0.35)
Vocational degree	0.286	(0.45)
A-level	0.117	(0.32)
GCSE	0.167	(0.37)

Table 2: Descriptive statistics; health

Variable	Mean	Standard Deviation
Health problem: arms, legs, hands, <i>etc.</i>	0.297	(0.46)
Health problem: sight	0.056	(0.23)
Health problem: hearing	0.093	(0.29)
Health problem: skin conditions/allergy	0.114	(0.32)
Health problem: chest/breathing	0.136	(0.34)
Health problem: heart/blood pressure	0.198	(0.40)
Health problem: stomach or digestion	0.086	(0.28)
Health problem: diabetes	0.049	(0.22)
Health problem: anxiety, depression, <i>etc.</i>	0.084	(0.28)
Health problem: migraine	0.070	(0.25)
Health problem: cancer	0.050	(0.22)
Number of cigarettes	3.598	(7.31)
Active	0.577	(0.49)

Table 3: Model selection metrics

	<i>BIC</i>	<i>AIC</i>	<i>CAIC</i>	<i>HQIC</i>	<i>Correlation</i>
Linear Regression	138,210	138,014	138,236	138,080	0.2759
2-class (restricted)	134,442	133,982	134,503	134,135	0.2970
3-class (unrestricted)	134,437	133,977	134,498	134,131	0.2975
3-class (restricted)	134,128	133,464	134,216	133,685	0.2001
3-class (unrestricted)	—	—	—	—	—
4-class (restricted)	134,088	133,221	134,203	133,510	0.3045
4-class (unrestricted)	—	—	—	—	—
5-class (restricted)	134,160	133,089	134,302	133,446	0.3087
5-class (unrestricted)	—	—	—	—	—

Note: preferred model for each metric in **bold**.

Table 4: Expected values, averaged posterior probabilities and dispersion parameters

	<i>Q = 5; OP</i>			<i>Q = 2; MNL</i>		
	<i>Expected</i>	<i>Posterior</i>	<i>ln (Dispersion)</i>	<i>Expected</i>	<i>Posterior</i>	<i>ln (Dispersion)</i>
	<i>Value</i>	<i>probability</i>		<i>Value</i>	<i>probability</i>	
Class 1	20.37 (0.24)**	0.06	0.667 (0.06)**	25.09 (0.05)**	0.28	1.184 (0.01)**
Class 2	23.24 (0.17)**	0.27	0.687 (0.05)**	31.91 (0.20)**	0.72	1.767 (0.01)**
Class 3	26.46 (0.22)**	0.39	0.929 (0.04)**	—	—	—
Class 4	31.27 (0.48)**	0.22	1.333 (0.04)**	—	—	—
Class 5	37.61 (0.86)**	0.06	1.767 (0.05)**	—	—	—
Overall	26.90 (0.04)**	—	—	27.00 (0.04)**	—	—

Notes: ** and * denote significant at 5, and 10% size, respectively. The preferred model for each metric in **bold**.

Table 5: Class membership equation; preferred specification

Variable	Estimated coefficient	Standard error
Female	-0.252	(0.02)**
Birth cohort 1940	0.488	(0.04)**
Birth cohort 1950	0.685	(0.06)**
Birth cohort 1960	0.862	(0.08)**
Birth cohort 1970	0.908	(0.10)**
Birth cohort 1980	0.775	(0.13)**
Birth cohort 1990	0.826	(0.23)**
White	0.142	(0.06)**
μ_1	-1.032	(0.10)**
μ_2	0.162	(0.13)**
μ_3	1.231	(0.14)**
μ_4	2.221	(0.14)**

Notes: ** and * denote significant at 5, and 10% size, respectively.

Table 6: Class-specific partial effects; demographics

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
(Log of) age	1.278 (0.23)**	2.226 (0.20)**	3.355 (0.41)**	3.894 (0.80)**	4.417 (1.21)**
Married	0.612 (0.14)**	0.605 (0.09)**	0.447 (0.10)**	0.752 (0.21)**	-0.012 (0.49)
Number of children	0.011 (0.08)	-0.084 (0.05)	0.045 (0.06)	-0.179 (0.12)	0.120 (0.21)
(Log of) household income	0.004 (0.11)	0.123 (0.08)	0.099 (0.09)	-0.219 (0.14)	-0.521 (0.35)
Employed	0.759 (0.22)**	0.438 (0.13)**	0.271 (0.17)	0.483 (0.37)	0.362 (0.96)
Not in the labour force	-0.434 (0.27)	-0.243 (0.16)	-0.027 (0.18)	0.147 (0.40)	0.984 (1.02)
Unemployed	-0.614 (0.65)	-0.015 (0.30)	0.134 (0.32)	0.475 (0.62)	2.208 (1.23)*
Degree	-0.421 (0.25)*	-0.683 (0.15)**	-0.771 (0.19)**	-1.571 (0.37)**	-2.433 (0.77)**
Vocational degree	0.235 (0.19)	-0.362 (0.11)**	-0.206 (0.12)*	-0.168 (0.26)	-0.555 (0.57)
A-level	0.197 (0.25)	-0.132 (0.15)	-0.526 (0.20)**	-0.742 (0.34)**	-0.762 (0.72)
GCSE	-0.083 (0.20)	-0.234 (0.12)*	0.054 (0.14)	-0.011 (0.33)	-2.067 (0.83)**

Notes: ** and * denote significant at 5, and 10% size, respectively.

Table 7: Class-specific partial effects; health

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
Health problem: arms, legs, hands, <i>etc.</i>	0.001 (0.20)	0.589 (0.09)**	0.675 (0.11)**	1.314 (0.23)**	1.033 (0.58)*
Health problem: sight	-0.381 (0.24)	-0.398 (0.16)**	-0.097 (0.17)	-0.257 (0.39)	0.273 (0.99)
Health problem: hearing	-0.081 (0.18)	-0.301 (0.13)**	-0.354 (0.15)**	-0.218 (0.31)	0.355 (0.78)
Health problem: skin conditions/allergy	0.212 (0.18)	0.221 (0.12)*	0.128 (0.14)	0.484 (0.26)*	-0.791 (0.80)
Health problem: chest/breathing	-0.274 (0.17)	0.166 (0.11)	0.393 (0.12)**	0.649 (0.25)**	1.956 (0.58)**
Health problem: heart/blood pressure	1.109 (0.16)**	1.104 (0.10)**	1.482 (0.14)**	2.053 (0.26)**	1.807 (0.71)**
Health problem: stomach or digestion	-0.548 (0.21)**	-0.434 (0.13)**	-0.132 (0.15)	0.275 (0.28)	-0.429 (0.88)
Health problem: diabetes	0.735 (0.25)**	1.047 (0.18)**	1.816 (0.18)**	2.580 (0.35)**	2.149 (1.21)*
Health problem: anxiety, depression, <i>etc.</i>	-0.343 (0.21)	-0.198 (0.14)	0.230 (0.15)	0.245 (0.31)	1.442 (0.71)**
Health problem: migraine	-0.302 (0.23)	-0.166 (0.15)	0.094 (0.16)	1.334 (0.32)**	-2.449 (1.50)
Health problem: cancer	-0.029 (0.35)	-0.014 (0.16)	0.304 (0.21)	0.543 (0.37)	1.867 (0.91)**
Number of cigarettes	-0.042 (0.01)**	-0.059 (0.01)**	-0.032 (0.01)**	-0.027 (0.01)**	-0.081 (0.03)**
Active	0.500 (0.13)**	-0.017 (0.08)	-0.473 (0.10)**	-1.049 (0.19)**	-2.461 (0.44)**

Notes: ** and * denote significant at 5, and 10% size, respectively.

Table 8: Overall partial effects

	$Q = 5; OP$		$Q = 2; MNL$	
(Log of) age	3.113	(0.38)**	0.915	(0.20)**
Married	0.539	(0.08)**	0.611	(0.08)**
Number of children	-0.036	(0.04)	0.026	(0.04)
(Log of) household income	-0.001	(0.06)	0.039	(0.06)
Employed	0.392	(0.16)**	1.057	(0.13)**
Not in the labour force	-0.015	(0.18)	0.431	(0.15)**
Unemployed	0.241	(0.34)	1.015	(0.25)**
Degree	-0.991	(0.12)**	-0.888	(0.12)**
Vocational degree	-0.238	(0.10)**	-0.106	(0.10)
A-level	-0.441	(0.14)**	-0.286	(0.13)**
GCSE	-0.161	(0.12)	-0.180	(0.11)
Health problem: arms, legs, hands, <i>etc.</i>	0.774	(0.09)**	0.748	(0.08)**
Health problem: sight	-0.209	(0.19)	-0.309	(0.15)**
Health problem: hearing	-0.258	(0.16)*	-0.288	(0.13)**
Health problem: skin conditions/allergy	0.185	(0.11)*	0.099	(0.11)
Health problem: chest/breathing	0.436	(0.11)**	0.514	(0.10)**
Health problem: heart/blood pressure	1.499	(0.10)* *	1.406	(0.10)**
Health problem: stomach or digestion	-0.164	(0.14)	-0.165	(0.12)
Health problem: diabetes	1.732	(0.24)**	1.851	(0.17)**
Health problem: anxiety, depression, <i>etc.</i>	0.152	(0.14)	0.160	(0.15)
Health problem: migraine	0.132	(0.16)	-0.120	(0.14)
Health problem: cancer	0.336	(0.19)*	0.267	(0.16)*
Number of cigarettes	-0.042	(0.01)**	-0.038	(0.00)**
Active	-0.530	(0.07)**	-0.478	(0.07)**

Notes: ** and * denote significant at 5, and 10% size, respectively.

Table 9: Model selection metrics; comparison with constants-only approach

	BIC	AIC	$CAIC$	$HQIC$	$Correlation$
Linear Regression	138,210	138,014	138,236	138,080	0.2759
5-class (restricted)	134,160	133,089	134,302	133,446	0.3087
2-class (constants)	134,526	134,126	134,579	134,260	0.2748
3- class (constants)	<i>134,319</i>	133,715	<i>134,399</i>	133,916	<i>0.2755</i>
4-class (constants)	134,349	133,542	134,456	<i>133,810</i>	0.2754
5-class (constants)	134,522	<i>133,511</i>	134,656	133,848	0.2753

Note: preferred model for each metric in **bold**. Preferred model for the constants-only versions in *italics*.

Table 10: Overall partial effects; comparison with constants-only approach

	5-class restricted	3-class constants only	Difference (%)
(Log of) age	3.113**	1.729**	80.0
Married	0.539**	0.713**	-24.4
Number of children	-0.036	0.079**	-146.0
(Log of) household income	-0.001	0.054	-101.6
Employed	0.392**	1.747**	-77.6
Not in the labour force	-0.015	1.002**	-101.5
Unemployed	0.241	1.787**	-86.5
Degree	-0.991**	-0.844**	17.3
Vocational degree	-0.238**	-0.036	554.3
A-level	-0.441**	-0.335**	31.4
GCSE	-0.161	-0.205*	-21.6
Health problem: arms, legs, hands, etc.	0.774**	0.732**	5.8
Health problem: sight	-0.209	-0.267*	-21.9
Health problem: hearing	-0.258*	-0.306**	-15.6
Health problem: skin conditions/allergy	0.185*	0.095	95.0
Health problem: chest/breathing	0.436**	0.532**	-17.9
Health problem: heart/blood pressure	1.499**	1.317**	13.8
Health problem: stomach or digestion	-0.164	-0.044	271.4
Health problem: diabetes	1.732**	1.914**	-9.5
Health problem: anxiety, depression, etc.	0.152	0.266*	-42.9
Health problem: migraine	0.132	-0.066	-298.0
Health problem: cancer	0.336*	0.257	30.4
Number of cigarettes	-0.042**	-0.028**	47.4
Active	-0.530**	-0.479**	10.6

Notes: ** and * denote significant at 5, and 10% size, respectively.