Article:

Brown, S. orcid.org/0000-0002-4853-9115, Greene, W.H. and Harris, M. (2020) A novel approach to latent class modelling: identifying the various types of body mass index individuals. Journal of the Royal Statistical Society: Series A, 183 (3). pp. 983-1004. ISSN 0964-1998

https://doi.org/10.1111/rssa.12552

Online Appendix: A novel approach to latent class modelling: Identifying the various types of Body Mass Index individuals

October 2019

1 Finite sample performance

To examine the validity of our modelling approach, we undertake a Monte Carlo (MC) analysis. We generate under numerous scenarios, for up to $Q = 6$ classes (our preferred specification in the empirical application). For each Q we consider a data generating process (DGP) for both the OP and MNL specifications. In each experiment we searched for models up to two classes in excess of the true DGP (as convergence issues were frequently encountered otherwise), up to a maximum of $Q = 7$. Due to the number of models estimated in each repetition and the time they took to run, the number of Monte Carlo repetitions was set equal to 100 in all instances.

A range of outputs was collected, but only for the correct DGP class model (OP and MNL variants): the proportion of times the respective ICs selected the correct model over all models and approaches considered; the same but just considering whether the correct class model was selected within only OP or MNL variants ($Within\ IC$); three $Vuong$ tests (as above, based on the BIC favoured OP and MNL models; the same for AIC favoured ones; and finally, the two preferred models, irrespective of approach, based on BIC and AIC metrics); the average proportion of correct class predictions based on the maximum - posterior - probability rule ($Correct$) ; and finally differences between the average estimated EVs and the true ones, evaluated at both an individual level and then averaged, $EV_q(x_i)$, and at sample means, $EV_q(\overline{x})$. We should note that as with any MC experiment these results cannot necessarily be generalised to all other situations; however, they do clearly demonstrate the validity of the approach in the current context and moreover, give greater confidence to the empirical findings.

In Table 1 we present the findings from the 2-class and 3-class MNL and OP DGPs, respectively. As, arguably, the most important metrics here are the model selection ones, these are highlighted in **bold**. Thus we can see that for a simple 2-class model, all of the IC and $Vuong$ statistics do an excellent job in correctly selecting the 2-class model. Indeed, only the AIC does not correctly select this model in all instances. Choosing the preferred model on the basis of the top two OP and MNL performing models (on the basis of AIC) the $Vuong$ test, $Vuong\ (AIC)$, correctly selects in 99% of instances; whereas those based on the best two performing models with respect to the both BIC, $Vuong\ (BIC)$, and best BIC and AIC models, $Vuong\ (BIC, AIC)$, have a 100% record here. Moreover, the percentage of correct class predictions appears high (at 84%), and all EVs are extremely accurately estimated.

Generating for a 3-class model now allows us to consider both the MNL and OP variants. Thus we can see that when the DGP is a 3-class MNL model, all ICs correctly select the MNL $Q = 3$ model over all other Q models, including all possible OP variants. All of the $Vuong$ statistics also correctly select the $Q = 3$ MNL model; all EVs are extremely accurately estimated; and the model correctly predicts class membership in some 98% of instances. The (mis-specified) OP $Q = 3$ model, does an equally good job at predicting class membership (also at 98%); estimation of class-specific EVs are on a par with those from the (correct) MNL $Q = 3$ model; and finally for all OP variants only, the $Within$ ICs show that especially if using BIC and $CAIC$, one would still correctly select the 3-class model.

These results are effectively mirrored when generating with a OP $Q = 3$ model (last two columns, Table 1). Here, apart from the AIC (at 88%), the ICs always correctly select the true model across all models considered; and these favourable results are also reflected in the performance of the $Vuong$ tests. The remaining performance across OP and MNL $Q = 3$ models is quite similar, with the former performing slightly better in terms of predicting class membership (69 compared to 68%), whereas estimation of EV_q is very accurate and similar across both. Finally, if only looking at model selection within all MNL models, the ICs do a reasonable job at correctly selecting the correct class model, in particular the $HQIC$ selects this in 100% of instances (although $CAIC$ and BIC are much lower).

Table 2 presents the results when $Q = 4$. When the DGP is MNL (first two columns), both the OP and MNL approaches have similar performances with respect to class membership (at 77% each), and estimation of EV_q is both very accurate and similar across approaches (marginally better in the MNL model). In terms of model selection though, somewhat surprisingly, both the BIC and $CAIC$ metrics seem to have difficulty in distinguishing between the $Q = 4$ MNL and OP models: for example, BIC only correctly predicts the MNL model in 30% of instances, whereas in all other instances (70%) it favours the OP one. Moreover, this relatively poor performance of the BIC metric is reflected in the relatively poor performance of the $Vuong$ tests here. However, it is interesting to note that the $HQIC$ metric continues to perform excellently here, correctly selecting MNL $Q = 4$ in all instances. The difficulty here for the ICs in choosing across the OP and MNL variants is further evidenced by the $Within$ IC results for the MNL model, where success rates are at 100% (except

Table 1: Monte Carlo class 2 and 3 results

DGP	2-class	3-class MNL		3-class OP	
	MNL	OP	MNL	OP	MNL
$Correct$	0.84	0.98	0.98	0.69	0.68
$EV_1(\overline{x})$	0.00	0.07	0.06	0.01	0.08
$EV_2(\overline{x})$	0.04	0.14	0.04	0.16	0.07
$EV_3(\overline{x})$	–	0.15	0.01	0.15	0.31
$EV_1(x_i)$	0.01	0.32	0.18	0.14	0.06
$EV_2(x_i)$	0.33	0.03	0.02	0.08	0.11
$EV_3(x_i)$	–	0.27	0.94	0.31	0.24
BIC	**1.00**	0.00	**1.00**	**1.00**	0.00
AIC	**0.83**	0.00	**1.00**	**0.88**	0.00
$CAIC$	**1.00**	0.00	**1.00**	**1.00**	0.00
$HQIC$	**1.00**	0.00	**1.00**	**1.00**	0.00
$Within\ IC$					
BIC	–	1.00	1.00	1.00	0.43
AIC	–	0.58	1.00	0.88	0.90
$CAIC$	–	1.00	1.00	1.00	0.07
$HQIC$	–	0.87	1.00	1.00	1.00
$Vuong\ (BIC)$	**1.00**	0.00	**1.00**	**1.00**	0.00
$Vuong\ (AIC)$	**0.99**	0.00	**1.00**	**0.88**	0.06
$Vuong\ (BIC, AIC)$	**1.00**	0.00	**1.00**	**1.00**	0.00

Table 2: Monte Carlo class 4 results

DGP	4-class MNL		4-class OP	
	OP	MNL	OP	MNL
$Correct$	0.77	0.77	0.64	0.63
$EV_1\,(\overline{x})$	0.02	0.00	0.06	0.19
$EV_2\,(\overline{x})$	0.04	0.02	0.18	0.19
$EV_3\,(\overline{x})$	0.58	0.04	0.27	0.44
$EV_4\,(\overline{x})$	0.83	0.07	0.01	0.45
$EV_1\,(x_i)$	0.02	0.00	0.15	0.24
$EV_2\,(x_i)$	0.06	0.08	0.10	0.39
$EV_3\,(x_i)$	0.66	0.05	0.44	0.29
$EV_4\,(x_i)$	0.39	0.10	0.19	0.48
BIC	0.70	**0.30**	**0.98**	0.00
AIC	0.00	**0.95**	**0.93**	0.00
$CAIC$	0.94	**0.06**	**0.84**	0.00
$HQIC$	0.00	**1.00**	**1.00**	0.00
$Within\ IC$				
BIC	1.00	1.00	0.98	0.00
AIC	0.79	0.95	0.95	0.75
$CAIC$	1.00	1.00	0.84	0.00
$HQIC$	1.00	1.00	1.00	0.75
$Vuong(BIC)$	0.09	**0.00**	**0.98**	0.00
$Vuong(AIC)$	0.11	**0.20**	**0.93**	0.00
$Vuong(BIC, AIC)$	0.12	**0.30**	**0.98**	0.00

AIC at 95%). It is hard to know what is causing the slightly below par performance of the IC metrics here, this may have something to do with the particular DGP considered.

The strong performance of the IC metrics return when the DGP is $Q = 4$ OP (columns 3 and 4, Table 2). $HQIC$ once again achieves a perfect score, with BIC just behind at 98%. The good performance of these is mirrored in the near 100% performance of the $Vuong$ statistics. However, this time the only IC metrics to have any power in correctly detecting the correct class MNL model ($Within\ IC$), are the AIC and $HQIC$ ones (both at 75%), with both BIC and $CAIC$ never selecting the $Q = 4$ MNL model. Once more EVs are very accurately estimated across both approaches (marginally more-so for OP) and class membership prediction rates are high for both (64 and 63%, respectively, for the OP and MNL models).

Table 3 presents the $Q = 5$ results. When the true DGP is MNL (columns 2 and 3) all ICs and $Vuong$ tests correctly select the MNL $Q = 5$ model. This DGP seems to adversely affect the performance of the OP model though, with class predictions much inferior (52 compared to 87%); and whilst the EV_q quantities are closely estimated by the MNL model, those for the OP are out now by over just decimal places, as has been the case in all previous experiments (those at $\overline{x}$ seem to be worst affected, and are out by up to 3.5 BMI units for EV_4). $Within\ IC$ metrics behave reasonably well for the OP approach, peaking at 98% for AIC, but with $HQIC$ also performing well at 75%.

For the OP DGP here (columns, 3 and 4, Table 3), again all IC metrics and $Vuong$ tests correctly select the true model, and class prediction across models is very similar (at 80%, OP and 78%, MNL). Interestingly, once more this class DGP seems to adversely affect the ability of the OP model to accurately estimate EV_q, even when it is the true DGP. Although not as biased as before, they are out, on average, by just over 2 BMI units for $EV_5(\overline{x})$ and by nearly 2 for $EV_4(x_i)$. It is unclear why these appear to be adversely affected in the $Q = 5$ scenario(s), especially so in light of the metrics doing a very good job in correctly identifying the $Q = 5$ OP model. The $Within\ IC$ metrics for the MNL model here do an exceptionally good job of correctly identifying the $Q = 5$ model, with AIC being the lowest at some 99%.

Finally, Table 4 presents the $Q = 6$ results; when the true DGP is MNL (columns 2 and 3) all ICs and $Vuong$ tests correctly select the MNL $Q = 6$ model. All class-specific EVs are accurately estimated by the MNL model, and moreover it correctly estimates 85% of class membership. However, the $Within\ IC$ metrics show that these infrequently fail to correctly identify the correct class OP model; it only correctly predicts class membership 42% of the time; and, especially for the lower class EV_q can be out up to as much as 7.55 BMI units. For the OP DGP (Table 4, columns 3 and 4), all metrics effectively correctly select the OP $Q = 6$ model all of the time; and the model correctly predicts class membership 80% of the time (as compared to the MNL model at 55%). Interestingly, the $Within\ IC$ metrics never select the correct class model for the MNL approach. The EV_q quantities are very closely estimated by the OP approach (especially so at the individual-averaged level). However, those for the MNL can be quite biased, up to 7.55 ($EV_5(\overline{x})$) and 5.88 ($EV_6(x_i)$).

1.1 Summary of finite sample results

The above experiments show, with only a few exceptions, that the model selection metrics all have exceptionally good performance in correctly selecting the correct/true model. In particular, the $HQIC$ one has 100% performance in all scenarios considered. Indeed, these

Table 3: Monte Carlo class 5 results

DGP	5-class MNL		5-class OP	
	OP	MNL	OP	MNL
$Correct$	0.52	0.87	0.80	0.78
$EV_1(\overline{x})$	0.27	0.43	0.06	0.07
$EV_2(\overline{x})$	1.60	0.56	0.36	0.05
$EV_3(\overline{x})$	1.68	0.92	1.24	0.03
$EV_4(\overline{x})$	3.52	0.07	1.57	0.38
$EV_5(\overline{x})$	3.37	0.25	2.05	0.71
$EV_1(x_i)$	0.04	0.25	0.07	0.04
$EV_2(x_i)$	1.60	0.04	0.03	0.33
$EV_3(x_i)$	0.57	0.09	0.01	0.87
$EV_4(x_i)$	1.69	0.01	1.94	1.95
$EV_5(x_i)$	0.21	0.16	0.45	0.52
BIC	0.00	**1.00**	**1.00**	0.00
AIC	0.00	**1.00**	**1.00**	0.00
$CAIC$	0.00	**1.00**	**1.00**	0.00
$HQIC$	0.00	**1.00**	**1.00**	0.00
$Within\ IC$				
BIC	0.21	1.00	1.00	1.00
AIC	0.98	1.00	1.00	1.00
$CAIC$	0.10	1.00	1.00	0.99
$HQIC$	0.75	1.00	1.00	1.00
$Vuong(BIC)$	0.00	**1.00**	**1.00**	0.00
$Vuong(AIC)$	0.00	**1.00**	**1.00**	0.00
$Vuong(BIC, AIC)$	0.00	**1.00**	**1.00**	0.00

Table 4: Monte Carlo class 6 results

DGP	6-class MNL		6-class OP	
	OP	MNL	OP	MNL
$Correct$	0.42	0.85	0.80	0.55
$EV_1\,(\overline{x})$	4.09	0.00	0.07	0.53
$EV_2\,(\overline{x})$	7.55	0.02	0.45	4.56
$EV_3\,(\overline{x})$	6.61	0.19	1.02	3.18
$EV_4\,(\overline{x})$	4.49	0.31	1.42	1.80
$EV_5\,(\overline{x})$	0.27	0.03	1.88	7.55
$EV_6\,(\overline{x})$	2.09	0.01	2.51	4.58
$EV_1\,(x_i)$	4.09	0.08	0.01	0.43
$EV_2\,(x_i)$	6.96	0.00	0.19	4.42
$EV_3\,(x_i)$	6.05	0.16	0.77	2.82
$EV_4\,(x_i)$	3.83	0.46	0.10	1.62
$EV_5\,(x_i)$	0.45	0.01	0.32	4.18
$EV_6\,(x_i)$	3.05	0.01	0.21	5.88
BIC	0.00	**1.00**	**1.00**	0.00
AIC	0.00	**1.00**	**1.00**	0.00
$CAIC$	0.00	**1.00**	**0.99**	0.00
$HQIC$	0.00	**1.00**	**1.00**	0.00
$Within\ IC$				
BIC	0.33	1.00	1.00	0.00
AIC	0.33	1.00	1.00	0.00
$CAIC$	0.33	1.00	0.99	0.00
$HQIC$	0.67	1.00	1.00	0.00
$Vuong(BIC)$	0.00	**1.00**	**1.00**	0.00
$Vuong(AIC)$	0.00	**1.00**	**1.00**	0.00
$Vuong(BIC, AIC)$	0.00	**1.00**	**1.00**	0.00

results lead to great confidence in relying on these metrics and tests, both in general, and in particular for modelling BMI. Typically EV_q values are quite closely estimated, even if the wrong model, but correct class, approach is chosen. Indeed, the metrics quite often will correctly select the right number of classes, even for the wrong approach. Correct class predictions are generally very high, with a value of around 80% being common. There is some evidence that model performance for both approaches, in particular with respect to estimation of EV_q, declines as the true number of classes increases, but intuitively this is to be expected.

It is interesting to relate these findings back to the analysis of BMI. Reassuringly in the empirical analysis, there was complete consensus across all IC variants and $Vuong$ tests over the superiority of the OP $Q = 6$ model. These findings give us great confidence in the results of our suggested approach and the consequent findings. We do note though, that the OP $Q = 6$ experimental results suggest that it might be prudent to take into account the estimated standard errors of EV_q in interpretation of the empirical results.

The fact that these MC experiments show that both approaches often have quite similar performance across DGP's is not to be taken as an indication that the form of the approach taken will be inconsequential in practice. Indeed, the results from modelling BMI make this quite clear, with many quantities of interest being quite distinct across approaches. In reality, it is likely that neither of these approaches represent an exact description of the true DGP, but the choice is more so of which one more closely mimics this reality in a parsimonious manner, as compared with the "clinical laboratory" conditions of the MC experiments. Given the results presented in this paper, it is our conjecture that this will, more often than not, be provided by the newly suggested OP approach.

2 Estimation considerations

2.1 Model Identification

The within class model is identified: it is an OP model with a nonlinear index function. OP models are identified in the absence of multicollinearity. Here, there is no issue of identification within class. As long as there is variation across classes, the whole model is identified. The issue of non-identification only arises if the number of classes in the specified model exceeds what is supported by the data – *i.e.*, the DGP. This possibility of failure due to over-specification (too many classes) is actually useful, as it becomes evident in the results if the model is specified with too many classes – essentially infinite standard errors and extremely small class probabilities for some of the classes.

A mathematical proof of identifiability would require, first, a definition of identification. Most observers agree that a sufficient condition is full rank of the Hessian. We can only verify this based on the logical argument above and on the nonsingularity of the estimator of the asymptotic covariance matrix of the estimated coefficients. We do not encounter any issues in the second case save for the aforementioned over-specification. Under this latter condition, the unidentification seems to reveal itself via the asymptotic covariance matrix. (We note, this aspect of latent class models is discussed in Heckman and Singer (1984); Heckman, J. and B. Singer, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration," *Econometrica*, 52, 1984, pp. 271-320.)

2.2 Estimation algorithms; maximum likelihood *versus* EM

Model parameters were estimated by maximum likelihood. The algorithm used is predominantly the $BFGS$ gradient method. Likelihoods for latent class models are sometimes maximized by the EM algorithm. However, this method cannot be used for this model because the class specific functions are not separable: due to the imposed ordering across classes, β_1 appears in the conditional mean function of *all* Q class specific functions. The so-called M step of the EM algorithm involves computation for each class separately, which would not impose the cross-class equality constraints required here. On the other hand, maximum likelihood estimation is generally routine and conveniently allows the construction of the full model. In estimation we also used algorithmic derivatives, whereas analytical ones are likely to improve convergence performance. We also used numerical procedures where appropriate, to evaluate relevant quantities of interest, and corresponding standard errors were obtained using the Delta method. Robust standard errors were calculated using the usual outer product of the gradient (OPG) estimator for the parameters of the model.

2.3 Start values

Starting values for the OP procedure were obtained in the following manner.

1. The MNL 2-class model was firstly estimated, using OLS values for the regression and variance terms (perturbed for one of the classes), β_q and $V(\varepsilon_q) = \sigma_q^2$, $q = 1, 2$. In estimation to ensure well-defined variances/standard deviations, these entered the

likelihood functions as $\sigma_q = \exp(\omega_q)$, where ω_q is freely estimated. Starting values for the single parameter vector γ required for a 2-class model, were obtained by a random draw from $N(0,1)/10$. Note that here, and elsewhere where appropriate, the user-written *Gauss* code was benchmarked against the available commercial software (*c.f., Limdep/Nlogit* and *Stata*).

2. Based on $\hat{\gamma}$ from 1., a 2-class restricted variant was estimated where start values for $\beta_{q=2}\,(restricted)$ were given by $\widehat{\beta}_{q=2}\,(unrestrricted)/100$. We note here that we do not consider this as a valid OP variant, as due to the 2-class nature of the model, one class by definition must embody a higher (lower) EV than the other one; and moreover the probabilistic expressions for both will be identical. However, we use it simply as a tool for providing sensible start values for the 3-class variant.

3. For the 3-class OP variant, we require start values for γ, μ_1, μ_2, β_q and ω_q, $q = 1, 2, 3$. We set the $\boldsymbol{\mu}$ values to simply split the standard normal distribution into equal parts: $\mu_1 = \Phi^{-1}(1/3)$ and $\mu_2 = \Phi^{-1}(1/3)$. We set the start value for $\beta_1 = \widehat{\beta}_1$ from 2.; all other start values were set equal to zero. Note that in estimation we used the in-built *Gauss cmlMT* inequality constraint function to ensure the requisite ordering in the μ_q parameters throughout. If such a function is unavailable, one could equivalently use $\mu_q = \mu_{q-1} + \exp(a_q)$, where a_q would be freely estimated.

4. For the 4-class model, a similar progression was followed for start values: $\mu_1 = \Phi^{-1}(1/4)$; $\mu_2 = \Phi^{-1}(1/2)$; $\mu_3 = \Phi^{-1}(3/4)$; $\beta_1 = \widehat{\beta}_1$ (from 3.); all other start values were set to zero.

5. Start values for the 5-class model continued this progression and so on.

We should note that, so long as sensible start values were given, the maximum likelihood estimates ended-up at the same values, but speed of convergence was sometimes affected. However, the procedures described above may not necessarily be optimal for all applications. In practice it might be advisable to try a range of different start values, and to enter previously solved final estimates as new start values to ensure that the likelihood has achieved a global, and not local, maximum.

Note also that *Gauss* code is freely available at:

https://drive.google.com/drive/folders/1rtoYfs5qfwcI4NcFpq0-pdhLJOZZEa56?usp=sharing.

Usual LCMs with the MNL form the class probabilities, can be routinely estimated in packages such as *Latent Gold*, *NLOGIT*, *Stata v15* and some packages in *R*; see Grun, B., and F. Leisch (2007): "FlexMix: An R Package for Finite Mixture Modelling,", Discussion paper, Faculty of Commerce, University of Wollongong, Faculty of Commerce, University of Wollongong.

3 Illustrative likelihood function

Below is the Gauss script likelihood function for a 3-class model:

```
proc LC3_MLE(struct PV p, y, x_lc, x_reg, ind);
local gama, mu, beta_1, beta_2, beta_3, sigma_1, sigma_2, sigma_3, zgama, pC1, pC2,
pC3i, pC2i, pC1i, Li, xb1, xb2, xb3, u1, u2, u3, regL1, regL2, regL3, mu1, mu2,
ln_sigma_1, ln_sigma_2, ln_sigma_3, count_i, start_i, stop_i, L1_temp, L2_temp,
L3_temp, L1_i, L2_i, L3_i;
    struct modelResults mm;
    gama = pvUnpack(p,"gama"); mu = pvUnpack(p,"mu");
    beta_1 = pvUnpack(p,"beta_1");
    beta_2 = pvUnpack(p,"beta_2");
    beta_3 = pvUnpack(p,"beta_3");
    ln_sigma_1 = pvUnpack(p,"ln_sigma_1");
    ln_sigma_2 = pvUnpack(p,"ln_sigma_2");
    ln_sigma_3 = pvUnpack(p,"ln_sigma_3");
    sigma_1 = exp(ln_sigma_1);
    sigma_2 = exp(ln_sigma_2);
    sigma_3 = exp(ln_sigma_3);
    /* regression EVs */
    xb1 = x_reg*beta_1;
    xb2 = xb1 + exp(x_reg*beta_2);
    xb3 = xb2 + exp(x_reg*beta_3);
    /* class probs */
    zgama = x_lc*gama;
    mu1 = mu[1];
```

```
mu2 = mu[2];
pC1i = cdfn(mu1-zgama);
pC2i = cdfn(mu2-zgama)-cdfn(mu1-zgama);
pC3i = (1 - cdfn(mu2-zgama));
u1 = y - xb1;
u2 = y - xb2;
u3 = y - xb3;
regL1 = (1/sigma_1) .* pdfn(u1 ./ sigma_1);
regL2 = (1/sigma_2) .* pdfn(u2 ./ sigma_2);
regL3 = (1/sigma_3) .* pdfn(u3 ./ sigma_3);
if panel;
    start_i = 1;
    stop_i = count[1];
    count_i = count[1];
    L1_temp = regL1[start_i:stop_i,.];
    L1_temp = prodc(L1_temp);
    L2_temp = regL2[start_i:stop_i,.];
    L2_temp = prodc(L2_temp);
    L3_temp = regL3[start_i:stop_i,.];
    L3_temp = prodc(L3_temp);
    L1_i = L1_temp;
    L2_i = L2_temp;
    L3_i = L3_temp;
    for jrep (2,rows(count),1);
        count_i = count[jrep];
        start_i = stop_i + 1;
        stop_i = start_i + count_i - 1;
        L1_temp = regL1[start_i:stop_i,.];
        L1_temp = prodc(L1_temp);
        L1_i = L1_i|L1_temp;
        L2_temp = regL2[start_i:stop_i,.];
        L2_temp = prodc(L2_temp);
```

```
            L2_i = L2_i|L2_temp;
            L3_temp = regL3[start_i:stop_i,.];
            L3_temp = prodc(L3_temp);
            L3_i = L3_i|L3_temp;
        endfor;
        regL1 = L1_i;
        regL2 = L2_i;
        regL3 = L3_i;
    endif;
    Li = (pC1i .* regL1) + (pC2i .* regL2) + (pC3i .* regL3);
    mm.Function = ln(Li);
retp(mm);
endp;
```