Economics Letters 123 (2014) 1–5

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

A latent class model for obesity

William Greene[a], Mark N. Harris[b], Bruce Hollingsworth[c,*], Pushkar Maitra[d]

[a] Economics Department, Stern Business School, New York University, United States
[b] School of Economics and Finance, Curtin University, Australia
[c] Division of Health Research, Lancaster University, United Kingdom
[d] Department of Economics, Monash University, Australia

HIGHLIGHTS

- Obesity is modelled using a latent class model.
- A latent variable for class membership is defined as a function of observables and unobservables.
- Equations defining the class membership and observed outcomes are allowed to be correlated.
- There are significant correlations between these equations.
- The model can easily be applied to more classes and/or to models other than OP.

ARTICLE INFO

Article history:
Received 17 May 2012
Received in revised form
16 July 2013
Accepted 6 January 2014
Available online 20 January 2014

Keywords:
Latent class
Class membership
Obesity

ABSTRACT

We extend the discrete data latent class literature by explicitly defining a latent variable for class membership as a function of both observables and unobservables, thereby allowing the equations defining the class membership and observed outcomes to be correlated. The procedure is then applied to modelling observed obesity outcomes, based upon an underlying ordered probit equation.

1. Introduction and background

Latent class models are increasingly popular across both the physical and social sciences. With regard to economics, their use is particularly widespread in the health economics literature (for example see Deb and Trivedi, 2002, Bago D'Uva, 2005a,b). The approach involves probabilistically splitting the population into a set of *unobserved* homogeneous segments; within each class an appropriate econometric model applies. This yields a parsimonious way of introducing heterogeneity into a model. *Ex post*, it is then possible to assign individuals into their most likely class, typically defined by the outcome variable in each class.

In the latent class literature however, there is an implicit assumption that any unobservables driving class membership and those in the main econometric model are independent. The refinement here is to explicitly specify a latent variable for class membership, as a function of both observables and unobservables, and via the latter, explicitly model this aspect of the correlation between class membership and observed outcomes. This framework bears some resemblance to the switching regressions model and the *mover/stayer* model (Greene, 2008). However, here, the individual is not observed to be in either particular state (the true type of the individual is unobserved); this has to be identified using data. We illustrate this by modelling discrete observations of female obesity levels.

2. Econometric framework

Given our dependent variable, a useful starting point is the ordered probit (OP) model for the $j = 1, \ldots, J$ outcomes

$$y^* = z'\gamma + u$$
$$y = j \quad \text{if } \mu_{j-1} < y^* \leq \mu_j,\ j = 1, \ldots, J,\ \mu_0 = -\infty,\ \mu_J = +\infty.$$

[*] Corresponding author.
E-mail addresses: wgreene@stern.nyu.edu (W. Greene), mark.harris@curtin.edu (M.N. Harris), b.hollingsworth@lancaster.ac.uk (B. Hollingsworth), pushkar.maitra@monash.edu (P. Maitra).

0165-1765/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.econlet.2014.01.004

With normally distributed disturbances (u), this implies

$$\Pr(y) = \left\{ \begin{array}{c} \Pr(y=1) = \Phi\left(\mu_1 - z'\gamma\right) \\ \Pr(y=j) = \Phi\left(\mu_j - z'\gamma\right) - \Phi\left(\mu_{j-1} - z'\gamma\right); \\ \text{for } 1 < j < J \\ \Pr(y=J) = \Phi\left(\mu_{J-1} - z'\gamma\right) \end{array} \right\} \tag{1}$$

where Φ is the standard normal cumulative distribution function (*cdf*); μ are cut-off points; and z are covariates with unknown weights γ; and where y is the observed BMI range.

Herbert et al. (2006) find evidence that an obesity predisposing geno-type is present in 10% of individuals. Given that about 25% of our sample are categorized as obese, this supports a hypothesis that factors other than genetics impact upon the probability of being obese. Individuals in the population are broadly segmented into two classes: consider two individuals in the same observed obesity range; one may be there due to time-invariant, or fixed, characteristics (such as genetics) while the other because of lifestyle or behavioural choices.

Indeed, these two distinct sets of individuals are likely to have completely different reaction curves to alternative policy measures and therefore not taking this latent decomposition into account could result in biased estimates and erroneous policy conclusions. Let the latent variable c^* determine class membership, based on a function of a vector of observed characteristics x, with unknown weights β and a random disturbance term ε such that

$$c^* = x'\beta + \varepsilon. \tag{2}$$

Under normality, the probability of an individual belonging to class 1 (and one minus this for class 0) is given by

$$\Pr(c=1|\mathbf{x}) = \Pr\left(c^* > 0|\mathbf{x}\right) = \Phi\left(x'\beta\right).$$

$$\Pr(c=1|\mathbf{x}) = \Pr\left(c^* > 0|\mathbf{x}\right) = \Phi\left(x'\beta\right).$$

$$\Pr(c=1|\mathbf{x}) = \Pr\left(c^* > 0|\mathbf{x}\right) = \Phi\left(\mathbf{x'\boldsymbol{\beta}}\right).$$

Note that neither c^* nor c, is observed. The latent class framework implies that *conditional* on being in class 0 or 1, outcomes are determined by the relevant OP model: that is, we have a different OP equation for each class. The overall probability of an outcome is simply the sum of those from the two latent classes, such that

$$\begin{aligned} \Pr(y=j|\mathbf{x},\mathbf{z}) &= \Pr(c=0|\mathbf{x})\Pr(y=j|\mathbf{z},c=0) \\ &\quad + \Pr(c=1|\mathbf{x})\Pr(y=j|\mathbf{z},c=1). \end{aligned}$$

For those belonging to class 0 we have

$$\Pr = \left\{ \begin{array}{l} \Pr(y=1,c=0|\mathbf{x},\mathbf{z}) \\ \quad = \left(1-\Phi\left(x'\beta\right)\right)\left[\Phi\left(\mu_{0,1} - z'\gamma_0\right)\right] \\ \Pr(y=j,c=0|\mathbf{x},\mathbf{z}) \\ \quad = \left(1-\Phi\left(x'\beta\right)\right)\left[\Phi\left(\mu_{0,j} - z'\gamma_0\right)\right. \\ \quad\quad \left. -\Phi\left(\mu_{0,j-1} - z'\gamma_0\right)\right]; \quad 1 < j < J \\ \Pr(y=J,c=0|\mathbf{x},\mathbf{z}) \\ \quad = \left(1-\Phi\left(x'\beta\right)\right)\left[1-\Phi\left(\mu_{0,J-1} - z'\gamma_0\right)\right]. \end{array} \right. \tag{3}$$

We expand the usual specification by allowing ε and u to be freely correlated, with respective correlation coefficients ρ_0 and ρ_1. The respective probabilities are now defined by a bivariate standard normal distribution. Therefore, for membership in class 1 ($c = 1$), for example, the joint probabilities for the class membership and the obesity outcome are given by

$$\begin{aligned} &\Pr(y=j,c=1) \\ &= \left\{ \begin{array}{l} \Pr(y=1,c=1|\mathbf{x},\mathbf{z}) = \Phi_2\left(x'\beta, \mu_{1,1} - z'\gamma_1; \rho_1\right) \\ \Pr(y=j,c=1|\mathbf{x},\mathbf{z}) = \Phi_2\left(x'\beta, \mu_{1,j} - z'\gamma_1; \rho_1\right) \\ \quad -\Phi_2\left(x'\beta, \mu_{1,j-1} - z'\gamma_1; \rho_1\right); \quad 1 < j < J \\ \Pr(y=j,c=1|\mathbf{x},\mathbf{z}) = \Phi_2\left(x'\beta, z'\gamma_1 - \mu_{1,J-1}; \rho_1\right) \end{array} \right. \end{aligned} \tag{4}$$

where $\Phi_2(.,.;\rho)$ denotes the cumulative distribution function of the standardized bivariate normal distribution. We note, the specification of the correlation between the unobservables in the equations adds a dimension to the familiar latent class model. The class memberships and the observed outcomes are jointly determined by both the observables and the unobservables now added to the model.

The log-likelihood function for the observed data for a random sample of N individuals is constructed under the constraint that c is unobserved. Thus, the contribution to the log-likelihood for individual i is

$$\begin{aligned} \log L_i(\theta) &= \log \text{Prob}\left(y_i = j|x_i, z_i\right) \\ &= \log\left(\text{Prob}\left(y_i = j|z_i, c_i = 0\right)\text{Prob}\left(c_i = 0|x_i\right)\right. \\ &\quad \left. + \text{Prob}\left(y_i = j|z_i, c_i = 1\right)\text{Prob}\left(c_i = 1|x_i\right)\right) \\ &= \log\left(\text{Prob}\left(y_i = j, c_i = 0|z_i, x_i\right)/\text{Prob}\left(c_i = 0|x_i\right)\right. \\ &\quad \times \text{Prob}\left(c_i = 0|x_i\right) \\ &\quad + \text{Prob}\left(y_i = j, c_i = 1|z_i, x_i\right)/\text{Prob}\left(c_i = 1|x_i\right) \\ &\quad \left. \times \text{Prob}\left(c_i = 1|x_i\right)\right). \end{aligned} \tag{5}$$

The resulting contribution to the log likelihood is the sum of the logs of the joint probabilities:

$$\begin{aligned} \log L_i(\theta) &= \log \text{Prob}(y_i = j|x_i, z_i) \\ &= \log\left(\text{Prob}(y_i = j, c = 0|x_i, z_i)\right. \\ &\quad \left. + \text{Prob}(y_i = j, c = 1|x_i, z_i)\right). \end{aligned} \tag{6}$$

The log-likelihood for the sample is obtained by summing the terms in (6) over the individuals in the sample. Combining terms for the OP model

$$\begin{aligned} \log L_i(\theta) &= \sum_{i=1}^{N} \log \text{Prob}(y_i = j|x_i, z_i) \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{J} h_{ij}\left(\text{Prob}(y_i = j, c_i = 0|z_i, x_i)\right. \\ &\quad \left. + \text{Prob}(y_i = j, c_i = 1|z_i, x_i)\right) \end{aligned} \tag{7}$$

where h_{ij} is the usual indicator function. Tests of $\rho_c = 0$ are tests of independence of the respective error terms.

As a further refinement to the basic OP specification, we also allow for the fact that (in our obesity example) strict adherence to the World Health Organisation (WHO) defined boundaries may be too strict: athletes may have relatively high BMI levels due to a high percentage of muscle mass, rather than fat, for example. To account for this we adopt a generalized OP variant (Pudney and Shields, 2000), where the boundary parameters are functions of observed personal characteristics. To aid in the identification, and to ensure the proper ordering of the boundary parameters, they are specified as

$$\begin{aligned} \mu_{ci1} &= \alpha_{c1} + \exp\left(w_i'\delta_c\right) \\ \mu_{ci2} &= \mu_{ci1} + \exp\left(\alpha_{c2} + w_i'\delta_c\right) \\ \mu_{ci3} &= \mu_{ci2} + \exp\left(\alpha_{c2} + w_i'\delta_c\right) \\ &\vdots \end{aligned} \tag{8}$$

where the w are variables (excluding a constant term) that affect the position of the boundary parameters with unknown weights δ.

3. Data and variable selection

Using the US National Health Interview Survey (2007), and focusing on females due to space constraints,[1] we have a sample size

[1] Full results, including those for simpler nested sub-models, can be found in the Working Paper version at http://ideas.repec.org/p/ste/nystbu/08-18.html.

Table 1
Descriptive statistics. Female sample.

Description	Mean	*Variable inclusion*		
		Splitting equation (x)	OP equations (z)	Boundary equations (w)
Age/10 (scaled for convergence)	4.7354 (1.8085)		×	×
Age 10 squared	25.6946 (18.5138)		×	×
Duration of strength (weight training) exercise	0.7131 (2.1068)			×
Born in the US	0.8086 (0.3934)	×		
Born in South America	0.1210 (0.3262)	×		
Born in Europe or Russia	0.0195 (0.1382)	×		
Hispanic	0.1845 (0.3879)	×		
White	0.5789 (0.4938)	×		
Black	0.1780 (0.3825)	×		
Born between 1925 and 1942	0.1657 (0.3718)	×		
Born between 1943 and 1953	0.1572 (0.3640)	×		
Born between 1954 and 1965	0.2225 (0.4160)	×		
Born between 1966 and 1980	0.2779 (0.4480)	×		
Born between 1981 and 1995	0.1425 (0.3496)	×		
Married	0.4662 (0.4989)		×	
Income category	0.9682 (0.6465)		×	
Square of income category	1.3554 (1.5155)		×	
Years of schooling	14.5292 (3.4713)		×	
Own house	0.6121 (0.4873)		×	
Conducted moderate exercise in the last week	0.3171 (0.4654)		×	
Number of times vigorous exercise undertaken in the last week	1.2445 (2.8057)		×	
Normal weight	0.4294			
Overweight	0.3001			
Obese	0.2212			
Morbidly obese	0.0494			
Sample size	11 244			

of 11,244. Four WHO BMI categories are considered[2]: 43% are normal weight ($\text{BMI} \in (18.5, 25)$); 30% are overweight ($\text{BMI} \in (25, 30)$); 22% are obese ($\text{BMI} \in (30, 40)$); and 5% are morbidly obese ($\text{BMI} > 40$). While using this kind of ordinal measure of BMI does not use all available information, it has two distinct advantages. First, height and weight of individuals are potentially sensitive personal issues such that there is likely to be mis-reporting (in addition to recall bias and/or imperfect knowledge) of true height and weight levels resulting in measurement error in the (self-reported) BMI numerical values (see, for example, Gorber et al., 2007). It is not clear what the direction of this measurement error is. However, one can assume without significant loss of generality, that while the true BMI may not always be correctly "measured" (when self-reported, as is typically the case), the BMI *category* is likely to be correct. While this is more likely to be true within each category, the potential problem arising at the extremes in the form of mis-categorization is also taken into account in our analysis, since we allow the boundary parameters to vary with observed characteristics. The second advantage of using ordinal BMI levels is that policy makers are arguably more interested in movement across these categories, rather than marginal changes within them. For the purpose of this paper we have four categories: normal weight; overweight; obese; and morbidly obese.

Table 1 presents the sample averages. The average woman in the sample is around 47 years old, likely to be White (58%), born in the US (81%), born between 1954 and 1980 (50%), unmarried (54%), likely to own a house (61%) and having some college education.

Here we choose latent class covariates akin to proxies for an individual's 'fixed effect' (Greene, 2008): where the individual was born; whether White, Black, Hispanic, or 'other'; and a set of broad time cohort dummies.[3] Following the literature, the set of explanatory variables included in z are time-varying variables, which typically represent the lifestyle choices of the individual. Finally, variables included in the boundary parameters w include variables

[2] We drop underweight women ($\text{BMI} < 18.5$).

[3] Our approach is flexible enough to accommodate various forms of this, including null vectors in $\mathbf{x}$ for example.

Table 2
Parameter estimates.

Panel A: Splitting function parameters		
Constant	−0.72*	
	(0.37)	
Born in US	0.59***	
	(0.18)	
Born in South America	0.21	
	(0.16)	
Born in Europe	0.34*	
	(0.20)	
Hispanic	0.56***	
	(0.18)	
White	0.29**	
	(0.13)	
Black	0.83***	
	(0.21)	
Born between 1925 and 1942	0.41***	
	(0.13)	
Born between 1943 and 1953	0.56***	
	(0.20)	
Born between 1954 and 1965	0.22	
	(0.20)	
Born between 1966 and 1980	0.05	
	(0.22)	
Born between 1981 and 1995	−0.46*	
	(0.25)	
Panel B: OP parameters	Class 0 (Inherently non-obese)	Class 1 (Inherently obese)
Age/10	1.05	0.85
	(4.59)	(2.82)
$(\text{Age}/10)^2$	−0.02	−0.44
	(4.50)	(2.48)
Married	0.39***	−0.07
	(0.13)	(0.06)
Income category	0.30	0.05
	(0.32)	(0.15)
$(\text{Income category})^2$	−0.20	−0.06
	(0.14)	(0.07)
Years of schooling	−0.05***	−0.01*
	(0.02)	(0.01)
Own home	−0.11	−0.08
	(0.11)	(0.05)
Conducted moderate exercise in the last week	0.17	−0.19***
	(0.15)	(0.07)
Number of times vigorous exercise undertaken in the last week	−0.04	−0.01
	(0.03)	(0.01)
Panel C: Boundary parameters	Class 0 (Inherently non-obese)	Class 1 (Inherently obese)
μ_0	−0.64	−1.84***
	(0.62)	(0.62)
μ_1	−0.69	−0.19
	(0.46)	(0.35)
μ_2	0.29	0.08
	(0.43)	(0.36)
Age/10	−0.01	−0.58
	(2.21)	(1.14)
$(\text{Age}/10)^2$	1.13	1.01
	(2.39)	(1.04)
Duration of strength (weight training) exercise	0.76**	0.03
	(0.32)	(0.05)
Correlation	−0.72**	−0.66***
	(0.29)	(0.14)
Average outcome probabilities		
Normal weight	0.6141	0.3093
Overweight	0.2295	0.3516
Obese	0.1522	0.2871
Morbidly obese	0.0042	0.0520
Log likelihood	−8370.3054	

Standard errors in parentheses.

* Significance: 10%.

** Significance: 5%.

*** Significance: 1%.

that can potentially cause the boundaries to shift at the margin (here taken to be the number of times the respondent weight/strength trains per week, and a quadratic in age). The list of variables included in x, z and w is summarized in Table 1 as well.

4. Results

In such a bivariate latent class model, it is not obvious how to compute the posterior class probabilities independently from the choice probabilities. Indeed, it is in this way that classes are usually *labelled* (Bago d'Uva et al., 2009). However, it is possible to compute (post-estimation), for each individual, the probabilities of them being in each BMI-category by class, using the expressions in Eqs. (4) and (8). Averaging the posterior class probabilities over individuals produces the overall average outcome probabilities. We find in class 0 that the probabilities are skewed away from being in either the overweight, obese, or morbidly obese categories; respective probabilities are 0.2295, 0.1522 and 0.0042. Thus we label this the *inherently non-obese* class. Compare this to 0.3516, 0.2871 and 0.0520 respectively, the probabilities we find in class 1 (consequently, the *inherently obese* class). Additionally both ρ_0 and ρ_1 are highly statistically significant, indicating significant correlations between the unobservables in the two equations driving both class and observed BMI outcome.

The regression results are presented in Table 2. With regard to the latent class equation (Panel A), it is primarily determined by the country of birth, race and a set of birth cohort variables. The OP estimates (Panel B) show that irrespective of class, given the other factors, age, income and wealth do not appear to affect BMI levels. In the inherently non-obese category, increased educational attainment is negatively associated with the probability of being morbidly obese—the partial effects (available upon request) indicate that for an inherently non-obese female an additional year of schooling is associated with a 0.9% point increase in the probability of being of normal weight, and a 0.4% and 0.5% point reduction in the probability of being overweight or obese. The results are qualitatively similar for females in the inherently obese category (the magnitude is smaller). An increase in the duration of exercise significantly increases the probability (by 4.9% points) that a woman is of normal weight for inherently obese females; matched by a 5.2% point reduction in the probability that an inherently obese woman is in the obese or morbidly obese category.

Finally, turning to the boundary equations, (Panel C), only the frequency of weight training seems to have a statistically significant effect. But this suggests that for females in the inherently non-obese category, strict interpretation of the WHO boundaries may be inappropriate for some individuals.

5. Conclusions

This paper extends the finite mixture/latent class model literature by explicitly defining a latent variable for class membership as a function of both observables and unobservables, thereby allowing the equations defining the class membership and observed outcomes to be correlated. The procedure was illustrated with an application to an OP model with two classes. Indeed, the results show that there are significant correlations between these equations. With obvious generalizations, the model can easily be applied to more classes and/or to models other than OP.

References

Bago D'Uva, T., 2005a. Latent class models for utilisation of health care. Health Econ. 15 (4), 329–343.

Bago D'Uva, T., 2005b. Latent class models for utilisation of primary care: evidence from a British panel. Health Econ. 14 (9), 873–892.

Bago d'Uva, T., Jones, A., et al., 2009. Measurement of horizontal inequity in health care utilisation using European panel data. J. Health Econ. 28, 280–289.

Deb, P., Trivedi, P., 2002. The structure of demand for health care: latent class versus two-part models. J. Health Econ. 21 (4), 601–625.

Gorber, S.C., Tremblay, M., et al., 2007. A comparison of direct versus self-report measures for assessing height, weight and body mass index: a systematic review. Obes. Rev. 8, 307–326.

Greene, W., 2008. Econometric Analysis. Prentice Hall, New York.

Herbert, A., Gerry, N., et al., 2006. A common genetic variant is associated with adult and childhood obesity. Science 312, 279–283.

Pudney, S., Shields, M., 2000. Gender, race, pay and promotion in the British nursing profession: estimation of a generalised ordered probit model. J. Appl. Econometrics 15, 367–399.