THE STATA JOURNAL

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Fitting adjusted limited dependent variable mixture models to EQ-5D

Mónica Hernández Alava
School of Health and Related Research
Health Economics and Decision Science
University of Sheffield
Sheffield, UK
monica.hernandez@sheffield.ac.uk

Allan Wailoo
School of Health and Related Research
Health Economics and Decision Science
University of Sheffield
Sheffield, UK
a.j.wailoo@sheffield.ac.uk

Abstract. In this article, we describe the `aldvmm` command for fitting adjusted limited dependent variable mixture models to either UK or U.S. tariff EQ-5D data. We present and explain the command and postestimation command through examples. The `aldvmm` command requires use of Stas Kolenikov's simulated annealing package (`simann()`), which can be easily installed by typing `net install simann.pkg, from(http://web.missouri.edu/~kolenikovs/stata)`.

Keywords: st0401, aldvmm, adjusted limited dependent variable mixture, EQ-5D, EQ-5D-3L mapping

1 Introduction

Quality adjusted life years (QALY) are used in many assessments for cost effectiveness of health interventions. However, often an evidence gap exists between clinical measures of effect that are available and the detailed preference-based information needed to construct QALY measures. QALY attaches a value of 1 to each year in full health and a value of 0 to death. These two values represent anchor points for any other health state. Instruments like the EQ-5D-3L (EQ-5D) have preference-based scoring systems and are favored by organizations such as the National Institute for Health and Care Excellence for the estimation of QALY. The EQ-5D questionnaire asks individuals to describe their health using five different dimensions: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression. Each dimension has three levels: no problems, some problems, and extreme problems. There are 243 theoretically possible health states described by this instrument, and each is assigned a value based on general public preferences (see Dolan et al. [1995] for the UK and Shaw, Johnson, and Coons [2005] for the United States).

EQ-5D is frequently absent from clinical studies of treatment effects, which prevents the direct calculation of QALY. Often this gap is bridged by "mapping"—estimating a relationship between observed clinical outcomes and preference-based measures with data from another dataset containing both types of information. However, the distribution of EQ-5D exhibits characteristics that make standard models inappropriate.

The adjusted limited dependent variable mixture model variable was first proposed by Hernández Alava, Wailoo, and Ara (2012) to deal with the distributional features presented by EQ-5D. The command `aldvmm` estimates the variant of the model presented in Hernández Alava et al. (2013) and Hernández Alava et al. (2014).

The article is organized as follows. In section 2, we briefly overview the adjusted limited dependent variable mixture model. In section 3, we describe the `aldvmm` syntax and options, including the syntax for `predict`. In section 4, we give some examples.

2   Adjusted limited dependent variable mixture model

The distribution of EQ-5D exhibits several characteristics that must be considered when fitting "mapping" models. EQ-5D values are limited both at the top and at the bottom. The highest attainable EQ-5D value is 1, which represents perfect health. At the other extreme, -0.594 corresponds to extreme problems in all five dimensions of the descriptive system in the UK tariff; the value is -0.109 in the U.S. tariff. EQ-5D attaches a value of 0 to death; thus a few health states described by EQ-5D are considered worse than death. There are usually a mass of observations at the upper limit (1). However, use of the standard tobit model is not appropriate for two reasons. First, there is a large gap between the mass at 1 and the next feasible EQ-5D value (0.883 and 0.860 for the UK and U.S. tariffs, respectively). Second, the rest of the distribution usually shows strong bimodality, often with a high degree of skewness. These characteristics often remain even after conditioning.

The adjusted limited dependent variable mixture model (Hernández Alava, Wailoo, and Ara 2012) was proposed as a flexible alternative to model EQ-5D data and has been shown to perform better than models used traditionally in this area. It is a mixture model of adjusted tobitlike distributions. A brief description of the model follows. A more detailed description and other variants can be found in Hernández Alava, Wailoo, and Ara (2012), Hernández Alava et al. (2013), and Hernández Alava et al. (2014).

It is assumed that EQ-5D (denoted by y_i) can be modeled as a mixture of C—components or classes. Conditional on an individual observation i belonging to component c $(c = 1, \ldots, C)$, EQ-5D can be written as

$$
y_i | c = \begin{cases} 1 & \text{if } y_i^* | c > \Psi_1 \\ \Psi_2 & \text{if } y_i^* | c <= \Psi_2 \\ y_i^* | c & \text{otherwise} \end{cases} \tag{1}
$$

where $\Psi_1 = 0.883$ and $\Psi_2 = -0.594$ for the UK, and $\Psi_1 = 0.860$ and $\Psi_2 = -0.109$ for the United States. For each mixture component c,

$$
y_i^* | c = \mathbf{x}_i^{'} \boldsymbol{\beta}_c + \varepsilon_{ic} \tag{2}
$$

$\boldsymbol{\beta}_c$ is a $(k \times 1)$ vector of coefficients including an intercept term, $\mathbf{x}_i^{'}$ is a row vector of covariates, and ε_{ic} is independent and identically distributed $N\left(0, \sigma_c^2\right)$. A multinomial logit model for the probability of latent class membership is assumed as

$$P\left(c|\mathbf{w}_i^{'}\right) = \frac{\exp\left(\mathbf{w}_i^{'}\boldsymbol{\delta}_c\right)}{\sum\limits_{s=1}^{C} \exp\left(\mathbf{w}_i^{'}\delta_s\right)} \tag{3}$$

where $\mathbf{w}_i^{'}$ is a vector of variables that affect the probability of component membership, $\boldsymbol{\delta}_c$ is the vector of corresponding coefficients, and C is the number of classes used in the analysis. One set of coefficients, $\boldsymbol{\delta}_c$, is normalized to zero for identification. If no variables are included, then the probabilities of component membership are constant for all individuals.

The log likelihood of the model defined by (1), (2), and (3) can be written as

$$\begin{aligned}
\ln l \quad = \quad & \sum_{i=1}^{n} \ln \left(\sum_{c=1}^{C} \frac{\exp\left(\mathbf{w}_i^{'}\boldsymbol{\delta}_c\right)}{\sum\limits_{s=1}^{C} \exp\left(\mathbf{w}_i^{'}\delta_s\right)} \left[\mathbf{1}\left(y_i > \Psi_1\right) \left\{ 1 - \Phi\left(\frac{\Psi_1 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \right. \right. \\
& + \mathbf{1}\left(y_i \leq \Psi_2\right) \left\{ \Phi\left(\frac{\Psi_2 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \\
& \left. \left. + \mathbf{1}\left(\Psi_2 < y_i < \Psi_1\right) \left\{ \frac{1}{\sigma_c} \phi\left(\frac{y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \right] \right)
\end{aligned} \tag{4}$$

where $\mathbf{1}\left(\cdot\right)$ is the indicator function, $\phi\left(\cdot\right)$ is the standard normal density function, and $\Phi\left(\cdot\right)$ is the standard cumulative normal.

After fitting the model, one can use the following conditional expectation to predict EQ-5D:

$$\begin{aligned}
E\left(y_i|\mathbf{x}_i^{'}\mathbf{w}_i^{'}\right) \quad = \quad & \sum_{c=1}^{C} \frac{\exp\left(\mathbf{w}_i^{'}\boldsymbol{\delta}_c\right)}{\sum\limits_{s=1}^{C} \exp\left(\mathbf{w}_i^{'}\delta_s\right)} \left[\left\{ 1 - \Phi\left(\frac{\Psi_1 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \right. \\
& + \left\{ \Phi\left(\frac{\Psi_2 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \Psi_2 \\
& + \left\{ \Phi\left(\frac{\Psi_1 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) - \Phi\left(\frac{\Psi_2 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) \right\} \\
& \left. \left\{ \mathbf{x}_i^{'}\boldsymbol{\beta}_c + \sigma_c \frac{\phi\left(\frac{\Psi_1 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) - \phi\left(\frac{\Psi_2 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right)}{\Phi\left(\frac{\Psi_2 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right) - \Phi\left(\frac{\Psi_1 - \mathbf{x}_i^{'}\boldsymbol{\beta}_c}{\sigma_c} \right)} \right\} \right]
\end{aligned}$$

Note that this is an average of the predictions for each component weighted by the corresponding probability of component membership.

3   The aldvmm command

3.1   Syntax

aldvmm *depvar* $\big[$*indepvars*$\big]$ $\big[$*if*$\big]$ $\big[$*in*$\big]$ $\big[$*weight*$\big]$, <u>nc</u>omponents(*#*)
   $\big[$<u>prob</u>abilities(*varlist*) country(*country*) llim(*#*) ulim(*#*)
   <u>c</u>onstraints(*numlist*) vce(*vcetype*) <u>l</u>evel(*#*) <u>ini</u>method(*inimethod*)
   <u>saopts</u>(*matrix*) *maximize_options* <u>se</u>arch(*spec*) <u>r</u>epeat(*#*)$\big]$

3.2   Description

aldvmm is a user-written program that fits an adjusted limited dependent variable mix-
ture model using maximum likelihood estimation. It is implemented as an ll ml eval-
uator. The model is a C-component mixture of densities adjusted to deal with EQ-5D
data. The mean of a density within a component as well as the mixing probabilities may
be functions of covariates. The default model allows the variances of the components to
be different, but they can be constrained to be the same via the constraints() option.

3.3   Options

ncomponents(*#*) specifies the number of mixture components. Strictly, a mixture
   model has a minimum of two components, but aldvmm does allow the estimation of
   a model with only one component. This one-component model is similar to a tobit
   model but can reflect the gap found in EQ-5D. ncomponents() is required.

probabilities(*varlist*) specifies a set of variables to be used to model the probability
   of component membership. The probabilities are specified using a multinomial logit
   parameterization. The default is to use constant probabilities.

country(*country*) specifies the EQ-5D tariff. The string *country* may be UK or US. The
   default is country(UK). This option is ignored if llim(*#*) and ulim(*#*) are supplied
   by the user.

llim(*#*) specifies the user-supplied lower limit of EQ-5D (Ψ_2). llim() and ulim()
   must be provided together.

ulim(*#*) specifies the user-supplied highest EQ-5D index value below 1 (Ψ_1). Setting
   # to 1 fits a model without a gap, that is, a mixture of tobit models. llim() and
   ulim() must be provided together.

constraints(*numlist*); see [R] **estimation options**.

vce(*vcetype*) specifies how to estimate the variance–covariance matrix corresponding to
   the parameter estimates. The supported options are oim, opg, <u>robust</u>, or <u>cluster</u>
   clustvar. The current version of the command does not allow bootstrap or jacknife
   estimators; see [R] *vce_option*.

level(*#*); see [R] **estimation options**.

inimethod(*inimethod*) specifies the method for choosing starting values for parameters. *inimethod* may be single, cons, or simann. The default is inimethod(single), which lets ml find starting values. cons fits first a constant-only model and uses those parameters as starting values in the estimation of the full model. simann runs simulated annealing first to find appropriate starting values. Simulated annealing can be slow depending on the arguments used (see help simann()). The default arguments for simann() can be changed by using the saopts(*matrix*) option.

saopts(*matrix*) specifies the name of the matrix with the following simann() arguments: *count*, *ftol*, *steps*, *cooling*, *start*, and *loglevel*.

maximize_options: <u>diffic</u>ult, <u>techn</u>ique(*algorithm_spec*), <u>iter</u>ate(*#*), [<u>no</u>]<u>log</u>, <u>trace</u>, <u>grad</u>ient, showstep, <u>hess</u>ian, <u>showtol</u>erance, <u>tol</u>erance(*#*), <u>ltol</u>erance(*#*), <u>gtol</u>erance(*#*), <u>nrtol</u>erance(*#*), <u>nonrtol</u>erance, and from(*init_specs*); see [R] **maximize**.

search(*spec*) specifies whether to use ml's initial search algorithm or not. *spec* may be on or off.

repeat(*#*) specifies the number of random attempts to be made to find a better initial-value vector. This option is used in conjunction with search(on).

4   predict

4.1   Syntax

predict *newvar* [*if*] [*in*] [, <u>out</u>come(*outcome*)]

4.2   Description

Stata's standard predict command can be used following aldvmm to obtain predicted probabilities for the dependent variable as well as predicted means and associated probabilities for each component in the mixture.

4.3   Option

outcome(*outcome*) specifies the predictions to be stored. *outcome* can be y or all. The default, outcome(y), stores only the dependent variable prediction in *newvar*. Use all to additionally obtain the predicted means and probabilities for each component in the mixture. These are stored as *newvar*_y1, *newvar*_y2, ... and *newvar*_p1, *newvar*_p2, ..., respectively.

5   The aldvmm command in practice

We now show how to use the `aldvmm` command to model EQ-5D data. We use UK tariff data from the Patient Reported Outcome Measures in England, April 2011 to March 2012 (Health and Social Care Information Centre). The data are freely available and can be downloaded from http://www.hscic.gov.uk/catalogue/PUB11359. For this example, we select a 30% random sample of individuals with data on age and gender (age and gender are excluded from the dataset for those patients who could be identified because of low numbers). We use postoperative data on EQ-5D and the Oxford hip score of patients who have undergone a hip replacement. The Oxford hip score questionnaire combines a patient's answers to 12 multiple-choice questions relevant to hips into one score, and it is designed to assess symptoms and function in patients undergoing hip replacements. Each question has 4 possible response categories; a score of 4 is assigned to the category representing the least or no symptoms, and a score of 0 is attached to the category representing symptoms of the greatest severity. The individual scores are then added together to one score with 0 denoting the worst possible symptoms and function and 48 denoting the best. Further details of the dataset can be found in Wineberg (2014).

Figure 1 shows a histogram of EQ-5D exhibiting the usual characteristics: a mass of observations at 1, a gap where no EQ-5D values are possible, and then a bimodal distribution.
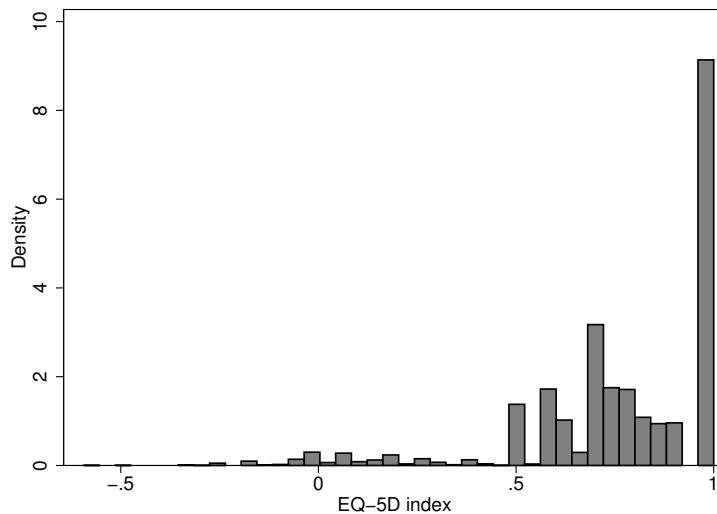

Figure 1. Histogram of EQ-5D data

Mixture models are extremely flexible and are a convenient semiparametric way to model data with characteristics not easily accommodated by known distributions. Mixtures of normal distributions can generate multimodality, strong skewness in a unimodal

distribution, and kurtotic densities; in fact, they can generate an incredibly large number of distributional shapes. It is important to emphasize that bimodality does not necessarily imply a model with two components. The optimal model might have three or possibly more components if the distribution presents asymmetries or peaks.

We recommend that readers become familiar with the idiosyncrasies of fitting mixture models (McLachlan and Peel 2000) before attempting to estimate one. We will briefly describe the two main issues that researchers are likely to encounter when trying to fit models of EQ-5D data. One of the problems of fitting mixture models relates to the presence of several local maximums in the likelihood function. We cannot assume that by running the model and getting some estimated parameters, the consistent solution has been found. To identify the global maximizer, we need, at the very least, to try different sets of random starting values and to select the solution with the highest likelihood function. Alternatively, a global optimization algorithm such as simulated annealing can be used. The `aldvmm` command can use Stas Kolenikov's `simann()` Mata function for simulated annealing. We recommend using this option when fitting only a few components because it could be time consuming and because it cannot restrict the parameter space. Another problem arises when estimating mixtures with different σ_c across components: the likelihood function becomes unbounded as the variance of a component tends to zero. It is not a "real" problem (Aitkin 1997); rather, it is due to the inability of the normal distribution to characterize the likelihood when the variances tend to zero. In essence, as the variance of one component becomes very small, the component turns into a conditional probability mass. However, the likelihood contribution of that component becomes infinite in (4) because we are dividing by a very small number. In this situation, we cannot trust the value of the likelihood. Usually, provided that certain regularity conditions are met, the consistent solution will correspond to a local maximizer. EQ-5D data usually have a mass of observations at one corresponding to individuals who are in full health and then have no immediately adjacent observations. If we try to estimate a standard mixture of normal distributions, we will quickly encounter problems with unbounded likelihoods as the model tries to fit the mass of observations. The adaptation to the mixture of normals used by the `aldvmm` command ensures that the likelihood value is correct even if a component becomes a probability mass at one. However, as in the standard mixture of normals, the likelihood function of a model displaying a component with a near-zero variance in the interior of the EQ-5D range should not be relied upon for model selection.

When one fits mixture models, it is important to start with a few components and use them as stepping stones to fit models with more components. We begin by fitting a simple "mapping" function of the Oxford hip score (divided by 10) to EQ-5D using a two-component model.

```
. use ohr11_12.dta
. generate male = sex == 1
. generate hr10 = hr/10
. aldvmm eq5d hr10, ncomponents(2)
initial:        log likelihood = -14123.606
  (output omitted )
Iteration 14:  log likelihood = -577.37808
2 component Adjusted Limited Dependent Variable Mixture Model
                                          Number of obs    =      10,565
                                          Wald chi2(1)     =           .
Log likelihood = -577.37808               Prob > chi2      =           .
```

eq5d	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Comp_1						
hr10	.2307964	.0022443	102.84	0.000	.2263977	.2351951
_cons	-.0883435	.0083791	-10.54	0.000	-.1047663	-.0719208
Comp_2						
hr10	885120.8	.	.	.	.	.
_cons	3930876	.	.	.	.	.
Prob_C1						
_cons	7.320611	.731422	10.01	0.000	5.88705	8.754171
/lns_1	-1.646109	.0089019	-184.92	0.000	-1.663556	-1.628661
/lns_2	-164.8187	.	.	.	.	.
sigma1	.1927987	.0017163			.189464	.196192
sigma2	2.63e-72	.			.	.
pi1	.9993387	.0004834			.9972325	.9998422
pi2	.0006613	.0004834			.0001578	.0027675

The output signals that something is wrong. The constant and the estimated co-efficient for the Oxford hip score are very large, and the standard errors are missing. The large estimated coefficients coupled with a very small standard deviation for that component effectively translate into a probability mass at one. The likelihood of the model is reliable in this case, and the missing standard errors signal that the parameters are not identified because small changes will still produce the same likelihood. If we believe that this is the consistent solution, we could use the `constraints()` option to fix the parameters to create the probability mass.

```
. matrix a = e(b)
. constraint 1 [Comp_2]:hr10 = 0
. constraint 2 [Comp_2]:_cons = 100
. constraint 3 [lns_2]:_cons = 1e-30
. aldvmm eq5d hr10, ncomponents(2) from(a) constraints(1 2 3)
initial:       log likelihood = -577.37808
rescale:       log likelihood = -577.37808
rescale eq:    log likelihood = -577.37808
Iteration 0:   log likelihood = -577.37808
Iteration 1:   log likelihood = -577.37808
2 component Adjusted Limited Dependent Variable Mixture Model
                                              Number of obs    =      10,565
                                              Wald chi2(1)     =    10575.71
Log likelihood = -577.37808                   Prob > chi2      =      0.0000
 ( 1)   [Comp_2]hr10 = 0
 ( 2)   [Comp_2]_cons = 100
 ( 3)   [lns_2]_cons = 1.00e-30
```

eq5d	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Comp_1						
hr10	.2307964	.0022443	102.84	0.000	.2263977	.2351951
_cons	-.0883435	.0083791	-10.54	0.000	-.1047663	-.0719208
Comp_2						
hr10	0	(omitted)				
_cons	100	(constrained)				
Prob_C1						
_cons	7.320611	.731422	10.01	0.000	5.88705	8.754171
/lns_1	-1.646109	.0089019	-184.92	0.000	-1.663556	-1.628661
/lns_2	1.00e-30	(constrained)				
sigma1	.1927987	.0017163			.189464	.196192
sigma2	1	(constrained)				
pi1	.9993387	.0004834			.9972325	.9998422
pi2	.0006613	.0004834			.0001578	.0027675

The fit model has the same value of the likelihood function, suggesting that our choice of parameters has not changed the specification. Component 2 is a component of 1s, but note that the probability of component membership (pi2) is very small. As highlighted earlier, it is well known that the likelihood functions of mixtures have multiple optima, and the usual local maximization algorithms might get stuck at a local maximum. When using these models, one should use a range of starting values to ascertain that the global maximum has been found. But first, one should take advantage of some of the options that have been programmed in the aldvmm command. One option that sometimes works well is to fit a constant-only model first and use the estimated parameters in the full-model specification. This can be accomplished using the inimethod(cons) option of the aldvmm command.

```
. aldvmm eq5d hr10, ncomponents(2) inimethod(cons)
Fitting constant-only model:
initial:       log likelihood = -14123.606
  (output omitted)
Iteration 9:   log likelihood = -3737.8838
Fitting full model:
initial:       log likelihood = -3737.8838
  (output omitted)
Iteration 15:  log likelihood =  685.78629
2 component Adjusted Limited Dependent Variable Mixture Model
```

			Number of obs	=	10,565
			LR chi2(2)	=	8847.34
Log likelihood = 685.78629			Prob > chi2	=	0.0000

eq5d	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Comp_1						
hr10	.3050275	.0063324	48.17	0.000	.2926162	.3174389
_cons	-.4029312	.0215507	-18.70	0.000	-.4451698	-.3606925
Comp_2						
hr10	.1480158	.0019441	76.13	0.000	.1442053	.1518263
_cons	.2261472	.0069948	32.33	0.000	.2124377	.2398566
Prob_C1						
_cons	-.7075574	.061444	-11.52	0.000	-.8279855	-.5871293
/lns_1	-1.263205	.0211453	-59.74	0.000	-1.304649	-1.221761
/lns_2	-2.45414	.0177415	-138.33	0.000	-2.488913	-2.419367
sigma1	.2827464	.0059788			.2712677	.2947107
sigma2	.0859371	.0015247			.0830002	.0889779
pi1	.3301388	.0135882			.3040712	.3572938
pi2	.6698612	.0135882			.6427062	.6959288

This model has a higher likelihood than the last model, confirming that the first set of estimated parameters related to only a local solution. In this solution, we find that all parameters are significant, and the two components now have sizable associated probabilities. In both components, EQ-5D increases as the Oxford hip score increases, but the sizes of the parameters are quite different. Based on these parameters, we could do a further search for a higher likelihood by randomly perturbing the parameters and refitting the model; alternatively, we could use a global optimization algorithm such as simulated annealing to check convergence to the global maximum (see accompanying do-file for examples).

After fitting the model, we can store the model and the estimated parameters and use `predict` to get the model predictions.

```
. estimates store c2consp

. matrix start2lc=e(b)

. predict predc, outcome(all)

. summarize predc*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
predc	10,565	.7730816	.1789941	.0323573	.9445987
predc_y1	10,565	.7083382	.2454649	-.360848	.9236923
predc_y2	10,565	.8049902	.1467324	.2261472	.9549023
predc_p1	10,565	.3301388	0	.3301388	.3301388
predc_p2	10,565	.6698612	0	.6698612	.6698612

We use the option `outcome(all)` so that in addition to the individual EQ-5D predictions (`predc`), we get the predictions for each component (`predc_y1` and `predc_y2`) and the predicted probabilities for each component (`predc_p1` and `predc_p2`). Because this model has constant probabilities of component membership, `predc_p1` and `predc_p2` are the same for all individuals and correspond to `p1` and `p2` reported in the estimation output. The means of the two components are located toward the top of EQ-5D (0.7083 and 0.8050).

In many cases, it is likely that the probabilities of the components will vary with observable characteristics. The variables may or may not be different from those used in the individual components. For simplicity, here we augment the model to include the Oxford hip score in the probabilities of component membership. We use the parameters of the constant probability model as initial values for the coefficients.

```
. matrix start = start2lc[1,1..4] , 0, start2lc[1,5..7]

. matrix list start

start[1,8]
        Comp_1:     Comp_1:     Comp_2:     Comp_2:                 Prob_C1:
          hr10       _cons        hr10       _cons         c5         _cons
y1   .30502755  -.40293118   .14801581   .22614716          0    -.70755741

          lns_1:      lns_2:
           _cons       _cons
y1    -1.2632051  -2.4541401
```

```
. aldvmm eq5d hr10, ncomponents(2) probabilities(hr10) from(start)
initial:        log likelihood =  685.78629
  (output omitted )
Iteration 10:  log likelihood =  942.71143
2 component Adjusted Limited Dependent Variable Mixture Model
                                           Number of obs    =    10,565
                                           Wald chi2(3)     =   7759.92
Log likelihood =  942.71143                Prob > chi2      =    0.0000
```

eq5d	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Comp_1						
hr10	.0887522	.0108455	8.18	0.000	.0674954	.110009
_cons	.0129138	.0269556	0.48	0.632	-.0399183	.0657459
Comp_2						
hr10	.1619008	.0019171	84.45	0.000	.1581433	.1656582
_cons	.1763049	.0075321	23.41	0.000	.1615422	.1910675
Prob_C1						
hr10	-1.395115	.0557792	-25.01	0.000	-1.50444	-1.28579
_cons	2.431909	.1748113	13.91	0.000	2.089285	2.774533
/lns_1	-1.319285	.0340906	-38.70	0.000	-1.386101	-1.252468
/lns_2	-2.271434	.012582	-180.53	0.000	-2.296095	-2.246774
sigma1	.2673265	.0091133			.2500484	.2857985
sigma2	.1031641	.001298			.1006511	.1057398

The additional parameter is significant, and the value of the likelihood function has increased considerably. We can see now that `pi1` and `pi2` no longer appear at the bottom of the table, because the probability of belonging to a component is now a function of the Oxford hip score. The probability of being in the first component decreases with the Oxford hip score. As patients show improved function and symptoms, they are less likely to be in the first component and more likely to be in the second component. Looking at the predictions below, we see that the first component has a much lower mean EQ-5D than the second component (0.352 versus 0.804), so the patients with a better Oxford hip score are also those with a better EQ-5D, as expected. There is considerable variation in the probabilities within each component, and on average, the individuals in the sample are less likely to be in the first component.

```
. estimates store c2varp
. predict predv, outcome(all)
. summarize predv*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
predv	10,565	.7687237	.1938134	.0271162	.9487147
predv_y1	10,565	.3523726	.0830315	.0140072	.4392129
predv_y2	10,565	.80352	.1552294	.1763049	.9558767
predv_p1	10,565	.0992993	.1502761	.013862	.9192284
predv_p2	10,565	.9007007	.1502761	.0807716	.986138

Information criteria can be displayed as usual:

```
. estimates stats *
Akaike´s information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
c2consp	10,565	-3737.884	685.7863	7	-1357.573	-1306.715
c2varp	10,565	.	942.7114	8	-1869.423	-1811.3

Note: N=Obs used in calculating BIC; see [R] BIC note.

Tests such as the likelihood-ratio test can also be carried out as usual:

```
. lrtest c2varp c2consp
Likelihood-ratio test                          LR chi2(1)  =    513.85
(Assumption: c2consp nested in c2varp)         Prob > chi2 =    0.0000
```

The number of components can be increased further. Of course, the analyst must exercise judgment in determining the appropriate number of components. Likelihood-ratio tests cannot be used to test models with different numbers of components because they involve testing at the edge of the parameter space ($\sigma_c = 0$), which distorts the distribution of the statistic. The Bayesian information criterion has been proposed as a useful indicator of the number of appropriate components, but other approaches also exist (McLachlan and Peel 2000).

6   Acknowledgments

7   References

Aitkin, M. 1997. Contribution to the discussion of paper by S. Richardson and P. J. Green. *Journal of the Royal Statistical Society, Series B* 59: 764–768.

Dolan, P., C. Gudex, P. Kind, and A. Williams. 1995. A social tariff for EuroQol: Results from a UK population survey. Discussion Paper No. 138, University of York, Centre for Health Economics. http://www.york.ac.uk/che/pdf/DP138.pdf.

Hernández Alava, M., A. Wailoo, F. Wolfe, and K. Michaud. 2013. The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. *Rheumatology* 52: 944–950.

———. 2014. A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. *Medical Decision Making* 34: 919–930.

Hernández Alava, M., A. J. Wailoo, and R. Ara. 2012. Tails from the peak district: Adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health* 15: 550–561.

McLachlan, G., and D. Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Shaw, J. W., J. A. Johnson, and S. J. Coons. 2005. US valuation of the EQ-5D health states: Development and testing of the D1 valuation model. *Medical Care* 43: 203–220.

Wineberg, A. 2014. *Finalised Patient Reported Outcome Measures (PROMs) in England: April 2011 to March 2012, Version V2.0.* Leeds, UK: Health and Social Care Information Centre. http://www.hscic.gov.uk/catalogue/PUB11359/final-proms-eng-apr11-mar12-fin-report-v2.pdf.

About the authors

Mónica Hernández Alava is an applied microeconometrician in the Health Economics and Decision Science section in ScHARR, University of Sheffield, Sheffield, UK.

Allan Wailoo is a health economist in the Health Economics and Decision Science section in ScHARR, University of Sheffield, Sheffield, UK.