The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–13

ALDVMM: A command for 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. This article describes the `aldvmm` Stata command for fitting an Adjusted Limited Dependent Variable Mixture Model to either UK or US tariff EQ-5D data. The command and post estimation options are presented and explained through examples. The `aldvmm` command requires Stas Kolenikov's Simulated Annealing package (`simann`) which can be easily installed by typing `help aldvmm` after installation and clicking on the link provided.

Keywords: st0001, aldvmm, Adjusted Limited Dependent Variable Mixture, EQ-5D, EQ-5D-3L mapping

1 Introduction

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

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

© *yyyy* StataCorp LP

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

The article is organised as follows: section 2 gives a brief overview of the adjusted limited dependent variable mixture model, section 3 describes the `aldvmm` syntax and options including the syntax for `predict`, and section 4 shows some examples.

2 Adjusted Limited Dependent Variable Mixture Model

The distribution of EQ-5D exhibits several characteristics that need to be taken into account when estimating "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. This value is -0.109 in the US tariff. Death is attached a value of 0 and therefore there are a few health states described by EQ-5D which are considered worse than death. There tends to be a mass of observations at the upper limit (1). However, the standard tobit model is not appropriate for several reasons. First there is a large gap between the mass at one and the next feasible EQ-5D value (0.883 and 0.860 for the UK and US tariffs respectively). In addition, the rest of the distribution tends to show strong bimodality, often with a high degree of skewness. These characteristics tend to remain even after conditioning.

The Adjusted Limited Dependent Variable Mixture Model (Hernández-Alava et al. 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 tobit-like distributions. A brief description of the model follows. A more detailed description and other variants can be found in Hernández-Alava et al. (2012, 2013, 2014).

It is assumed that EQ-5D (denoted by y_i) can be modelled as a mixture of C–components or classes. Conditional on an individual observation i belonging to component c ($c = 1, ..., C$), 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 $\Psi_1 = 0.883$, $\Psi_2 = -0.594$ for the UK and $\Psi_1 = 0.860$, $\Psi_2 = -0.109$ for the US. For each mixture component c

$$y_i^*|c = x_i^{'}\beta_c + \varepsilon_{ic} \tag{2}$$

β_c is a $(k \times 1)$ vector of coefficients including an intercept term, $x_i^{'}$ is a row vector of covariates, ε_{ic} is IID $N\left(0, \sigma_c^2\right)$. A multinomial logit model for the probability of latent

class membership is assumed as follows:

$$P\left(c|w_i^{'}\right) = \frac{\exp\left(w_i^{'}\delta_c\right)}{\sum_{s=1}^{C}\exp\left(w_i^{'}\delta_s\right)} \tag{3}$$

where $w_i^{'}$ is a vector of variables that affect the probability of component membership, δ_c is the vector of corresponding coefficients and C is the number of classes used in the analysis. One set of coefficients δ_c is normalised to zero for identification. If no variables are included, then the probabilities of component membership are constant for all individuals.

The loglikelihood of the model defined by equations 1, 2 and 3 can be written as

$$\begin{aligned}\ln l \quad = \quad & \sum_{i=1}^{n}\ln\left\langle\sum_{c=1}^{C}\frac{\exp\left(w_i^{'}\delta_c\right)}{\sum_{s=1}^{C}\exp\left(w_i^{'}\delta_s\right)}\left\{\mathbf{1}\left(y_i>\Psi_1\right)\left[1-\Phi\left(\frac{\Psi_1-x_i^{'}\beta_c}{\sigma_c}\right)\right]+\right.\right. \\ & +\mathbf{1}\left(y_i\leq\Psi_2\right)\left[\Phi\left(\frac{\Psi_2-x_i^{'}\beta_c}{\sigma_c}\right)\right]+ \\ & \left.\left.+\mathbf{1}\left(\Psi_2<y_i<\Psi_1\right)\left[\frac{1}{\sigma_c}\phi\left(\frac{y_i-x_i^{'}\beta_c}{\sigma_c}\right)\right]\right\}\right\rangle\end{aligned} \tag{4}$$

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

After estimating the model one can use the conditional expectation below to predict EQ-5D:

$$\begin{aligned}E\left(y_i|x_i^{'}w_i^{'}\right) \quad = \quad & \sum_{c=1}^{C}\frac{\exp\left(w_i^{'}\delta_c\right)}{\sum_{s=1}^{C}\exp\left(w_i^{'}\delta_s\right)}\left\{\left[1-\Phi\left(\frac{\Psi_1-x_i^{'}\beta_c}{\sigma_c}\right)\right]+\right. \\ & +\left[\Phi\left(\frac{\Psi_2-x_i^{'}\beta_c}{\sigma_c}\right)\right]\Psi_2+ \\ & \left.+\left[\Phi\left(\frac{\Psi_1-x_i^{'}\beta_c}{\sigma_c}\right)-\Phi\left(\frac{\Psi_2-x_i^{'}\beta_c}{\sigma_c}\right)\right]\left[x_i^{'}\beta_c+\sigma_c\frac{\phi\left(\frac{\Psi_1-x_i^{'}\beta_c}{\sigma_c}\right)-\phi\left(\frac{\Psi_2-x_i^{'}\beta_c}{\sigma_c}\right)}{\Phi\left(\frac{\Psi_2-x_i^{'}\beta_c}{\sigma_c}\right)-\Phi\left(\frac{\Psi_1-x_i^{'}\beta_c}{\sigma_c}\right)}\right]\right\}\end{aligned}$$

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

3 Command syntax

3.1 aldvmm

Syntax

`aldvmm` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] , `ncomponents(#)` [
`probabilities(`*varlist*`)` `country(`*country*`)` `llim(#)` `ulim(#)`
`constraints(`*numlist*`)` `vce(`*vcetype*`)` `level(#)` `inimethod(`*inimethod*`)`
`saopts(`*matrix*`)` *maximize_options*]

Description

`aldvmm` is a user-written program which fits an adjusted limited dependent variable mixture model using maximum likelihood estimation. It is implemented as a `l1 ml` evaluator. 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 can be constrained to be the same via the `constraints` option.

Options

`ncomponents(#)` is required and specifies the number of mixture components. Strictly a mixture model has a minimum of 2 components, however the command does allow the estimation of a model with only one component. This one component model is similar to a tobit model but is able to reflect the gap found in EQ-5D.

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

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

`llim(#)` user supplied lower limit of EQ-5D (Ψ_2). If `llim(#)` is used `ulim(#)` must also be provided.

`ulim(#)` user supplied highest EQ-5D index value below 1 (Ψ_1). Setting #to 1 estimates a model without a gap, that is, a mixture of tobit models. If `ulim(#)` is used `llim(#)` must also be provided.

`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`, `robust` or `cluster`. 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 the parameters. The string *inimethod* may be `single`, `cons` or `simann`. The default is `single` which lets `ml` find starting values. Using `cons` fits first a constant only model and uses those parameters as starting values in the estimation of the full model. Specifying `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 where *matrix* is the name of the matrix with the following `simann` arguments: (*count*, *ftol*, *steps*, *cooling*, *start*, *loglevel*).

maximize_options: difficult, technique(*algorithm_spec*), iterate(#), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), gtolerance(#), nrtolerance(#), nonrtolerance, from(*init_specs*); see [R] **maximize**

3.2 predict

Syntax

`predict` *varname* [*if*] [*in*] [, `outcome(`*outcome*`)`]

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.

Options

`outcome(`*outcome*`)` specifies the predictions to be stored. There are two options for *outcome* `y` or `all`. The default, `y`, stores only the dependent variable prediction in *newvar*. Use `all` to, in addition, 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.

4 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 (PROMs) in England April 2011 to March 2012 (Health and Social Care Information Centre). The data is freely available and can be downloaded from http://www.hscic.gov.uk/catalogue/PUB11359. For the purpose of this example we select a 30% random sample of individuals who have data on age and gender (age and gender are excluded from the dataset for those patients that could be identified due to low numbers). We use post-operative 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 a single 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 symptomas and a 0 score attached to the greatest severity. The individual scores are then added together to a single score with 0 denoting the worst possible symptoms and function and 48 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.

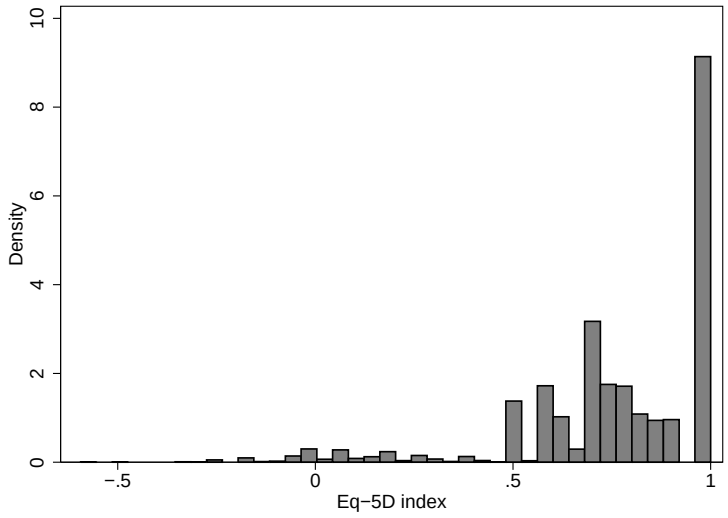

Figure 1: Histogram of EQ-5D data

Mixture models are extremely flexible and are a convenient semiparametric way to model data showing characteristics not easily accomodated by known distributions. Mixtures of normal distributions can generate multimodality, strong skewness in a unimodal distribution, 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 and/or peaks.

We recommend the reader to familiarise herself/himself with the idiosyncracies of fitting mixture models (McLachlan and Peel 2000) before attempting to estimate one. We will describe briefly here the two main issues that researchers trying to estimate models of EQ-5D are likely to encounter. One of the problems of fitting mixture models relates to the presence of several local maxima in the likelihood function. One cannot assume that by running the model and getting some estimated parameters, the consis-

tent solution has been found. To identify the global maximizer, at the very least, it is important to try different sets of random starting values and to select the solution with the highest likelihood function. Alternatively, a global optimisation algorithm such as simulated annealing can be used. The `aldvmm` command can use Stas Kolenikov `simann` command for simulated annealing. We recommend using this option only when fitting a small number of components as it could be time-consuming and since it is not able to restrict the parameter space can run more easily into the issue described next. A second problem arises when estimating mixtures with different σ_c across components as 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 but the likelihood contribution of that component becomes infinite in equation 4 because we are dividing by a very small number. In this situation, we cannot trust the value of the likelihood. In most cases, provided certain regularity conditions are met, the consistent solution will correspond to a local maximizer. EQ-5D data has in most cases a mass of observations at 1 corresponding to individuals who are in full health and no immediately adjacent observations. If we try to estimate a standard mixture of normal distributions we will run into problems of unbounded likelihoods very quickly 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 one of the components becomes a probability mass at 1. 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.

It is important when fitting mixture models to start with simple models with a small number of components and use them as a stepping stone to estimate 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 2-component model.

```
. aldvmm eq5d hr10, ncomp(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   =      10565
                                                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	885110.9	.	.	.	.	.
_cons	3930832	.	.	.	.	.
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	-164.8169	.	.	.	.	.
sigma1	.1927987	.0017163			.189464	.196192
sigma2	2.64e-72	.			.	.
pi1	.9993387	.0004834			.9972325	.9998422
pi2	.0006613	.0004834			.0001578	.0027675

The output signals that something is wrong. The constant and the estimated coefficient 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 translates into a probability mass at 1. The likelihood of the model is reliable in this case and the missing standard errors only signal that the parameters are not identified as 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, ncomp(2) from(a) c(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   =      10565
                                                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 estimated 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 ones 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 there is a risk that the usual local maximization algorithms get stuck at a local maximum. It is important when using these models to use a range of starting values to ascertain that the global maximum has been found. Before embarking on a search it is worth first taking advantage of some of the options that have been programmed in the `aldvmm` command. One option that sometimes works well is to estimate a constant only model first and use the estimated parameters in the full model specification. This can be accomplished using the `inim(cons)` option of the `aldvmm` command.

```
. aldvmm eq5d hr10, ncomp(2) inim(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   =      10565
                                                  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 only to a local solution. In this solution we find that all parameters are significant and the two components have now sizeable associated probabilities. In both components EQ-5D increases as the Oxford Hip Score increases but the size of the parameters is quite different. Based on these parameters we could do a further search for a higher likelihood by randomly perturbing the parameters and re-estimating the model or, alternatively, we could use a global optimisation algorithm such as simulated annealing to check convergence to the global maximum (see accompanying do file for examples).

After estimating 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)
. sum predc*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
predc	10565	.7730816	.1789941	.0323573	.9445987
predc_y1	10565	.7083382	.2454649	-.360848	.9236923
predc_y2	10565	.8049902	.1467324	.2261472	.9549023
predc_p1	10565	.3301388	0	.3301388	.3301388
predc_p2	10565	.6698612	0	.6698612	.6698612

We use the option `outcome(all)` so that in addition to the individual EQ-5D predictions (predc) we also get the predictions for each component (predc_y1 and predc_y2) and the predicted probabilities for each component (predc_p1 and predc_p2). Since 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 towards the top end 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 to 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:      lns_1:
          hr10       _cons        hr10       _cons          c5       _cons       _cons
y1   .30502755  -.40293118   .14801581   .22614716           0  -.70755741  -1.2632051
         lns_2:
         _cons
y1  -2.4541401
. aldvmm eq5d hr10, ncomp(2) prob(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   =      10565
                                                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 since 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, the first component has a much lower mean EQ-5D than the second component (0.352 vs 0.804) so that 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)
. sum predv*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
predv	10565	.7687237	.1938134	.0271162	.9487147
predv_y1	10565	.3523726	.0830315	.0140072	.4392129
predv_y2	10565	.80352	.1552294	.1763049	.9558767
predv_p1	10565	.0992993	.1502761	.013862	.9192284
predv_p2	10565	.9007007	.1502761	.0807716	.986138

Information criteria can be displayed in the usual way:

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

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
c2consp	10565	-3737.884	685.7863	7	-1357.573	-1306.715
c2varp	10565	.	942.7114	8	-1869.423	-1811.3

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

and tests such as the Likelihood Ratio test can also be carried out in the usual way:

```
. 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 judgement in determining the appropriate number of components. Likelihood ratio tests cannot be used to test models with different number of components because it involves 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)

5 Acknowledgements

This work was supported by two funding sources the Medical Research Council under grant MR/L022575/1 and the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this article are of the authors only. The authors also wish to thank David Trueman for helping to test the code, and an anonymous referee for helpful suggestions.

6 References

Aitkin, M. 1997. Contribution to the discussion of paper by S. Richardson and P.J. Green. *Journal of the Royal Statistical Society 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*. University of York, Centre for Health Economics: Discussion Paper 138.

Hernández-Alava, M., A. Wailoo, and R. Ara. 2012. Tails from the peak district: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Health State Utility Values. *Value in Health* 15(3): 550–561.

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

McLachlan, G. J., and D. Peel. 2000. *Finite mixture models*. New York: Wiley.

Shaw, J. W., J. A. Johnson, and S. Coons. 2005. US Valuation of the EQ-5D Health States Development and Testing of the D1 Valuation Model. *Medical Care* 43(3): 203–220.

Wineberg, A. 2014. Finalised Patient Reported Outcome Measures (PROMs) in England April 2011 to March 2012 v2.0. Technical report, Health & 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, UK.

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