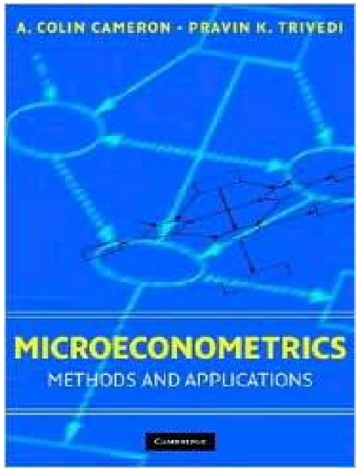
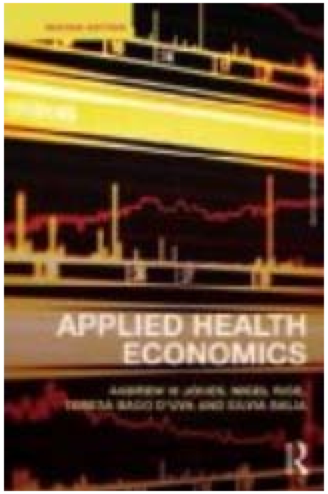
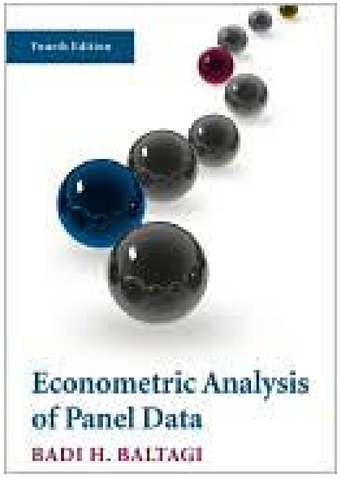
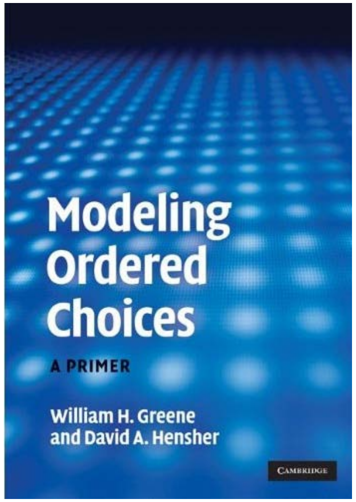
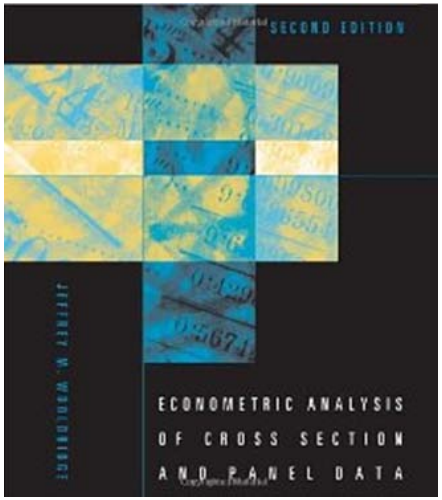


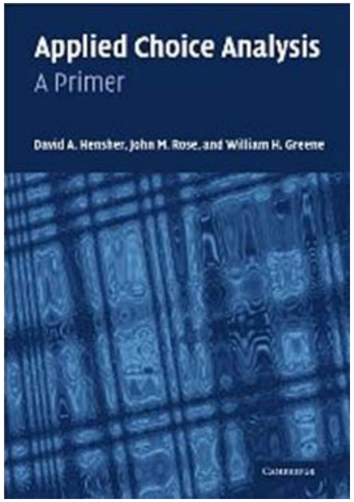




9. Heterogeneity: Latent Class Models







Science

AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

Science Magazine

AAAS

NEWS | SCIENCE JOURNALS | CAREERS | BLOGS & COMMUNITIES

Science The World's Leading Journal of Original Scientific Research, Global News, and Comme

Science Home | Current Issue | Previous Issues | *Science* Express | *Science* Products | My *Science* | Abou

Published Online April 12 2007
Science 11 May 2007:
Vol. 316 no. 5826 pp. 889-894
DOI: 10.1126/science.1141634

‹ Prev | Table of Contents | N

REPORT

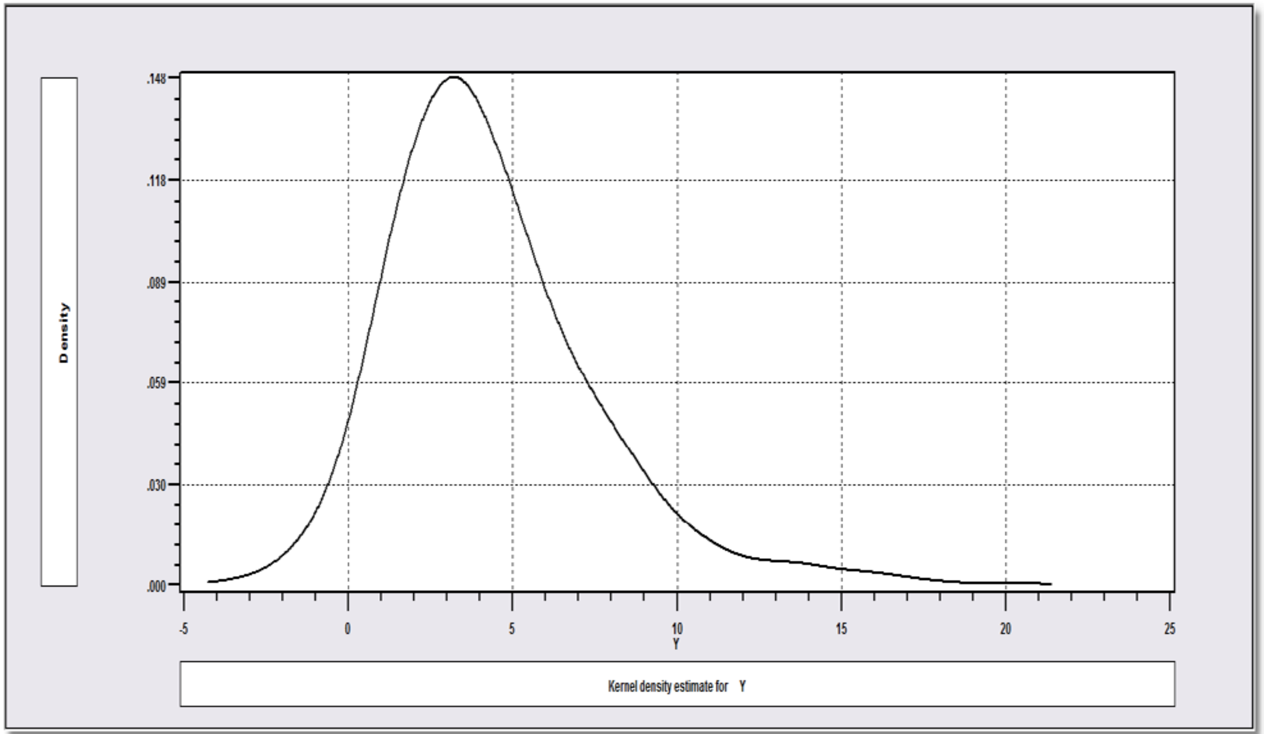
A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity

Latent Classes

- A population contains a mixture of individuals of different types (classes)
- Common form of the data generating mechanism within the classes
- Observed outcome y is governed by the **common process** $F(y|\mathbf{x},\theta_j)$
- Classes are distinguished by the parameters, θ_j.

How Finite Mixture Models Work

Y
4.21855
1.20367
2.45719
0.470427
16.4708
0.428376
1.56961
5.93268
3.83085
4.10209
7.29334
14.278
9.12016
1.57473
5.19982
3.84372
-3.57989
2.32862
2.85411
5.23678
2.25915
3.22748
11.0248
4.31525
3.55592
7.30238
8.61563
1.31486
5.6779
11.3807



Density? Note significant mass below zero. Not a gamma or lognormal or any other familiar density.

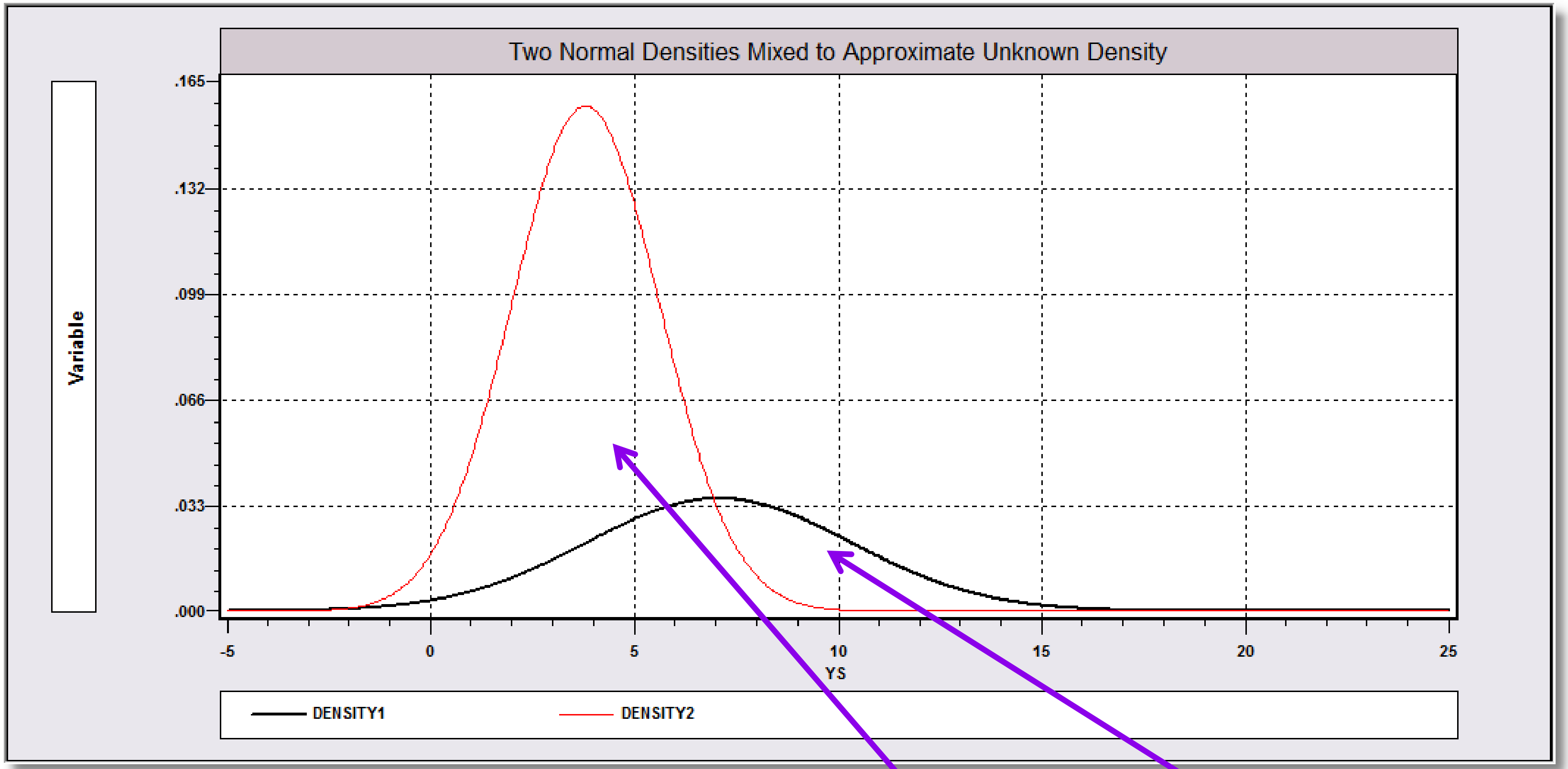
Find the ‘Best’ Fitting Mixture of Two Normal Densities

$$\text{LogL} = \sum_{i=1}^{1000} \log\left(\sum_{j=1}^{2} \pi_j \frac{1}{\sigma_j} \phi\left(\frac{y_i - \mu_j}{\sigma_j} \right) \right)$$

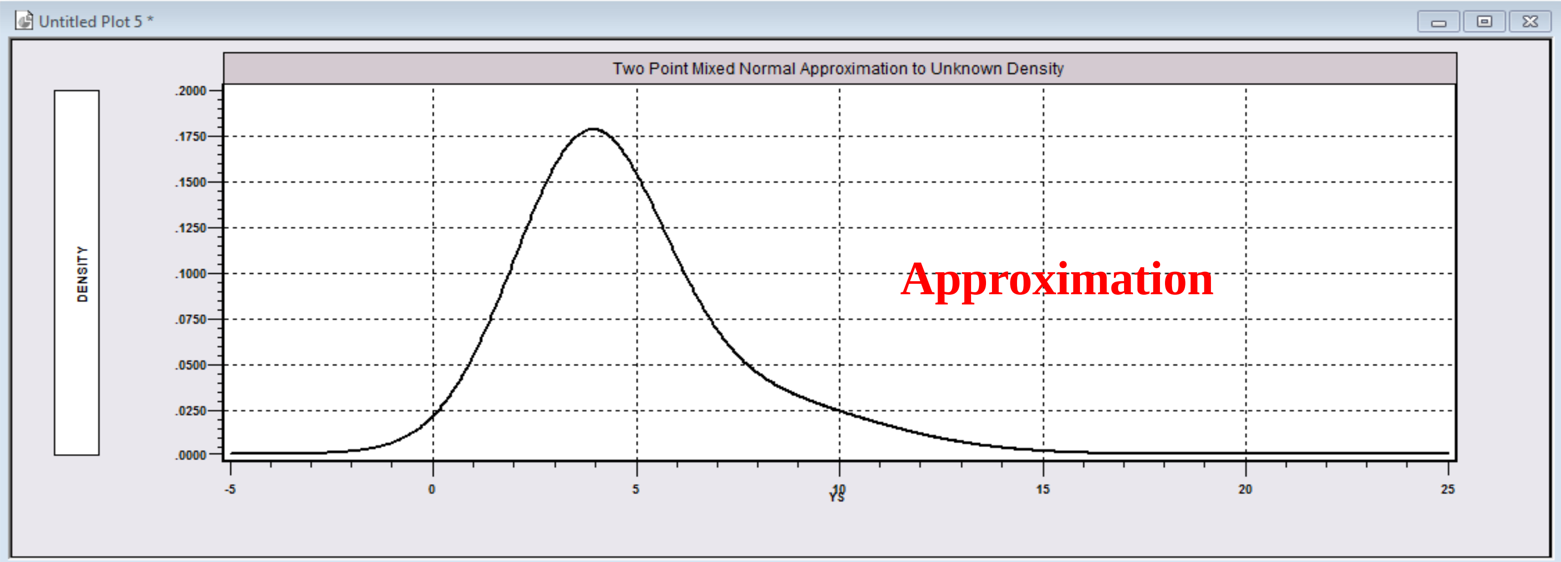
Maximum Likelihood Estimates

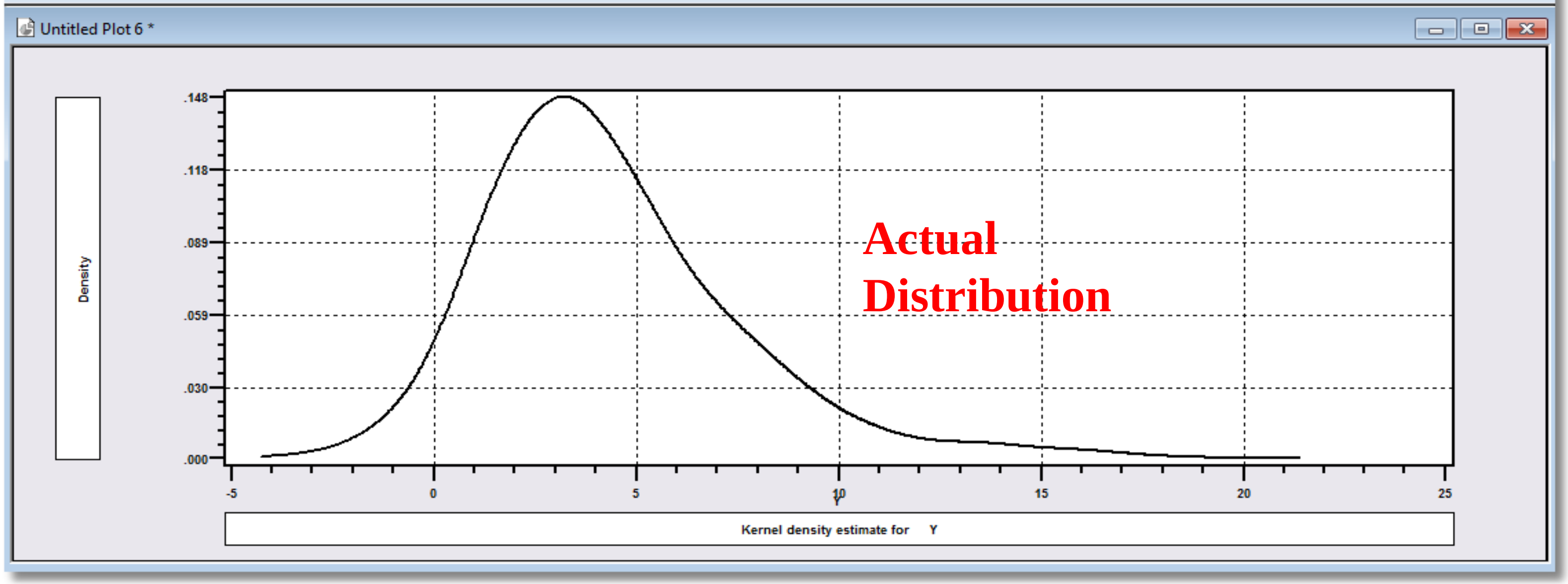
	Class 1 Estimate	Class 1 Std. Error	Class 2 Estimate	Class 2 Std. error
μ	7.05737	.77151	3.25966	.09824
σ	3.79628	.25395	1.81941	.10858
π	.28547	.05953	.71453	.05953

$$\hat{F}(y) = .28547\left[\frac{1}{3.79628}\phi\left(\frac{y-7.05737}{3.79628}\right)\right] + .71453\left[\frac{1}{1.81941}\phi\left(\frac{y-3.25966}{1.81941}\right)\right]$$



Mixing probabilities .715 and .285


Untitled Plot 5 *
Two Point Mixed Normal Approximation to Unknown Density
DENSITY
.2000
.1750
.1500
.1250
.1000
.0750
.0500
.0250
.0000
-5
0
5
10
15
20
25
YS
Approximation


Untitled Plot 6 *
Density
.148
.118
.089
.059
.030
.000
-5
0
5
10
15
20
25
Y
Kernel density estimate for Y
Actual
Distribution

A Practical Distinction

- **Finite Mixture (Discrete Mixture):**
  - Functional form strategy
  - Component densities have no meaning
  - Mixing probabilities have no meaning
  - There is no question of “class membership”
  - The number of classes is uninteresting – enough to get a good fit
- **Latent Class:**
  - Mixture of subpopulations
  - Component densities are believed to be definable “groups” (**Low Users** and **High Users** in Bago d’Uva and Jones application)
  - The classification problem is interesting – who is in which class?
  - Posterior probabilities, P(class|y,**x**) have meaning
  - Question of the number of classes has content in the context of the analysis

The Latent Class Model

(1) There are Q classes, unobservable to the analyst

(2) Class specific model: $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_q, \text{class} = q) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}_q)$

(3) Conditional class probabilities

Common multinomial logit form for prior class probabilities

$$P(\text{class}=q|\boldsymbol{\delta}) = \pi_{iq} = \frac{\exp(\boldsymbol{\delta}_{\mathbf{q}})}{\sum_{q=1}^{Q} \exp(\boldsymbol{\delta}_{\mathbf{q}})}, \quad \mathbf{Q}$$

$$\delta_q = \log(\pi_q / \pi_Q).$$

Log Likelihood for an LC Model

Conditional density for each observation is

$$P(y_{i,t} \mid \mathbf{x}_{i,t}, \text{class} = q) = f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)$$

Joint conditional density for T_i observations is

$$f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{X}_i, \boldsymbol{\beta}_q) = \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)$$

(T_i may be 1. This is not only a 'panel data' model.)

Maximize this for each class if the classes are known.

They aren't. Unconditional density for individual i is

$$f(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{X}_i) = \sum_{q=1}^{Q} \pi_q \left(\prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q) \right)$$

Log Likelihood

$$\boxed{\text{LogL}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_Q, \delta_1, \ldots, \delta_Q) = \sum_{i=1}^{N} \log \sum_{q=1}^{Q} \pi_q \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)}$$

Estimating Which Class

Prior class probability Prob[class=q]=π_q

Joint conditional density for T_i observations is

$$P(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{X}_i, \text{class} = q) = \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)$$

Joint density for data and class membership is the product

$$P(y_{i1}, y_{i2}, \ldots, y_{i,T_i}, \text{class} = q \mid \mathbf{X}_i) = \pi_q \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)$$

Posterior probability for class, given the data

$$P(\text{class} = q \mid y_{i1}, y_{i2}, \ldots, y_{i,T_i}, \mathbf{X}_i) = \frac{P(\mathbf{y_i}, \text{class} = q \mid \mathbf{X}_i)}{P(y_{i1}, y_{i2}, \ldots, y_{i,T_i} \mid \mathbf{X}_i)}$$

$$= \frac{P(\mathbf{y_i}, \text{class} = q \mid \mathbf{X}_i)}{\sum_{q=1}^{Q} P(\mathbf{y_i}, \text{class} = q \mid \mathbf{X}_i)}$$

Use Bayes Theorem to compute the **posterior (conditional) probability**

$$w(q \mid \mathbf{y_i}, \mathbf{X}_i) = P(\text{class} = j \mid \mathbf{y_i}, \mathbf{X}_i) = \frac{\pi_q \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)}{\sum_{q=1}^{Q} \pi_q \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{i,t}, \boldsymbol{\beta}_q)}$$

$$= w_{iq}$$

Best guess = the class with the largest posterior probability.

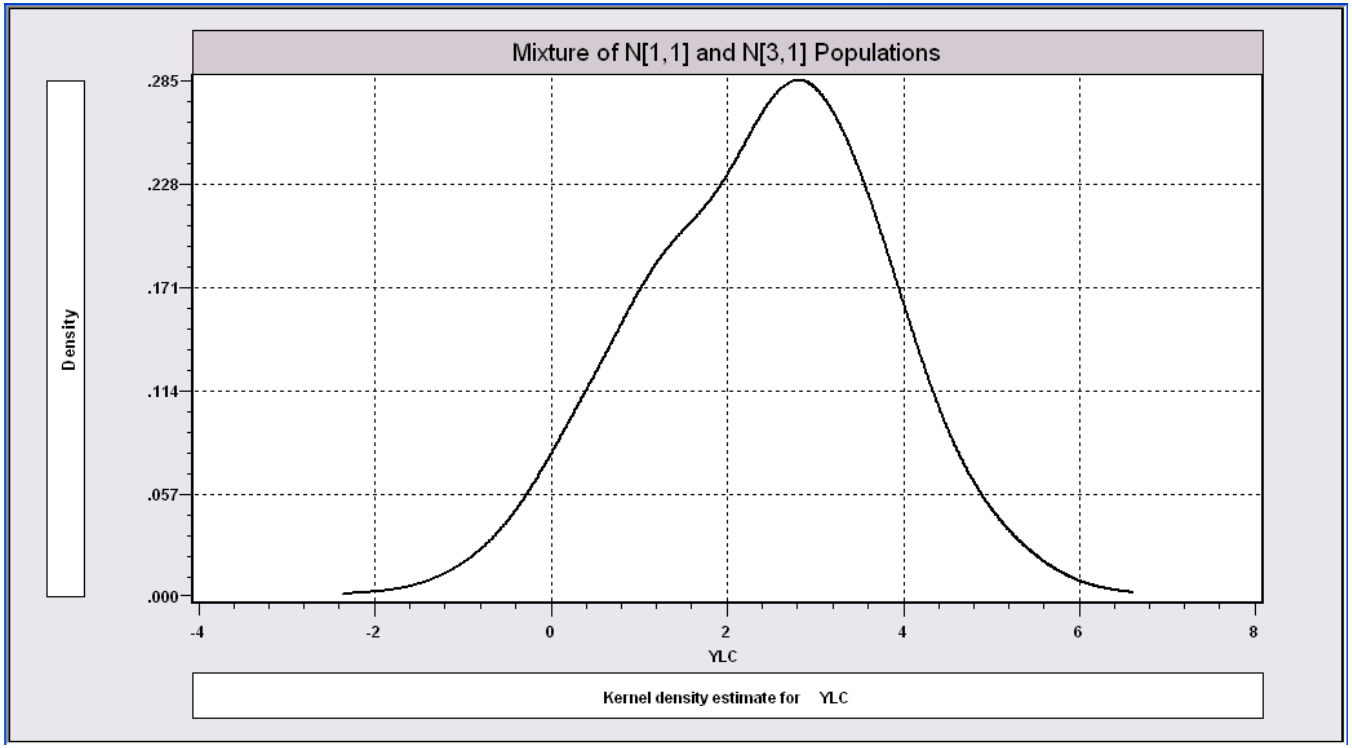
Posterior for Normal Mixture

$$\hat{w}(q \mid \text{data}_i) = \hat{w}(q \mid i) = \frac{\hat{\pi}_q \left[\prod_{t=1}^{T_i} \frac{1}{\hat{\sigma}_q} \phi\left(\frac{y_{it} - \hat{\mu}_q}{\hat{\sigma}_q} \right) \right]}{\sum_{q=1}^{Q} \hat{\pi}_q \left[\prod_{t=1}^{T_i} \frac{1}{\hat{\sigma}_q} \phi\left(\frac{y_{it} - \hat{\mu}_q}{\hat{\sigma}_q} \right) \right]}$$

Estimated Posterior Probabilities

G1	G2	YLCM
0.000380815	0.999619	0.749454
0.999998	2.4586e-006	6.00763
0.914497	0.0855032	3.3601
0.999983	1.65714e-005	5.53157
0.999998	2.1888e-006	6.03659
0.99946	0.000540485	4.65866
0.997469	0.00253128	4.27006
0.999999	1.40053e-006	6.14779
0.999863	0.000136592	5.00377
0.996985	0.00301525	4.2259
0.999983	1.72564e-005	5.52145
0.746364	0.253636	3.03334
0.860929	0.139071	3.22172
0.999932	6.81582e-005	5.1779
0.000111912	0.999888	0.434583
0.999933	6.66186e-005	5.18362
6.72545e-005	0.999933	0.303494
1	2.90138e-007	6.53927
0.985588	0.0144116	3.8288
0.279522	0.720478	2.51907

More Difficult When the Populations are Close Together



The Technique Still Works

```
----------------------------------------------------------------------
Latent Class / Panel LinearRg Model
Dependent variable                 YLC
Sample is  1 pds and   1000 individuals
LINEAR regression model
Model fit with  2 latent classes.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Model parameters for latent class 1
Constant|    2.93611***       .15813       18.568   .0000
   Sigma|    1.00326***       .07370       13.613   .0000
        |Model parameters for latent class 2
Constant|     .90156***       .28767        3.134   .0017
   Sigma|     .86951***       .10808        8.045   .0000
        |Estimated prior probabilities for class membership
Class1Pr|     .73447***       .09076        8.092   .0000
Class2Pr|     .26553***       .09076        2.926   .0034
--------+-------------------------------------------------------------
```

'Estimating' $\boldsymbol{\beta}_i$

(1) Use $\hat{\boldsymbol{\beta}}_j$ from the class with the largest estimated probability

(2) Probabilistic - in the same spirit as the 'posterior mean'

$$\hat{\boldsymbol{\beta}}_i = \sum_{q=1}^{Q} \text{Posterior Prob[class=q|data}_i\text{]}\,\hat{\boldsymbol{\beta}}_q$$

$$= \sum_{q=1}^{Q} \hat{w}_{iq} \hat{\boldsymbol{\beta}}_q$$

Note : This estimates $E[\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{X}_i]$, not $\boldsymbol{\beta}_i$ itself.

How Many Classes?

(1) Q is not a 'parameter' - can't 'estimate' Q with π and $\boldsymbol{\beta}$

(2) Can't 'test' down or 'up' to Q by comparing log likelihoods. Degrees of freedom for Q+1 vs. Q classes is not well defined.

(3) Use AKAIKE IC; AIC = $-2 \times \log L + 2\#\text{Parameters}$.

For our mixture of normals problem,

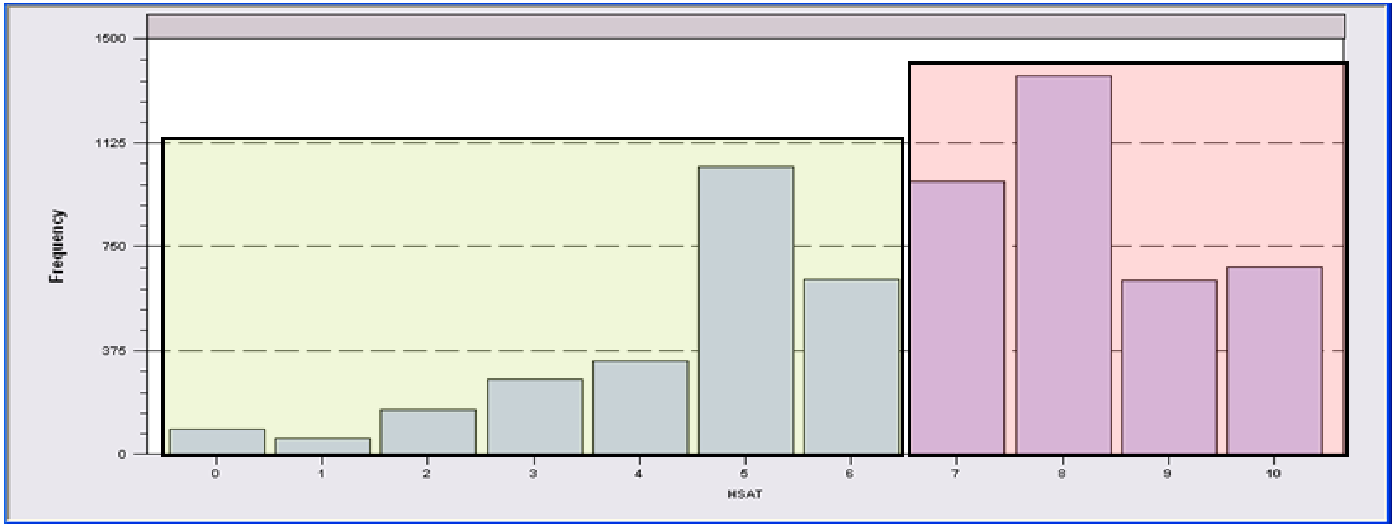
$AIC_1 = 10827.88$

$AIC_2 = 9954.268 \Leftarrow\!=\!=$

$AIC_3 = 9958.756$

LCM for Health Status

- Self Assessed Health Status = 0,1,…,10
- Recoded: Healthy = HSAT > 6
- Using only groups observed T=7 times; N=887
- Prob = Φ(Age,Educ,Income,Married,Kids)
- 2, 3 classes



Too Many Classes

```
----------------------------------------------------------------------
Latent Class / Panel Probit   Model
Dependent variable                HHEALTHY
Estimation based on N =    6209, K =  20
Unbalanced panel has     887 individuals
PROBIT (normal)  probability model
Model fit with  3 latent classes.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Model parameters for latent class 1
Constant|     .01265         .385900D+10      .000  1.0000
     AGE|     .16523         .138024D+09      .000  1.0000     44.3352
    EDUC|     .15327         .520918D+08      .000  1.0000     10.9409
  HHNINC|     .43195         .887276D+09      .000  1.0000      .34930
 MARRIED|     .06640         .153413D+09      .000  1.0000      .84539
  HHKIDS|     .17832         .152061D+09      .000  1.0000      .45482
        |Model parameters for latent class 2
Constant|     .32074           .29082        1.103   .2701
     AGE|    -.02690***        .00406       -6.622   .0000     44.3352
    EDUC|     .12215***        .01753        6.969   .0000     10.9409
  HHNINC|    -.03849           .17139        -.225   .8223      .34930
 MARRIED|     .20051***        .07749        2.588   .0097      .84539
  HHKIDS|     .05879           .06565         .895   .3705      .45482
        |Model parameters for latent class 3
Constant|     .00731           .26582         .027   .9781
     AGE|    -.03396***        .00446       -7.612   .0000     44.3352
    EDUC|     .02741*          .01466        1.869   .0616     10.9409
  HHNINC|     .73861***        .24133        3.061   .0022      .34930
 MARRIED|     .10671           .10520        1.014   .3104      .84539
  HHKIDS|     .16550**         .07838        2.111   .0347      .45482
        |Estimated prior probabilities for class membership
Class1Pr|     .12387***        .01676        7.390   .0000
Class2Pr|     .52530***        .02447       21.468   .0000
Class3Pr|     .35083***        .02268       15.466   .0000
--------+-------------------------------------------------------------
```

Two Class Model

```
----------------------------------------------------------------------
Latent Class / Panel Probit   Model
Dependent variable                 HEALTHY
Unbalanced panel has    887 individuals
PROBIT (normal)  probability model
Model fit with  2 latent classes.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Model parameters for latent class 1
Constant|     .61652**         .28620        2.154   .0312
     AGE|    -.02466***        .00401       -6.143   .0000      44.3352
    EDUC|     .11759***        .01852        6.351   .0000      10.9409
  HHNINC|     .10713           .20447         .524   .6003       .34930
 MARRIED|     .11705           .09574        1.223   .2215       .84539
  HHKIDS|     .04421           .07017         .630   .5287       .45482
        |Model parameters for latent class 2
Constant|     .18988           .31890         .595   .5516
     AGE|    -.03120***        .00464       -6.719   .0000      44.3352
    EDUC|     .02122           .01934        1.097   .2726      10.9409
  HHNINC|     .61039***        .19688        3.100   .0019       .34930
 MARRIED|     .06201           .10035         .618   .5367       .84539
  HHKIDS|     .19465**         .07936        2.453   .0142       .45482
        |Estimated prior probabilities for class membership
Class1Pr|     .56604***        .02487       22.763   .0000
Class2Pr|     .43396***        .02487       17.452   .0000
```

Partial Effects in LC Model

```
----------------------------------------------------------------------
Partial derivatives of expected val. with
respect to the vector of characteristics.
They are computed at the means of the Xs.
Conditional Mean at Sample Point    .6116
Scale Factor for Marginal Effects   .3832
B for latent class model is a wghted avrg.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]  Elasticity
--------+-------------------------------------------------------------
        |Two class latent class model
     AGE|    -.01054***       .00134       -7.860   .0000      -.76377
    EDUC|     .02904***       .00589        4.932   .0000       .51939
  HHNINC|     .12475**        .05598        2.228   .0259       .07124
 MARRIED|     .03570          .02991        1.194   .2326       .04934
  HHKIDS|     .04196**        .02075        2.022   .0432       .03120
--------+-------------------------------------------------------------
        |Pooled Probit Model
     AGE|    -.00846***       .00081      -10.429   .0000      -.63399
    EDUC|     .03219***       .00336        9.594   .0000       .59568
  HHNINC|     .16699***       .04253        3.927   .0001       .09865
        |Marginal effect for dummy variable is P|1 - P|0.
 MARRIED|     .02414          .01877        1.286   .1986       .03451
        |Marginal effect for dummy variable is P|1 - P|0.
  HHKIDS|     .06754***       .01483        4.555   .0000       .05195
--------+-------------------------------------------------------------
```

Conditional Means of Parameters

$$Est.\mathrm{E}[\boldsymbol{\beta} \,|\, \text{All information for individual i}] = \sum_{j=1}^{J} \hat{w}_{ij}\hat{\boldsymbol{\beta}}_j$$

using posterior (conditional) estimated class probabilities.

Matrix - BETA_I

[887, 6] Cell: 0.615759

	1	2	3	4	5	6
1	0.615759	-0.0246682	0.117419	0.108025	0.116947	0.0444756
2	0.604733	-0.0248373	0.114929	0.121032	0.115524	0.0483638
3	0.202056	-0.0310124	0.0239692	0.596029	0.0635754	0.190357
4	0.240936	-0.0304162	0.0327517	0.550166	0.0685913	0.176647
5	0.598638	-0.0249307	0.113552	0.128221	0.114738	0.0505129
6	0.447543	-0.0272478	0.0794216	0.306452	0.0952454	0.103792
7	0.616334	-0.0246594	0.117549	0.107347	0.117021	0.0442728
8	0.616489	-0.024657	0.117584	0.107164	0.117041	0.0442183
9	0.616015	-0.0246643	0.117477	0.107724	0.11698	0.0443856
10	0.194904	-0.0311221	0.0223537	0.604465	0.0626528	0.192879
11	0.369695	-0.0284416	0.0618367	0.398282	0.0852023	0.131244
12	0.198347	-0.0310693	0.0231313	0.600405	0.0630969	0.191665
13	0.440908	-0.0273496	0.0779228	0.31428	0.0943894	0.106132
14	0.430045	-0.0275161	0.075469	0.327094	0.0929879	0.109963
15	0.189896	-0.0311989	0.0212223	0.610373	0.0620066	0.194645
16	0.19017	-0.0311947	0.0212844	0.610049	0.062042	0.194548
17	0.190421	-0.0311908	0.0213409	0.609754	0.0620743	0.19446
18	0.189918	-0.0311985	0.0212274	0.610347	0.0620095	0.194637
19	0.414608	-0.0277529	0.0719819	0.345303	0.0909964	0.115406
20	0.194766	-0.0311242	0.0223225	0.604628	0.0626349	0.192928

An Extended Latent Class Model

Class probabilities relate to observable variables (usually demographic factors such as age and sex).

(1) There are Q classes, unobservable to the analyst

(2) Class specific model: $f(y_{it} \mid \mathbf{x}_{it}, \text{class} = q) = g(y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}_q)$

(3) Conditional class probabilities given some information, $\mathbf{z}_i$)

Common multinomial logit form for prior class probabilities

$$P(\text{class}=q|\mathbf{z}_i, \boldsymbol{\delta}) = \pi_{iq} = \frac{\exp(\mathbf{z}_i'\boldsymbol{\delta}_q)}{\sum_{q=1}^{Q}\exp(\mathbf{z}_i'\boldsymbol{\delta}_q)}, \quad \boldsymbol{\delta}_q = \mathbf{0}$$

Contents lists available at ScienceDirect

Journal of Health Economics

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

Health care utilisation in Europe: New evidence from the ECHP

Teresa Bago d'Uva [a,b,*], Andrew M. Jones [c]

[a] Department of Applied Economics (Room H13-09), Erasmus School of Economics, PB 1738, 3000 DR Rotterdam, The Netherlands
[b] Netspar, The Netherlands
[c] Department of Economics and Related Studies, University of York, York, YO10 5DD, United Kingdom

Most empirical applications of latent class models to health care utilisation take class membership probabilities as parameters $\pi_{ij} = \pi_j, j = 1, \ldots, C$ to be estimated along with $\theta_1, \ldots, \theta_C$ (e.g., Deb and Trivedi, 1997; Deb, 2001; Jiménez-Martín et al., 2002; Atella et al., 2004; Bago d'Uva, 2006). This is analogous to the hypothesis that individual heterogeneity is uncorrelated with the regressors in a random effects or random parameters specification. A more general approach is to parameterise the heterogeneity as a function of time invariant individual characteristics z_i, as in Mundlak (1978), thus accounting for the possible correlation between observed regressors and unobserved effects. This has been done in recent studies that consider continuous distributions for the individual effects, mostly by setting $z_i = \bar{x}_i$. To implement this approach in the case of the latent class model, class membership can be modelled as a multinomial logit (as in, e.g., Clark and Etilé, 2006; Clark et al., 2005; Bago d'Uva, 2005):

$$\pi_{ij} = \frac{\exp(z_i'\gamma_j)}{\sum_{g=1}^{C} \exp(z_i'\gamma_g)}, \quad j = 1, \ldots, C, \tag{3}$$

with $\gamma_C = 0$. This specification makes it possible to uncover the determinants of class membership (more commonly done by means of posterior analysis). The vectors of parameters $\theta_1, \ldots, \theta_C, \gamma_1, \ldots, \gamma_{C-1}$ are estimated jointly by maximum likelihood.

Health Satisfaction Model

Latent Class / Panel Probit Model
Dependent variable HEALTHY
Log likelihood function -3465.98697

Used mean AGE and FEMALE in class probability model

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Model parameters for latent class 1				
Constant	.60050**	.29187	2.057	.0396	
AGE	-.02002***	.00447	-4.477	.0000	44.3352
EDUC	.10597***	.01776	5.968	.0000	10.9409
HHNINC	.06355	.20751	.306	.7594	.34930
MARRIED	.07532	.10316	.730	.4653	.84539
HHKIDS	.02632	.07082	.372	.7102	.45482
	Model parameters for latent class 2				
Constant	.10508	.32937	.319	.7497	
AGE	-.02499***	.00514	-4.860	.0000	44.3352
EDUC	.00945	.01826	.518	.6046	10.9409
HHNINC	.59026***	.19137	3.084	.0020	.34930
MARRIED	-.00039	.09478	-.004	.9967	.84539
HHKIDS	.20652***	.07782	2.654	.0080	.45482
	Estimated prior probabilities for class membership				
ONE_1	1.43661***	.53679	2.676	.0074	(.56519)
AGEBAR_1	-.01897*	.01140	-1.664	.0960	
FEMALE_1	-.78809***	.15995	-4.927	.0000	
ONE_2	.000	......(Fixed Parameter)......			(.43481)
AGEBAR_2	.000	......(Fixed Parameter)......			
FEMALE_2	.000	......(Fixed Parameter)......			

The EM Algorithm

Latent Class is a '**missing data**' model

$d_{i,q} = 1$ if individual i is a member of class q

If $d_{i,q}$ were observed, the complete data log likelihood would be

$$\log L_c = \sum_{i=1}^{N} \log\left\{\sum_{q=1}^{Q} d_{i,q}\left[\prod_{t=1}^{T_i} f(y_{i,t} \mid data_{i,t}, class = q)\right]\right\}$$

(Only one of the Q terms would be nonzero.)

Expectation - Maximization algorithm has two steps

(1) Expectation Step: Form the 'Expected log likelihood' given the data and a prior guess of the parameters.

(2) Maximize the expected log likelihood to obtain a new guess for the model parameters.

(E.g., http://crow.ee.washington.edu/people/bulyko/papers/em.pdf)

Implementing EM for LC Models

Given initial guesses $\pi_q^0 = \pi_1^0, \pi_2^0, \ldots, \pi_Q^0$, $\boldsymbol{\beta}_q^0 = \boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2^0, \ldots, \boldsymbol{\beta}_Q^0$

E.g., use 1/Q for each π_q and the MLE of $\boldsymbol{\beta}$ from a one class model. (Must perturb each one slightly, as if all π_q are equal and all $\boldsymbol{\beta}_q$ are the same, the model will satisfy the FOC.)

(1) Compute $\hat{F}(q|i)$ = posterior class probabilities, using $\hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\delta}}^0$

Reestimate each $\boldsymbol{\beta}_q$ using a weighted log likelihood

Maximize wrt $\boldsymbol{\beta}_q$ $\sum_{i=1}^{N} \hat{F}_{iq} \sum_{t=1}^{T_i} \log f(y_{it} \mid \mathbf{x}_{it}\boldsymbol{\beta},\ {}_q)$

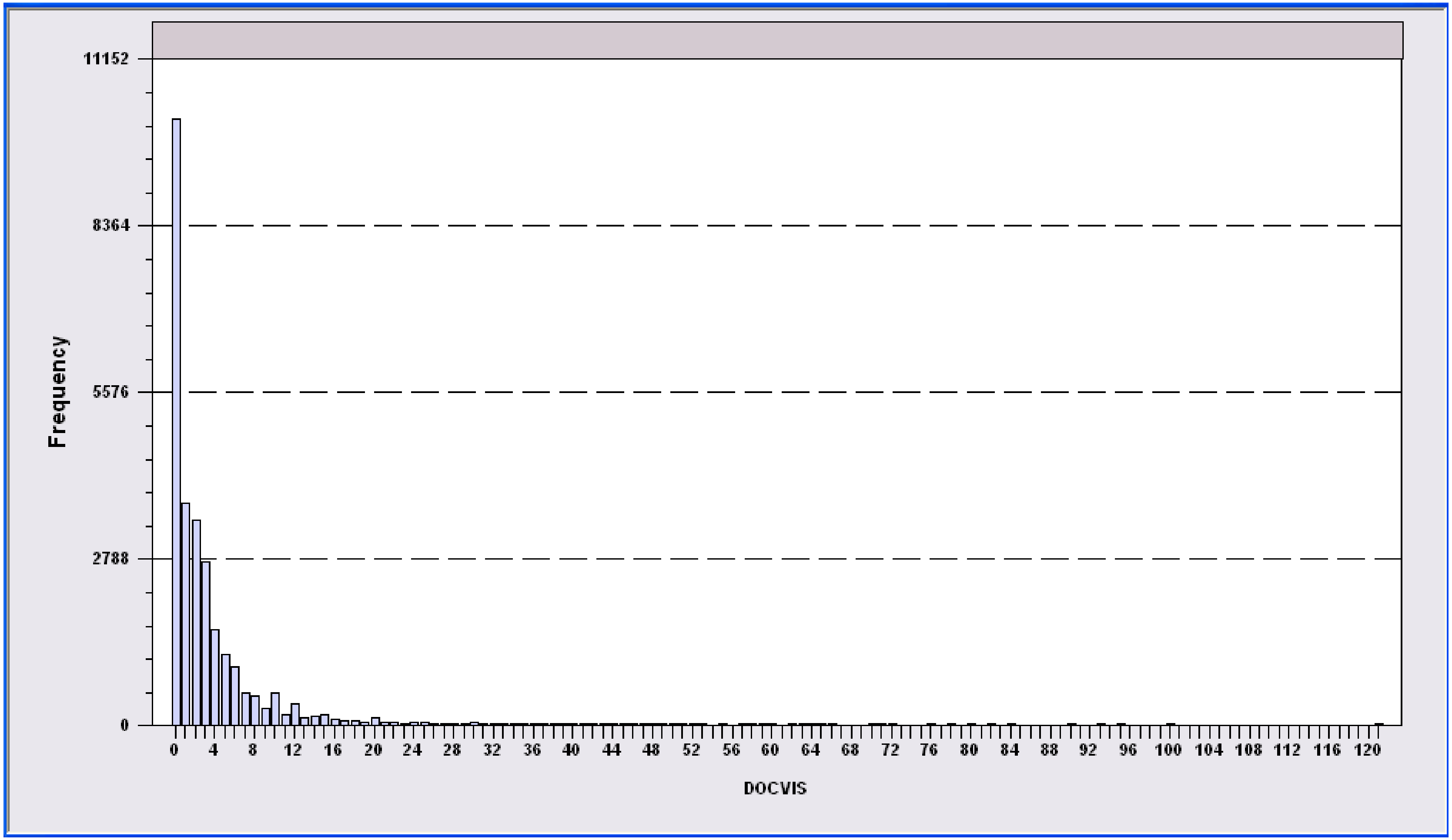
(2) Reestimate π_q by reestimating $\boldsymbol{\delta}$

$\hat{\pi}_q = (1/N)\Sigma_{i=1}^{N} \hat{F}(q|i)$ using old $\hat{\pi}$ and new $\boldsymbol{\beta}$

Now, return to step 1.

Iterate until convergence.

Zero Inflation?



Zero Inflation – ZIP Models

- Two regimes: (Recreation site visits)
  - Zero (with probability 1). (Never visit site)
  - Poisson with Pr(0) = exp[- $\boldsymbol{\beta}'\mathbf{x}_i$]. (Number of visits, including zero visits this season.)
- Unconditional:
  - Pr[0] = P(regime 0) + P(regime 1)*Pr[0|regime 1]
  - Pr[j | j >0] = P(regime 1)*Pr[j|regime 1]
- This is a “latent class model”

Hurdle Models

- Two decisions:
  - Whether or not to participate: y=0 or +.
  - If participate, how much. y|y>0
- One ‘regime’ – individual always makes both decisions.
- Implies different models for zeros and positive values
  - Prob(0) = 1 – F($\gamma'\mathbf{z}$), Prob(+) = F($\gamma'\mathbf{z}$)
  - Prob(y|+) = P(y)/[1 – P(0)]

Journal of Health Economics 28 (2009) 265–279

Contents lists available at ScienceDirect

Journal of Health Economics

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

ELSEVIER

Health care utilisation in Europe: New evidence from the ECHP

Teresa Bago d'Uva[a,b,*], Andrew M. Jones[c]

[a] Department of Applied Economics (Room H13-09), Erasmus School of Economics, PB 1738, 3000 DR Rotterdam, The Netherlands
[b] Netspar, The Netherlands
[c] Department of Economics and Related Studies, University of York, York, YO10 5DD, United Kingdom

Consider individuals i observed T_i times, where T_i can take values up to 5 and 6 for Finland and Austria, and values up to 7 for the remaining eight countries considered here. Denote the observations of the dependent variable over the panel as $y_i = [y_{i1}, \ldots, y_{iT_i}]$, where y_{it} represents the number of visits in year t. We assumed that each individual i belongs to a latent class $j, j = 1, \ldots, C$, and that individuals are heterogeneous across classes. The probability of belonging to class j is π_{ij}, where $0 < \pi_{ij} < 1$ and $\sum_{j=1}^{C} \pi_{ij} = 1$. Conditional on the class that individual i belongs to, the number of visits in a given year t, y_{it}, is distributed according to $f_j(y_{it}|x_{it}, \theta_j)$ and the θ_j are vectors of parameters specific to each class. Assuming independence, conditional on the latent class j, the joint density of y_{it} over the observed periods is obtained from the product of T_i independent densities $f_j(y_{it}|x_{it}, \theta_j)$. The unconditional (on the latent class) joint density of $y_i = [y_{i1}, \ldots, y_{iT_i}]$ derives from averaging out the individual unobserved heterogeneity represented by the latent classes:

$$g(y_i|x_i; \pi_{i1}, \ldots, \pi_C; \theta_1, \ldots, \theta_C) = \sum_{j=1}^{C} \pi_{ij} \prod_{t=1}^{T_i} f_j(y_{it}|x_{it}, \theta_j), \tag{1}$$

where x_i is a vector of covariates, including a constant, and θ_j are vectors of parameters.

Following Bago d'Uva (2006), the class-specific density of the number of visits in a given year, $f_j(y_{it}|x_{it}, \theta_j)$, is defined as in the standard hurdle model, using a negative binomial as the parent distribution in both stages. Formally, for each component $j, j = 1, \ldots, C$, the probability of zero visits and the probability of observing y_{it} visits, given $y_{it} > 0$, are given by

$$\begin{aligned} f_j(0|x_{it}; \theta_{j1}) &= P[Y_{it} = 0|x_{it}, \theta_{j1}] = (\lambda_{j1,it}^{1-k} + 1)^{-\lambda_{j1,it}^{k}} \\ f_j(y_{it}|y_{it} > 0, x_{it}; \theta_{j2}) &= \frac{\Gamma\left(y + (\lambda_{j2,it}^{k}/\alpha_j)\right)\left(\alpha_j \lambda_{j2,it}^{1-k} + 1\right)^{-(\lambda_{j2,it}^{k}/\alpha_j)} [1 + (\lambda_{j2,it}^{k-1}/\alpha_j)]^{-y_{i,t}}}{\Gamma(\lambda_{j2,it}^{k}/\alpha_j)\Gamma(y_{it} + 1)\left[1 - (\alpha_j \lambda_{j2,it}^{1-k} + 1)^{-(\lambda_{j2,it}^{k}/\alpha_j)}\right]}, \end{aligned} \tag{2}$$

where $\lambda_{j1,it} = \exp(x_{it}'\beta_{j1})$, $\lambda_{j2,it} = \exp(x_{it}'\beta_{j2})$, α_j are overdispersion parameters and k is an an arbitrary constant (most commonly set equal to 1 or 0, corresponding to the NegBin1 and NegBin2 models, respectively; we use the NegBin2 model). The vectors of parameters driving the probability of seeking care, β_{j1}, and the number of visits, given that this is positive, β_{j2}, are allowed to be different which means that the determinants of care may have different effects on the two stages of the

A Latent Class Hurdle NB2 Model

- Analysis of ECHP panel data (1994-2001)
- Two class Latent Class Model
  - Typical in health economics applications
- Hurdle model for physician visits
  - Poisson hurdle for participation and negative binomial intensity given participation
  - Contrast to a negative binomial model

Table 8. Estimated income coefficients and elasticities for GP and specialist visits—country-specific LC hurdle models.

Country		GPs				Specialists			
		Low users		High users		Low users		High users	
		Estimated coefficient	Estimated elasticity	Estimated coefficient	Estimated elasticity	Estimated coefficient	Estimated elasticity	Estimated coefficient	Estimated elasticity
Austria	*P*(Y>0)	−0.051 (−1.467)	−0.012	−0.109 (−0.872)	−0.005	**0.191** (3.743)	**0.110**	**0.211** (3.556)	**0.030**
	E(Y|Y>0)	0.012 (0.693)	0.009	**0.039** (2.167)	**0.035**	0.014 (0.210)	0.006	**0.105** (3.858)	**0.070**
Belgium	*P*(Y>0)	0.035 (1.002)	0.008	**0.292** (4.004)	**0.010**	0.054 (1.399)	0.036	0.079 (1.348)	0.014
	E(Y|Y>0)	**−0.052** (−3.125)	**−0.037**	**−0.055** (−4.030)	**−0.050**	−0.112 (−1.611)	−0.052	−0.049 (−1.920)	−0.035
Denmark	*P*(Y>0)	0.083 (1.746)	0.033	**0.261** (2.302)	**0.023**	0.053 (0.738)	0.045	0.079 (1.123)	0.034
	E(Y|Y>0)	0.042 (0.992)	0.021	−0.030 (−1.009)	−0.024	−0.053 (−0.434)	−0.022	−0.082 (−1.120)	−0.050
Finland	*P*(Y>0)	0.054 (1.358)	0.024	−0.030 (−0.263)	−0.003	**0.203** (3.525)	**0.155**	0.167 (1.909)	0.041
	E(Y|Y>0)	0.007 (0.237)	0.004	−0.048 (−1.706)	−0.037	**−0.229** (−2.985)	**−0.090**	0.025 (0.487)	0.014
Greece	*P*(Y>0)	0.012 (0.565)	0.006	0.015 (0.447)	0.004	**0.184** (7.641)	**0.128**	**0.148** (5.413)	**0.060**
	E(Y|Y>0)	−0.024 (−1.864)	−0.015	**0.026** (1.967)	**0.020**	0.017 (0.878)	0.010	**0.067** (4.192)	**0.055**
Ireland	*P*(Y>0)	**0.164** (4.754)	**0.064**	0.026 (0.339)	0.003	**0.172** (3.274)	**0.152**	**0.313** (4.367)	**0.144**
	E(Y|Y>0)	**−0.095** (−3.865)	**−0.057**	**−0.049** (−2.528)	**−0.043**	0.063 (0.738)	0.027	−0.091 (−1.838)	−0.057
Italy	*P*(Y>0)	−0.001 (−0.054)	0.000	**0.116** (3.766)	**0.011**	**0.136** (6.251)	**0.105**	**0.190** (7.918)	**0.063**
	E(Y|Y>0)	**−0.044** (−4.944)	−0.031	**−0.024** (−2.691)	**−0.021**	**−0.084** (−2.787)	**−0.035**	0.000 (−0.026)	0.000
The Netherlands	*P*(Y>0)	**0.082** (2.897)	**0.035**	0.094 (1.739)	0.009	**0.071** (2.085)	**0.055**	−0.055 (−1.084)	−0.016
	E(Y|Y>0)	−0.037 (−1.484)	−0.019	**−0.085** (−5.446)	**−0.068**	**−0.250** (−4.377)	**−0.129**	−0.008 (−0.299)	−0.006
Portugal	*P*(Y>0)	**0.223** (10.888)	**0.104**	**0.243** (8.070)	**0.036**	**0.252** (9.190)	**0.198**	**0.295** (9.454)	**0.099**
	E(Y|Y>0)	**0.027** (2.302)	**0.018**	0.001 (0.078)	0.001	**−0.087** (−3.292)	**−0.045**	**0.041** (2.340)	**0.028**
Spain	*P*(Y>0)	−0.015 (−0.997)	−0.006	0.037 (1.261)	0.005	**0.112** (5.680)	**0.080**	**0.138** (5.189)	**0.042**
	E(Y|Y>0)	**−0.053** (−4.401)	**−0.034**	**−0.025** (−2.324)	**−0.021**	**−0.070** (−2.460)	**−0.033**	0.017 (1.026)	0.012

Notes: *t*-statistics of coefficients in parentheses. Coefficients in bold are those significant at 5%. Elasticities are calculated for each individual and averaged over the sample. Elasticities in bold correspond to significant coefficients.

LC Poisson Regression for Doctor Visits

```
Latent Class / Panel Poisson  Model
Dependent variable              DOCVIS
Log likelihood function    -71088.40014
Restricted log likelihood -213471.86049
Chi squared [  20 d.f.]    284766.92070
Significance level              .00000
Unbalanced panel has   7293 individuals
POISSON regression model
Model fit with  3 latent classes.
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Model parameters for latent class 1					
Constant	3.17505***	.03833	82.84	.0000	3.09993	3.25017
AGE	.00734***	.00067	10.93	.0000	.00603	.00866
HSAT	-.16554***	.00271	-61.11	.0000	-.17085	-.16023
INCOME	-.49866***	.04988	-10.00	.0000	-.59642	-.40090
	Model parameters for latent class 2					
Constant	.22803***	.07453	3.06	.0022	.08195	.37410
AGE	.02479***	.00127	19.49	.0000	.02230	.02728
HSAT	-.23278***	.00454	-51.32	.0000	-.24167	-.22389
INCOME	.08285	.06470	1.28	.2003	-.04396	.20966
	Model parameters for latent class 3					
Constant	1.79613***	.03743	47.99	.0000	1.72278	1.86949
AGE	.01418***	.00063	22.48	.0000	.01294	.01542
HSAT	-.17543***	.00243	-72.17	.0000	-.18020	-.17067
INCOME	-.13569***	.03602	-3.77	.0002	-.20630	-.06509
	Estimated prior probabilities for class membership					
ONE_1	-1.17685***	.27028	-4.35	.0000	-1.70659	-.64710
FEMALE_1	.25804***	.09562	2.70	.0070	.07063	.44544
MARRIE_1	-.20921**	.09901	-2.11	.0346	-.40326	-.01516
EDUC_1	-.02913	.02113	-1.38	.1679	-.07054	.01228
ONE_2	.28202*	.16781	1.68	.0928	-.04688	.61093
FEMALE_2	-.71203***	.05922	-12.02	.0000	-.82809	-.59597
MARRIE_2	.02506	.06549	.38	.7020	-.10329	.15341
EDUC_2	.00599	.01290	.46	.6423	-.01929	.03127
ONE_3	0.0	.....(Fixed Parameter).....				
FEMALE_3	0.0	.....(Fixed Parameter).....				
MARRIE_3	0.0	.....(Fixed Parameter).....				
EDUC_3	0.0	.....(Fixed Parameter).....				

Prior class probabilities at data means for LCM variables

Class 1	Class 2	Class 3	Class 4	Class 5
.09543	.45860	.44596	.00000	.00000

Is the LCM Finding High and Low Users?

```
SETPANEL ; Group = ID ; Pds = TI $
CREATE   ; t = ndx(id,1) $ t = the within group index, 1,2,...,Ti
CREATE   ; edbar = Group Mean(educ,pds=ti) $
NAMELIST ; x=one,age,hsat,educ,married,hhkids $
POISSON  ; Lhs = docvis ; Rhs=x
         ; Lcm = female,edbar ; Par
         ; pts = 2
         ; panel$
CREATE   ; Prob1 = classp_i(id,1) ; Prob2 = 1-Prob1 $
CREATE   ; Class1= Prob1 > .5 ; Class2 = 1-Class1
         ; Class = class1 + 2*class2 $
DSTAT    ; For[class] ; Rhs = Docvis $
```

Is the LCM Finding High and Low Users? Apparently So.

```
--------------------------------------------------------------------------
Subsample analyzed for this command is CLASS    =      2

Descriptive Statistics for   1 variables
--------+-----------------------------------------------------------------
Variable|      Mean       Std.Dev.     Minimum      Maximum     Cases Missing
--------+-----------------------------------------------------------------
  DOCVIS|    1.567855    2.421592          0.0         44.0     19188       0
--------+-----------------------------------------------------------------
Subsample analyzed for this command is CLASS    =      1
Descriptive Statistics for   1 variables
--------+-----------------------------------------------------------------
Variable|      Mean       Std.Dev.     Minimum      Maximum     Cases Missing
--------+-----------------------------------------------------------------
  DOCVIS|    6.992996    8.614750          0.0        121.0      8138       0
--------+-----------------------------------------------------------------
```

Heckman and Singer's RE Model

- Random Effects Model
- Random Constants with Discrete Distribution

(1) There are Q classes, unobservable to the analyst

$\boxed{\text{(2) Class specific model: } f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}, \text{class}=q) = g(y_{it}, \mathbf{x}_{it}, \alpha_q, \boldsymbol{\beta})}$

(3) Conditional class probabilities π_q

Common multinomial logit form for prior class probabilities

to constrain all probabilities to (0,1) and ensure $\sum_{q=1}^{Q} \pi_q = 1$;

multinomial logit form for class probabilities;

$$P(\text{class}=q|\boldsymbol{\delta}) = \pi_q = \frac{\exp(\delta_q)}{\sum_{j=1}^{J} \exp(\delta_q)}, \ \delta_Q = 0$$

3 Class Heckman-Singer Form

Latent Class / Panel Poisson Model
Dependent variable DOCVIS
Log likelihood function -71303.62317
Restricted log likelihood -213471.86049
Chi squared [20 d.f.] 284336.47464
Significance level .00000

Log likelihood function -71088.40014

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Model parameters for latent class 1					
Constant	2.94456***	.02913	101.09	.0000	2.88747	3.00165
AGE	.01146***	.00053	21.56	.0000	.01042	.01250
HSAT	-.17724***	.00177	-100.23	.0000	-.18071	-.17377
INCOME	-.21880***	.03045	-7.19	.0000	-.27848	-.15913
	Model parameters for latent class 2					
Constant	.60562***	.03277	18.48	.0000	.54140	.66984
AGE	.01146***	.00053	21.56	.0000	.01042	.01250
HSAT	-.17724***	.00177	-100.23	.0000	-.18071	-.17377
INCOME	-.21880***	.03045	-7.19	.0000	-.27848	-.15913
	Model parameters for latent class 3					
Constant	1.94821***	.03015	64.61	.0000	1.88911	2.00731
AGE	.01146***	.00053	21.56	.0000	.01042	.01250
HSAT	-.17724***	.00177	-100.23	.0000	-.18071	-.17377
INCOME	-.21880***	.03045	-7.19	.0000	-.27848	-.15913
	Estimated prior probabilities for class membership					
ONE_1	-1.02248***	.26972	-3.79	.0002	-1.55112	-.49383
FEMALE_1	.24984***	.09529	2.62	.0087	.06307	.43662
MARRIE_1	-.31995***	.09893	-3.23	.0012	-.51385	-.12605
EDUC_1	-.03069	.02108	-1.46	.1455	-.07200	.01063
ONE_2	.39483**	.17044	2.32	.0205	.06077	.72889
FEMALE_2	-.73345***	.06013	-12.20	.0000	-.85129	-.61560
MARRIE_2	-.11502*	.06663	-1.73	.0843	-.24562	.01558
EDUC_2	.00774	.01310	.59	.5546	-.01793	.03341
ONE_3	0.0	.....(Fixed Parameter).....				
FEMALE_3	0.0	.....(Fixed Parameter).....				
MARRIE_3	0.0	.....(Fixed Parameter).....				
EDUC_3	0.0	.....(Fixed Parameter).....				

Prior class probabilities at data means for LCM variables

Class 1	Class 2	Class 3	Class 4	Class 5
.09900	.46041	.44059	.00000	.00000

Heckman and Singer Binary ChoiceModel – 3 Points

Latent Class / Panel Probit Model
Dependent variable HEALTHY
Log likelihood function -3442.77242
Sample is 7 pds and 887 individuals
PROBIT (normal) probability model

HEALTHY	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Model parameters for latent class 1					
Constant	1.82601***	.32656	5.59	.0000	1.18595	2.46606
AGE	-.03515***	.00423	-8.31	.0000	-.04344	-.02687
EDUC	.08945***	.01750	5.11	.0000	.05515	.12375
HHNINC	.49818***	.16176	3.08	.0021	.18113	.81523
MARRIED	.03473	.08490	.41	.6825	-.13167	.20113
HHKIDS	.18370***	.06214	2.96	.0031	.06190	.30550
	Model parameters for latent class 2					
Constant	.50135	.31077	1.61	.1067	-.10775	1.11045
AGE	-.03515***	.00423	-8.31	.0000	-.04344	-.02687
EDUC	.08945***	.01750	5.11	.0000	.05515	.12375
HHNINC	.49818***	.16176	3.08	.0021	.18113	.81523
MARRIED	.03473	.08490	.41	.6825	-.13167	.20113
HHKIDS	.18370***	.06214	2.96	.0031	.06190	.30550
	Model parameters for latent class 3					
Constant	-.75636**	.31567	-2.40	.0166	-1.37506	-.13767
AGE	-.03515***	.00423	-8.31	.0000	-.04344	-.02687
EDUC	.08945***	.01750	5.11	.0000	.05515	.12375
HHNINC	.49818***	.16176	3.08	.0021	.18113	.81523
MARRIED	.03473	.08490	.41	.6825	-.13167	.20113
HHKIDS	.18370***	.06214	2.96	.0031	.06190	.30550
	Estimated prior probabilities for class membership					
Class1Pr	.31094***	.03386	9.18	.0000	.24458	.37730
Class2Pr	.45267***	.03032	14.93	.0000	.39324	.51209
Class3Pr	.23639***	.03017	7.84	.0000	.17727	.29552

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

Heckman/Singer vs. REM

```
----------------------------------------------------------------------------
Random Effects Binary Probit Model
Sample is  7 pds and   887 individuals.
--------+-------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
 HEALTHY|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|     .33609         .29252     1.15  .2506     -.23723    .90941
(Other coefficients omitted)
     Rho|     .52565***      .02025    25.96  .0000      .48596    .56534
--------+-------------------------------------------------------------------
```

Rho = $\sigma^2/(1+s2)$ so σ^2 = rho/(1-rho) = 1.10814.
Mean = .33609, Variance = 1.10814

```
For Heckman and Singer model,
3 points        a1,a2,a3 = 1.82601, .50135, -.75636
3 probabilities p1,p2,p3 =  .31094, .45267,  .23639
Mean = .61593   variance = .90642
```

Modeling Obesity with a Latent Class Model

Mark Harris
Department of Economics, Curtin University

Bruce Hollingsworth
Department of Economics, Lancaster University

William Greene
Stern School of Business, New York University

Pushkar Maitra
Department of Economics, Monash University

Two Latent Classes: Approximately Half of European Individuals

Science AAAS.ORG | FEEDBACK | HELP | LIBRARIANS Science Magazine

AAAS NEWS SCIENCE JOURNALS CAREERS BLOGS & COMMUNITIES

Science The World's Leading Journal of Original Scientific Research, Global News, and Comme

Science Home Current Issue Previous Issues *Science* Express *Science* Products My *Science* Abou

< Prev | Table of Contents | N

Published Online April 12 2007
Science 11 May 2007:
Vol. 316 no. 5826 pp. 889-894
DOI: 10.1126/science.1141634

REPORT

A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity

An Ordered Probit Approach

A Latent Regression Model for "**True BMI**"

$BMI^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon,\ \varepsilon \sim N[0,\sigma^2],\ \sigma^2 = 1$

"**True BMI**" = a proxy for weight is unobserved

Observation Mechanism for **Weight Type**

WT	= 0 if	$BMI^* \leq 0$	Normal
	1 if	$0 < BMI^* \leq \mu$	Overweight
	2 if	$\mu < BMI^*$	Obese

Latent Class Modeling

- Several ‘types’ or ‘classes. Obesity be due to **genetic** reasons (the FTO gene) or **lifestyle** factors
- Distinct sets of individuals may have differing reactions to various policy tools and/or characteristics
- The observer does not know from the data which class an individual is in.
- Suggests a latent class approach for health outcomes (Deb and Trivedi, 2002, and Bago d’Uva, 2005)

Latent Class Application

- Two class model (considering FTO gene):
  - More classes make class interpretations much more difficult
  - Parametric models proliferate parameters
- Two classes allow us to correlate the unobservables driving class membership and observed weight outcomes.
- Theory for more than two classes not yet developed.

Correlation of Unobservables in Class Membership and BMI Equations

Class Membership: $C^* = \boldsymbol{\delta}'\mathbf{z}_i + \boxed{u_i}, \quad C = 1[C^* > 0] \quad \text{(Probit)}$

BMI|Class=0,1 $\quad BMI^* = \boldsymbol{\beta}_c'\mathbf{x}_i + \boxed{\varepsilon_{c,i}}, \quad \text{BMI group} = OP[BMI^*, \mu(\boldsymbol{\alpha}_c'\mathbf{w}_i)]$

Endogeneity: $\begin{pmatrix} u_i \\ \varepsilon_{c,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \boxed{\rho_c} \\ \boxed{\rho_c} & 1 \end{pmatrix}\right]$

Bivariate Ordered Probit (one variable is binary).

Full information maximum likelihood.

Outcome Probabilities

- Class 0 dominated by normal and overweight probabilities ‘normal weight’ class
- Class 1 dominated by probabilities at top end of the scale ‘non-normal weight’
- Unobservables for weight class membership, negatively correlated with those determining weight levels:

Male	Sample	Class 0	Class 1
Normal	0.315	0.306	0.009
Overweight	0.439	0.180	0.258
Obese	0.246	0.001	0.245
ρ		-0.018	0.036

Female	Sample	Class 0	Class 1
Normal	0.439	0.318	0.121
Overweight	0.298	0.084	0.214
Obese	0.264	0.002	0.261
ρ		-0.067	-0.131

Classification (Latent Probit) Model

	Male	Female
Constant	-2.233**	-2.758**
US born	0.532**	0.527**
Mother/Father born O/S	0.009	0.269**
Age	0.877**	1.153**
Age^2	-0.086**	-0.106**
White	-0.178**	-0.340**
Black	-0.094	0.350**

Inflated Responses in Self-Assessed Health

Mark Harris
Department of Economics, Curtin University
Bruce Hollingsworth
Department of Economics, Lancaster University
William Greene
Stern School of Business, New York University

SAH vs. Objective Health Measures

Favorable SAH categories seem artificially high.

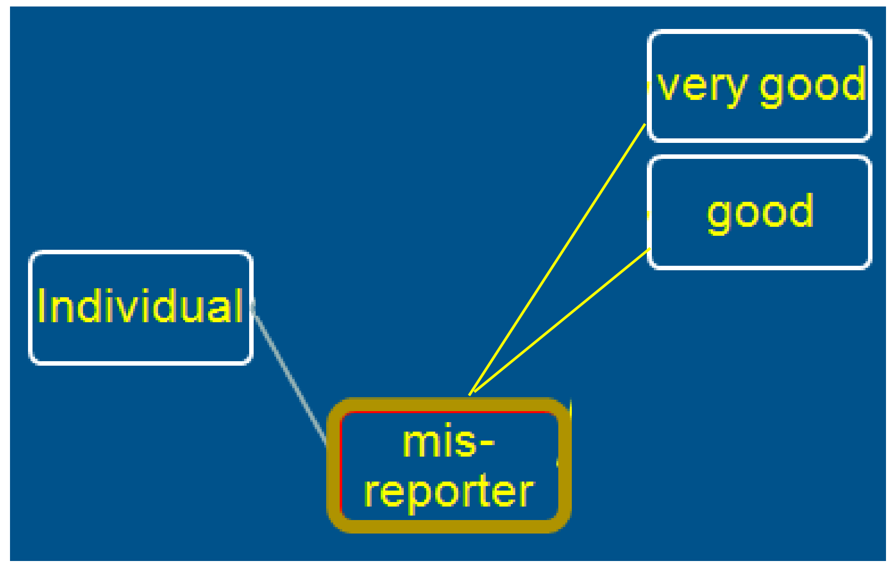
- 60% of Australians are either overweight or obese (Dunstan *et. al*, 2001)
- 1 in 4 Australians has either diabetes or a condition of impaired glucose metabolism
- Over 50% of the population has elevated cholesterol
- Over 50% has at least 1 of the “deadly quartet” of health conditions (diabetes, obesity, high blood pressure, high cholestrol)
- Nearly 4 out of 5 Australians have 1 or more long term health conditions (National Health Survey, Australian Bureau of Statistics 2006)
- Australia ranked #1 in terms of obesity rates

Similar results appear to appear for other countries

A Two Class Latent Class Model

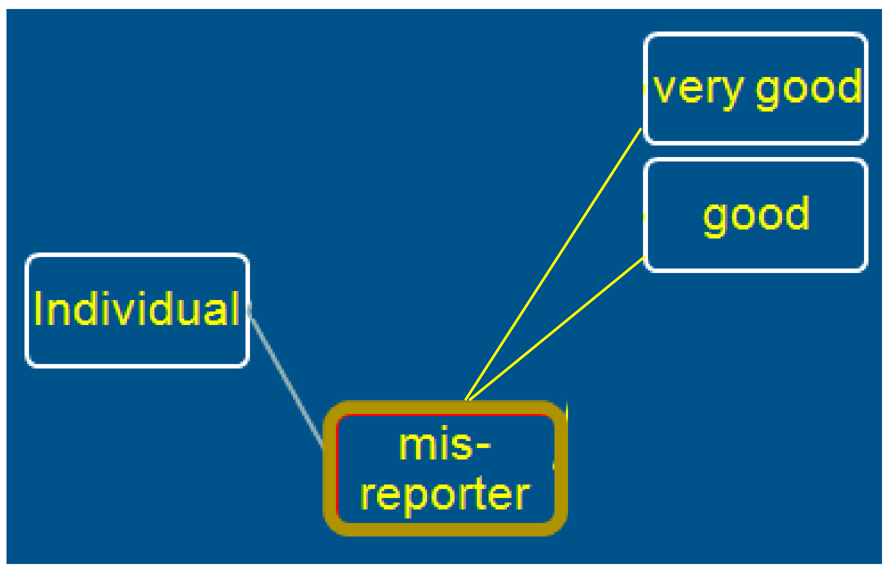
True Reporter

Misreporter



- Mis-reporters choose either good or very good
- The response is determined by a probit model

$$m^* = x_m'\beta_m + \varepsilon_m$$



Y=3

Y=2

- True-reporters can choose any outcome

- This actual response will be determined by a further latent variable equation, along the lines of the previous Ordered Probit set-up...

$$y^{*} = x_y' \beta_y + \varepsilon_y$$

Observed Mixture of Two Classes

For true-reporters the respective probabilities for each outcome are simply the ordered probit ones, weighted by the probability of being a truthful respondent (0=poor to 4=excellent):

$$
\Pr(true, y) = \begin{cases}
0 = \Phi(x_r'\beta_r) \times \left[\Phi(-x_y'\beta_y)\right] \\
1 = \Phi(x_r'\beta_r) \times \left[\Phi(\mu_1 - x_y'\beta_y) - \Phi(-x_y'\beta_y)\right] \\
2 = \Phi(x_r'\beta_r) \times \left[\Phi(\mu_2 - x_y'\beta_y) - \Phi(\mu_1 - x_y'\beta_y)\right] \\
3 = \Phi(x_r'\beta_r) \times \left[\Phi(\mu_3 - x_y'\beta_y) - \Phi(\mu_2 - x_y'\beta_y)\right] \\
4 = \Phi(x_r'\beta_r) \times \left[1\text{-}\Phi(\mu_3 - x_y'\beta_y)\right]
\end{cases}
$$

Pr(true,y) = Pr(true) * Pr(y | true)

- Putting all these elements together we have a latent class model:

$$\Pr(y)=\begin{cases} 0=\Phi(x_r'\beta_r)\times\left[\Phi(-x_y'\beta_y)\right] \\ 1=\Phi(x_r'\beta_r)\times\left[\Phi(\mu_1-x_y'\beta_y)-\Phi(-x_y'\beta_y)\right] \\ 2=\Phi(x_r'\beta_r)\times\left[\Phi(\mu_2-x_y'\beta_y)-\Phi(\mu_1-x_y'\beta_y)\right]+\Phi(-x_r'\beta_r)\Phi(-x_m'\beta_m) \\ 3=\Phi(x_r'\beta_r)\times\left[\Phi(\mu_3-x_y'\beta_y)-\Phi(\mu_2-x_y'\beta_y)\right]+\Phi(-x_r'\beta_r)\Phi(x_m'\beta_m) \\ 4=\Phi(x_r'\beta_r)\times\left[1\text{-}\Phi(\mu_3-x_y'\beta_y)\right] \end{cases}$$

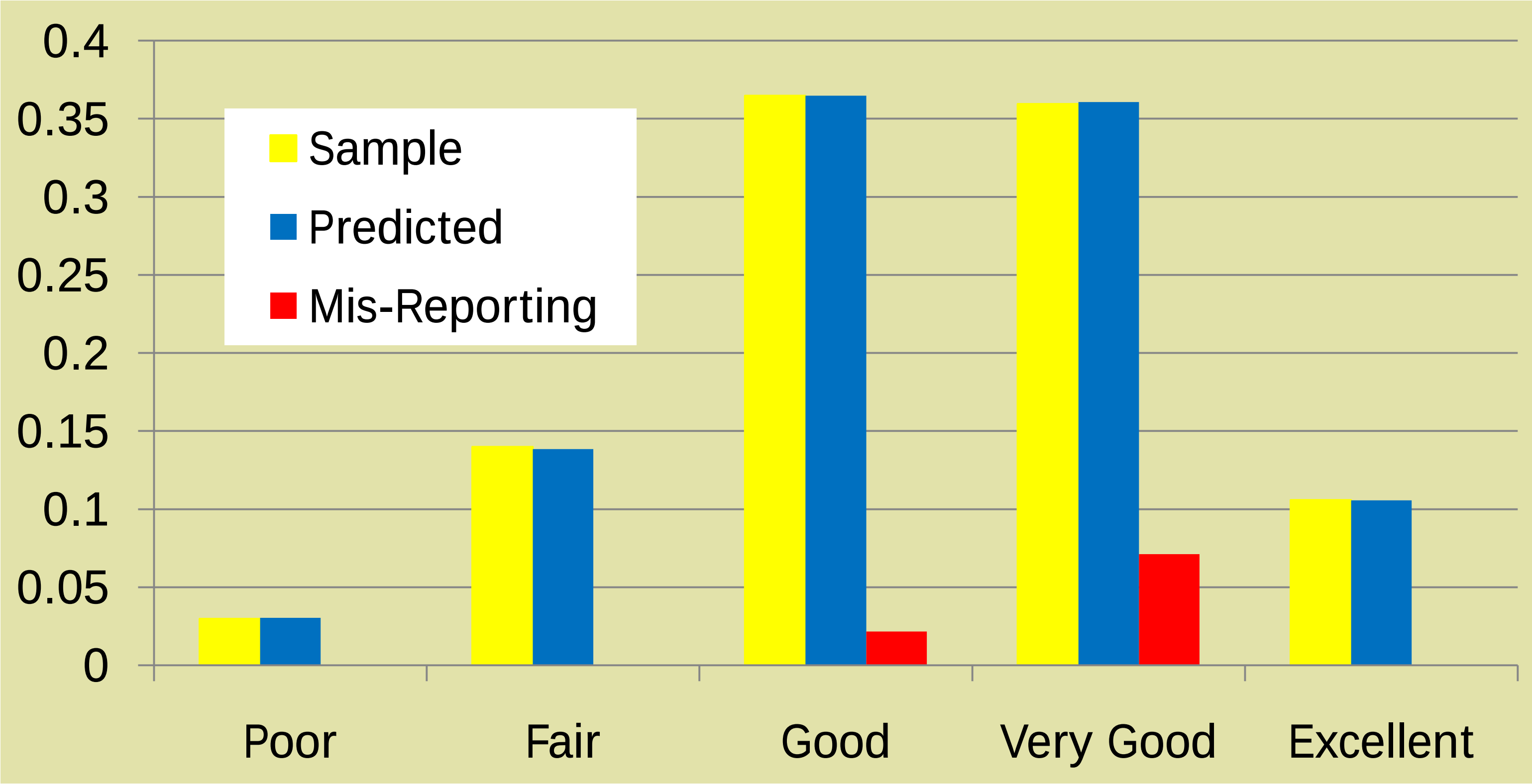
$$\Pr(y)=\Pr(\mathit{true})\Pr(y\mid\mathit{true})+\Pr(\mathit{misreporter})\Pr(y\mid\mathit{misreporter})$$

- With nonzero correlations, probabilities are now functions of bivariate normal cdf's:

$$\begin{cases} 0 = \Phi_2\left(x_r'\beta_r, -x_y'\beta_y; -\boxed{\rho_{ry}}\right) \\ 1 = \Phi_2\left(x_r'\beta_r, \mu_1 - x_y'\beta_y; -\boxed{\rho_{ry}}\right) - \Phi_2\left(x_r'\beta_r, -x_y'\beta_y, -\boxed{\rho_{ry}}\right) \\ 2 = \left[\Phi_2\left(x_r'\beta_r, \mu_2 - x_y'\beta_y; -\boxed{\rho_{ry}}\right) - \Phi_2\left(x_r'\beta_r, \mu_1 - x_y'\beta_y; -\boxed{\rho_{ry}}\right)\right] + \Phi_2\left(-x_r'\beta_r, -x_m'\beta_m; \boxed{\rho_{rm}}\right) \\ 3 = \left[\Phi_2\left(x_r'\beta_r, \mu_3 - x_y'\beta_y; -\boxed{\rho_{ry}}\right) - \Phi_2\left(x_r'\beta_r, \mu_2 - x_y'\beta_y; -\boxed{\rho_{ry}}\right)\right] + \Phi_2\left(-x_r'\beta_r, x_m'\beta_m; -\boxed{\rho_{rm}}\right) \\ 4 = \Phi_2\left(x_r'\beta_r, x_y'\beta_y - \mu_3; \boxed{\rho_{ry}}\right) \end{cases}$$

- Again inflates the good/very good categories

General Result



HEALTH ECONOMICS
Health Econ. **22**: 554–567 (2013)
Published online 20 April 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.2824

INVESTIGATING ATTRIBUTE NON-ATTENDANCE AND ITS CONSEQUENCES IN CHOICE EXPERIMENTS WITH LATENT CLASS MODELS

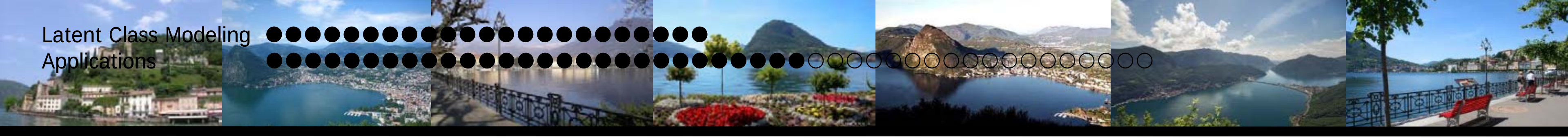
MYLENE LAGARDE*
Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK

Second, this article contributes to the study of heuristics in DCEs in the health literature. As underlined before, the majority of DCE studies in health have identified dominant preferences, which constitute special cases of ANA. This paper not only describes and accounts for dominant preferences, but it encompasses all other ANA response strategies. Doing that, it adds to a very limited literature in health, where ANA has been mostly ignored despite its critical importance and potential relevance. One qualitative study of the responses given by 18 respondents in a DCE exploring preferences for cancer screening identified that only five respondents seemed to consider all attributes, whereas the rest revealed that they employed various attribute non-attendance strategies (Ryan *et al*., 2009). In a recent study, Hole (2011b) used a two-step process to account for endogenous ANA of respondents and found that '*a substantial share of the respondents ignored one or more attributes when making their choices*'. Although his approach overcame the problems of LCMs mentioned above, it did not present a precise description of the various ANA rules employed by respondents. In contrast with Hole (2011b), which is, to date, the unique quantitative study of ANA in health DCEs, this paper provides a full description of all the response patterns used by respondents in the DCE, as well as their frequency.

… only five respondents seemed to consider all attributes, whereas the rest revealed that they employed various attribute nonattendance strategies …

The 2^K model

- The analyst believes some attributes are ignored. There is no definitive indicator.
- Classes distinguished by which attributes are ignored
- A latent class model applies. For K attributes there are 2^K candidate coefficient vectors



A Latent Class Model

$$\left\{ \begin{bmatrix} \text{Uncertainty} & \text{Toll Cost} & \text{Running Cost} \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix} \begin{bmatrix} \text{Free Flow} & \text{Slowed} & \text{Start / Stop} \\ 0 & 0 & 0 \\ \beta_4 & 0 & 0 \\ 0 & \beta_5 & 0 \\ 0 & 0 & \beta_6 \\ \beta_4 & \beta_5 & 0 \\ \beta_4 & 0 & \beta_6 \\ 0 & \beta_5 & \beta_6 \\ \beta_4 & \beta_5 & \beta_6 \end{bmatrix} \right\}$$

2.1. Using latent class models to account for attribute non-attendance

Through their econometric specification, LCMs provide an alternative approach to models such as multinomial logit and mixed logit to accommodate response heterogeneity. In LCMs, it is assumed that the population of respondents can be divided into a set number (Q) of classes, or groups of individuals, who will differ in their preferences. In other words, although the groups are different from each other (i.e. they are defined by different parameter vectors), all members of a same group share the same parameters. As the analyst ignores which observation is in which class, the model assumes that individuals belong to a certain group up to a probability. The logit choice probability function of choosing one particular alternative from J alternatives for an individual i belonging to a specific class q can be then written as

$$Pr\left(\mathrm{y_{it}} = 1 \mid class\ q\right) = \mathrm{P}_{it|}\mathrm{q} = \frac{e^{X'_{itj}\beta_q}}{\sum_{\mathrm{j=1}}^{J} e^{X'_{itj}\beta_q}} \tag{1}$$

The probability that an individual i belongs to class q (out of a total of Q classes) is given by

$$H_{iq} = \frac{e^{\theta_q}}{\sum_{\mathrm{q=1}}^{Q} e^{\theta_q}}, \mathrm{q} = 1, \ldots, \ \mathrm{Q \ and} \ \theta_{\mathrm{Q}} = 0 \tag{2}$$

The data used in this study come from a DCE designed to elicit preferences regarding the introduction of new guidelines to managing malaria in pregnancy in Ghana (Lagarde *et al*., 2011). The choice experiment was designed after a series of focus group discussions and in-depth interviews with healthcare providers and a pilot study.

Six attributes describing the conditions of malaria case management by ante-natal care were used (see Table I): the type of treatment approach to managing malaria in pregnancy, the drugs prescribed to pregnant women, the workload, the potential monthly bonus included in the policy and the likely health outcomes for mothers (incidence of severe anaemia) and babies (incidence of low birth weight). An orthogonal D-efficient experimental design of 16 choice sets was created using the macros developed for SAS (Kuhfeld, 2009). Each choice set consisted of two generic alternatives representing two policies that could be introduced to manage malaria in pregnancy (see Figure 1).

… a discrete choice experiment designed to elicit preferences regarding the introduction of new guidelines to managing malaria in pregnancy in Ghana …

Table I. Attributes and levels in the choice experiment

Attribute	Levels
The type of approach to managing malaria in pregnancy	▪ Preventive approach ▪ Curative approach (test and treat if parasite positive)
The anti-malarial drugs you have to prescribe to pregnant women	▪ SP (Fansidar) ▪ Artesunate-amodiaquine (AS–AQ)
Prevalence of anaemia for mothers treated with protocol	▪ 1% ▪ 15%
Prevalence of low birth weight among infants of mothers treated with the protocol	▪ 10% ▪ 15%
Staffing level for the ANC clinic	▪ Under-staffed ▪ Adequately staffed
The salary supplement included in the protocol	▪ GH. C10 ▪ GH. C20

Table II. Average proportion of respondents who ignored one attribute

Model	Attribute assumed to have been ignored in the second class	Average class membership (%)
Model 1	Low birth weight ignored	7.99
Model 2	Anaemia ignored	17.24
Model 3	Bonus ignored	21.75
Model 4	Workload ignored	46.92
Model 5	Drug ignored	62.57
Model 6	Treatment ignored	66.76