Paper presented at the joint CES/HESG conference, Paris, January 2004
Work in progress

Estimating population cardinal health state valuation models from individual ordinal (rank) health state preference data.

Christopher McCabe[†‡], John Brazier[†‡], Peter Gilks[‡], Aki Tsuchiya[†], Jennifer Roberts[†],

Anthony O'Hagan[‡Φ]

Sheffield Health Economics Group[†]
Centre for Bayesian Statistics in Health Economics[‡]
Department of Probability and Statistics[Φ]
University of Sheffield

Address for correspondence:

Christopher McCabe
Senior Lecturer in Health Economics
Sheffield Health Economics Group, ScHARR
University of Sheffield
Regent Court
31 Regent Street
Sheffield
S1 4DA
Email: c.mccabe@sheffield.ac.uk

Tel: 00 (44) 114 222 5454

Abstract

Ranking exercises have routinely been used as warm-up exercises within health state valuation surveys. Very little use has been made of the information obtained in this process. Instead, research has focussed upon the analysis of health state valuation data obtained using the category rating, standard gamble and time trade off methods.

Thurstone's law of comparative judgement, proposes a stable relationship between ordinal and cardinal preferences, based upon the information provided by pairwise choices. McFadden proposed that this relationship could be modelled by estimating conditional logistic regression models where alternatives had been ranked. In this paper we report the estimation of such models for the Health Utilities Index Mark 2 and the SF-6D. The results are compared to the conventional regression models estimated from standard gamble data, and to the observed mean standard gamble valuations.

For both the HUI2 and the SF-6D, the models estimated using rank data are broadly comparable to the models estimated on standard gamble data and the predictive performance of these models is closer to the standard gamble models than we would have expected. Our research indicates that rank data has the potential to provide useful insights into community health state preferences, however important questions remain to be addressed.

Introduction

As cost effectiveness analysis has become more important in health care decision making processes, the interest in how to value health outcomes has increased. There is a substantial body of research on the relative strengths and weaknesses of alternative methods. [1,2,3,4] Such research has focused primarily on three valuation methods; Time Trade Off (TTO); Standard Gamble (SG) and Visual Analogue Scales (VAS), also called category scaling.

Work that has attempted to identify a preferred method has tended to support the use of Time Trade Off and/or Standard Gamble.[2,5] VAS has been criticised on a number of points, both theoretical (does VAS capture strength of preference) and empirical (the data may be subject to end-point and context bias).[6,7] However, it is widely accepted that TTO and SG have significant limitations.[2] What is notable, is the degree to which the role of ordinal data in health state valuation has been largely ignored; notable exceptions to this observation being the work by Kind.[8,9]

Ranking exercises are conventionally included in health state valuation interviews as a warm-up exercise, in order to familiarise the interviewee with the health state classification system being valued and thinking about preferences between hypothetical health states.[10] The use of the data from these ranking exercises has generally been limited to checking the degree of consistency between the valuations obtained from the SG or TTO valuation exercises and the ranking exercise.

Kind identified Thurstone's model of comparative judgement as a potential theoretical basis for deriving cardinal preferences from rank preference data. [8] Thurstone's method considers the proportion of times that health state *i* is considered worse than health state *j*. The preferences over the health states represent a latent utility function, and the likelihood of health state I being ranked above health state j, when health state j is actually preferred to health state *i*, is a function of how close to each other the states lie on this latent utility function. Therefore, choice data provide information about the latent utility function. McFadden proposed the conditional logistic regression model as a means of modelling this latent utility function from ordinal data.[11]

Recently Salomon presented work that applied conditional logistic regression models to the rank data collected as part of the Measurement and Valuation of Health Study (MVH).[12] Salomon estimated a model equivalent to that reported by Dolan.[13] The model did not produce utilities on the 0-1 scale necessary for use in estimating Quality Adjusted Life Years. Salomon rescaled the model coefficients on to the full health-death (1-0) scale, using the mean measured TTO value for the PITS state in the EQ-5D classification (3,3,3,3,3). In this paper we present an approach that avoids the need for external health state value data, for such rescaling, by directly estimating a parameter for the state death, as part of the model. This method is applied to rank data from two health state valuation surveys; a UK based valuation survey for the Health Utilities Index Mark 2, and the UK valuation survey for the SF-6D. [14] [15]

Methods

Model specification

To model the predicted health state valuations using the ordinal preference data we used conditional logistic regression as outlined by McFadden.[11] To use this model it is necessary to assume that the ranking exercise is equivalent to the respondent making a series of individual selections from smaller and smaller groups. Thus, in ranking 10 health states we assume that the respondent first chooses the most preferred health state from all 10, before choosing the most preferred health state from the remaining 9 and so on, until all the health states have been assigned a rank between 1 and 10. To characterise this as equivalent to pair wise choice we must rely on the hypothesis of the independence of irrelevant alternatives.

The conditional logistic regression model assumes that respondent i has a latent utility value for state j, U_{ij}, and that given the choice of two states j and k the respondent will choose state j over state k if $U_{ij} > U_{ik}$. Hence given the task of choosing the preferred state from a finite group of different states, respondent i will choose state j if $U_{ij} > U_{ik}$ for all $k \neq j$.

Each individual's utility function for state j is $U_{ij} = \mu_j + \varepsilon_{ij}$ where μ_j is representative of the tastes of the population and ε_{ij} represents the particular taste of the individual. If the error term ε has an extreme value distribution, then the odds of choosing state j over state k are $exp\{\mu_j - \mu_k\}$.

For the analyses reported here, the expected value of each unobserved utility was assumed to be a linear function of the categorical levels on the domains of each dataset respectively. The model specification is:

$$\mu_{ij} = g(\boldsymbol{\beta}'\mathbf{x}_{\mathbf{ij}} + \theta D + u_{ij})$$

where i = 1, 2, …, n represents individual health state values and j = 1,2, …, m represents respondents. g is a function specifying the appropriate functional form, which is assumed here to the linear. u_{ij} is an error term whose autocorrelation structure and distributional properties depend on the assumptions underlying the particular model used.

x is a vector of dummy explanatory variables (x_{ij}) for each level *i* of dimension *j* of the instrument in question. SF-6D. For example for the SF-6D, x_{31} denotes dimension i = 3 (social functioning), level j = 1 (health limits social activities none of the time). For any given health state, x_{ij} will be defined as

x_{ij} = 1 if, for this state, dimension i is at level j

x_{ij} = 0 if, for this state, dimension i is not at level j

Level 1 is the baseline for each dimension.

D is a dummy variable for the state 'Death', which takes the value 1 for this health state. For all other health states the variable Death is always set at 0.

In these models, the dummy variables take the value 1 if the health state has that attribute at that level and 0 otherwise. Level 1 is the baseline for each dimension. For the state 'Death', only the dummy variable Death takes value 1, all other dummy variables take value 0. For all other health states the variable Death is always set at 0.

Rescaling model coefficients on to the death-full health (1-0) scale

The scale of the latent variable μ is arbitrarily defined by the identifying assumptions in the model. For QALY calculations it is necessary to have health state values on the scale where death is valued at zero and full health is valued at 1. Therefore, we rescaled the coefficients using the formula $\beta_{ri} = \beta_i / |\beta_d|$, where β_{ri} is the rescaled coefficient of β_i and β_d is the coefficient of death.

These rescaled coefficients provide predictions for health state values on the same scale as standard gamble or time trade off valuations. This method of rescaling anchors death at zero and full health at 1, whilst retaining the possibility of a health state having a value of <0; i.e. worse than death.

Model Assessment

Models are assessed in a number of stages. The first stage checks that the estimated model coefficients have the expected negative sign and that they are statistically significant. These coefficients are then rescaled on to the full health-death (1-0) scale and the rescaled coefficients are checked for logical inconsistencies; i.e, that lower levels of functioning are associated with greater decrements in health state value.

The rescaled coefficients are then compared to the coefficients from the preferred models estimated on the standard gamble data from the same valuation interviews.[15 16]

The predictive performance of the models is assessed using the following battery of measures:

- Root Mean Square Error (RMSE),
- Mean Absolute Error (MAE),
- Proportion of health state values predicted to within 0.05 of the observed mean of the standard gamble valuations
- Ljung-Box test for autocorrelation in the errors[17]
- Correlation between predicted and observed mean values

In addition we plot the health state values predicted by the models against the observed mean SG values and the values predicted by the original SG models. We also plot the errors against the observed mean values.

We report model coefficients, significance levels, diagnostic plots and tests of predictive performance for both the HUI2 and the SF6D models.

Surveys

Descriptions of both of these surveys have been reported elsewhere, thus, we will only provide a brief summary of the two valuation surveys here.

Health Utilities Index Mark 2

The Health Utilities Index Mark 2 is a six dimension health state classification (sensation, mobility, emotion, cognition, self care and pain) with either four or five levels for each dimension. It describes a total of 8,000 distinct health states. It was developed specifically for use with paediatric populations.[14]

One hundred and ninety eight respondents ranked 8 health states from the HUI2 classification plus Full Health and Immediate Death. The health states valued were sampled from a orthogonal array for the HUI2 classification. The interviewees then valued the same 8 health states using the McMaster version of the Standard Gamble question; i.e. the chance board prop was used to aid the respondent understand the question. The risk of death was varied in a ping-pong manner until the interviewee identified a risk of death at which they were indifferent between the impaired health state and the uncertain choice. Where health states were ranked as worse than immediate death, the worse than death version of the Standard Gamble question was used.[10]

The respondent was asked to imagine that they were a ten year old child who would live for another 60 years in the outcome health state.

SF-6D

The SF-6D has 6 dimensions: Physical functioning, Role Limitations, Social Functioning, Pain, Mental Health and Vitality. Each dimension has either 5 or 6 levels. The classification describes a total of 18,000 health states.[15]

A representative sample of 611 members of the UK population provided standard gamble valuations for a sample of 249 health states defined by the SF-6D classification.

The interview consisted of an exercise to rank 5 health states that they would be asked to value, plus the best and worst states defined by the SF-6D and immediate death. This was followed by a series of standard gamble questions. The standard gamble question asked the respondent to value one of 5 certain SF6D health states against the best and 'pits' health state. All respondents were then asked to provide a standard gamble valuation of the PITS state in relation to death. The form of the sixth standard gamble valuation depended upon whether the respondent has ranked the PITS state as better or worse than death, in the ranking exercise. The result of the sixth standard gamble exercise was then used to 'chain' the health state values in order to place them on to the 1-0, full health –death scale, required for the calculation of QALYs. The interviewers used the chance board prop and the ping-pong version of the standard gamble question.

The respondent was asked to answer the question for himself or herself, imagining that they would remain the outcome health state for the rest of their lives.[18]

Results

Health Utilities Index Mark 2

Table 1 reports the original and rescaled coefficients for the rank health state value models for the HUI2. It also gives the results for each of the diagnostic tests. For comparative purposes the same information is provided for the standard gamble health state valuation model.[15]

The similarity of the rank and standard gamble data models is quite striking. The rank model has marginally more inconsistencies than the standard gamble model, and does not distinguish as clearly between the different levels on the mobility dimension. However, this dimension is one of the weaker dimensions in the standard gamble model. Generally the decrement for the lowest levels of functioning is greater in the SG than the rank model. The predictive performance of the two models is closer than we would have expected given the difference in the level of information the two models were estimated from. This said the SG model does perform better than the rank model on all tests.

Figure 1 plots the observed and modelled values health state values, and the predictions errors. Figure 2 provides the same information for the Standard Gamble model. The plots confirm the similarity of the predictive performance of the rank and standard gamble models.

SF-6D

Table 2 reports the original and rescaled coefficients for the rank health state value models for the SF-6D. It also gives the results for each of the diagnostic tests. For comparative purposes the same information is provided for the standard gamble health state valuation model.

The Rank data model is quite different from the standard gamble model. It is notable that the number of inconsistencies is lower in the rank data model than the standard gamble model. Whilst there are inconsistencies in the coefficients for role physical, in both models, there are fewer in the rank model than the standard gamble model. The coefficients on the role and social functioning dimension are quite similar in both models, although the lowest level of functioning on the social role dimension has a considerably larger decrement in the rank model. The rank model also gives much greater weight to reductions on the pain dimension. The coefficients on the mental health dimensions are very similar in the two models. The vitality dimension in the standard gamble model has a number of inconsistencies, the rank model by contrast has none. This said, the predictive performance of the rank model is slightly worse than the standard gamble model, for most tests. The notable exception to this is the LB test. The LB test results suggest that the relationship between prediction error and observed health state value is less strong for the rank model than the standard gamble model.

Figure 3 plots the observed and modelled values health state values, and the predictions errors. Figure 4 provides the same information for the Standard Gamble model.

Discussion

In this paper we have reported the estimation of <u>population</u> cardinal health state valuation models for the HUI2 and the SF-6D, from individual ordinal preference data. In both cases the models bare comparison to the health state valuation models estimated from standard gamble (cardinal) data provided by the same respondents as the ordinal data. Although in both cases, the SG models are superior.

The impetus for this research was an analysis of rank data for the EQ-5D, presented by Solomon. It is notable that the degree of agreement between the rank model and the TTO model for the EQ-5D is considerably less than we report for the rank and standard gamble models we have estimated.

Our apparent success in estimating cardinal health state valuation models from ordinal data raises many questions. In describing our results as a success, we are assuming that the standard gamble data are the appropriate standard by which to judge these models. It is arguable that our results say as much about the limitations of standard gamble data as they do about the existence or otherwise of a latent utility function. Research is required to examine whether respondents expressed preferences are consistent with the models that are derived from the standard gamble (and Time Trade Off) values they provide. Such work is likely to be require qualitative rather than quantitative methods.

We have modelled the relationship between the standard gamble values and the rank values as being linear. There is no reason why this should be so. The ranking exercise does not involved risk, whilst the standard gamble explicitly incorporates risk into the valuation process. Standard models of risk attitude would suggest that a linear model would not be the best functional form.[19] Future work should look at the performance of alternative functional forms. Theoretical perspectives on the relationship between rank and standard gamble data should inform such research.

The application of the clustered logistic regression model requires that the rank data exercise be characterised as a sequential choice process. Whilst we believe that this assumption is defensible, we accept that other models of the ranking process are equally plausible. There is an increasing body of research suggesting that respondents apply decision heuristics to complex choice scenarios, and that lexicographic preferences are common in contingent valuation studies. Research on the thought processes of individual's undertaking ranking exercises would be a valuable contribution to this field. An alternative means of addressing this issue would be to design the ranking exercise to ensure consistency with the underlying assumptions of the model. Thus the respondent would be presented with the all health states to be ranked and asked to identify the highest ranked health state. This would be recorded and then the respondent would be presented with the remaining health states and again asked to identify the highest ranked health state from that set. This process would be repeated until all the states had been ranked. Work to establish the feasibility of undertaking this type of valuation exercise and to compare the results with those from the ranking exercises presented here would be interesting.

This work assumes that the rank data are preference data. The literature on health state preference elicitation has generally argued that Visual Analogue Scale data are not preferences because the valuation process does not require the respondent to trade. This same observation can be applied to ranking exercises. If rank data are reflecting an underlying utility function the utility functions may reflect Broome's concept of the relative 'goodness' of different health states, rather than the conventional expected utility, that the standard gamble is designed to measure.[20]

The analyses assume that the information content of the rank is unaffected by the order of the rank or indeed the number of states to be ranked. Hausman and Ruud have hypothesised that respondents may take more care with the initial ranking exercises than the later ones.[21] Thus the risk of ranking being incorrect would be systematically related to a health states position in the rank; i.e. the assumption of independence of irrelevant alternatives would not hold. Work is on-going to test this hypothesis in these datasets.

Should future research confirm the promise of ordinal data to support the modelling of cardinal health state preferences, it is by no means clear what the implications for future health state valuation work would be. It may be that ranking data may make it possible to incorporate the views of populations for whom the TTO and SG procedures are felt to be too arduous e.g. younger children.[22] However, the ranking tasks themselves are not simple and no research to date has examined children's ability to understand them.

An alternative benefit may be that the future valuation surveys may require fewer resources. In addition, ranking exercises may be more feasible in postal interviews than TTO and SG, again allowing more efficient implementation of health state valuation surveys. It might be that rank data offers the convenience of the visual analogue scale without the problems of context and end-point bias.[7]

The results raise questions about the relationship between discrete choice experiments and the conventional methods of obtaining health state preferences for calculating QALYs. The format of the discrete choice question fits more immediately within the comparative judgement framework than the ranking exercises described above. It seems reasonable to expect that discrete choice scenarios that included a dimension for alive/dead, would be suitable data sources for a similar modelling strategy to that described in this paper.

Summary

In this paper we have presented two models of cardinal health state preferences based upon ordinal health state preference data; one for the SF-6D health state classification, the other for the HUI2 health state classification. We have compared these models to models previously estimated on using standard gamble valuations, in terms of the degree of accuracy and bias in predicting mean observed standard gamble health state valuations in the estimation samples.

The ordinal rank models perform much better than might have been expected given the difference in the informational content between the standard gamble and ranking exercises. This said, the SG models were better than the ordinal models on the tests we applied.

The results are consistent with Thurstone’s law of comparative judgement, and the existence of a latent utility function. They justify further research on the potential for ordinal health state valuation data to reflect cardinal population preferences and suggest there is potential for research on the potential for discrete choice experiments to provide health state preference data on the Full health-death scale.

Table 1: Ordinal and OLS Health State Valuation Models for HUI2[1]

HUI2 Rank Model and SG Model(OLS)			
	RankCoeff	RescaledCoeff	SGCoeff
sens2	-0.9932932	-0.1156	-0.1151
sens3	-0.9350973	-0.1089	-0.1223
sens4	-2.116679	-0.2464	-0.2253
mobil2	-0.7287155	-0.0848	-0.0516
mobil3	-0.9887335	-0.1151	-0.1224
mobil4	-0.8041412	-0.0936	-0.1308
mobil5	-1.008526	-0.1174	-0.1103
emot2	-0.8122273	-0.0946	-0.0945
emot3	-1.0001	-0.1164	-0.1119
emot4	-1.429127	-0.1664	-0.1801
emot5	-1.43784	-0.1674	-0.1824
cogn2	-0.3222758	-0.0375	-0.0567
cogn3	-0.5438438	-0.0633	-0.0966
cogn4	-0.773194	-0.0900	-0.1676
sc2	-0.4409409	-0.0513	-0.0516
sc3	-0.692351	-0.0806	-0.1138
sc4	-0.7762394	-0.0904	-0.1158
pain2	-0.8131845	-0.0947	-0.1114
pain3	-0.940143	-0.1095	-0.1155
pain4	-1.216913	-0.1417	-0.1626
pain5	-1.76543	-0.2055	-0.2538
death	-8.589516	-1	
n states		51	51
MAE		0.062	0.051
No.>0.05		23	18
No.>0.10		12	5
RMSE		0.0775	0.0657
LB		36.11	25.78
Corr(means)		0.8814	0.921
No. of Logical Inconsistencies		2	1

[1] All coefficients for both models were significant at the p<0.1.

Table 2: Ordinal and Standard Gamble Health State Valuation Models for SF-6D[2]

SF6D Rank Model and SG model(Mean'6')			
	RankCoeff	RescaledCoeff	SGCoeff
pf2	**-0.363575**	**-0.0566**	-0.0532
pf3	**-0.431302**	**-0.0671**	-0.0106
pf4	**-0.9856325**	**-0.1534**	**-0.0402**
pf5	**-0.6340183**	**-0.0987**	**-0.0535**
pf6	**-1.447536**	**-0.2253**	**-0.1110**
rl2	**-0.3210761**	**-0.0500**	**-0.0530**
rl3	**-0.4069154**	**-0.0633**	**-0.0552**
rl4	**-0.4052777**	**-0.0631**	**-0.0503**
sf2	**-0.3626836**	**-0.0565**	**-0.0555**
sf3	**-0.4203095**	**-0.0654**	**-0.0668**
sf4	**-0.5737133**	**-0.0893**	**-0.0698**
sf5	**-0.8054821**	**-0.1254**	**-0.0866**
pain2	**-0.377161**	**-0.0587**	**-0.0467**
pain3	**-0.3635335**	**-0.0566**	-0.0250
pain4	**-0.6520135**	**-0.1015**	**-0.0561**
pain5	**-0.8187383**	**-0.1275**	**-0.0912**
pain6	**-1.191158**	**-0.1854**	**-0.1669**
mh2	**-0.2157184**	**-0.0336**	**-0.0490**
mh3	**-0.3371096**	**-0.0525**	**-0.0424**
mh4	**-0.7015521**	**-0.1092**	**-0.1092**
mh5	**-0.8992905**	**-0.1400**	**-0.1279**
vit2	**-0.173969**	**-0.0271**	**-0.0861**
vit3	**-0.2139943**	**-0.0333**	**-0.0606**
vit4	**-0.3226131**	**-0.0502**	**-0.0543**
vit5	**-0.5267463**	**-0.0820**	**-0.0907**
death	**-6.423983**	**-1.0000**	
n states		249	249
MAE		0.0882	0.0742
No.>0.05		169	118
No.>0.10		84	51
RMSE		0.1096	0.0976
LB		106.7200	169.5700
Corr(means)		0.7111	0.7377
No. of logical inconsistencies		3	8

[2] Coefficients in bold are significant at p<0.1

Figure 1: Observed and predicted health state values for the HUI2 rank model

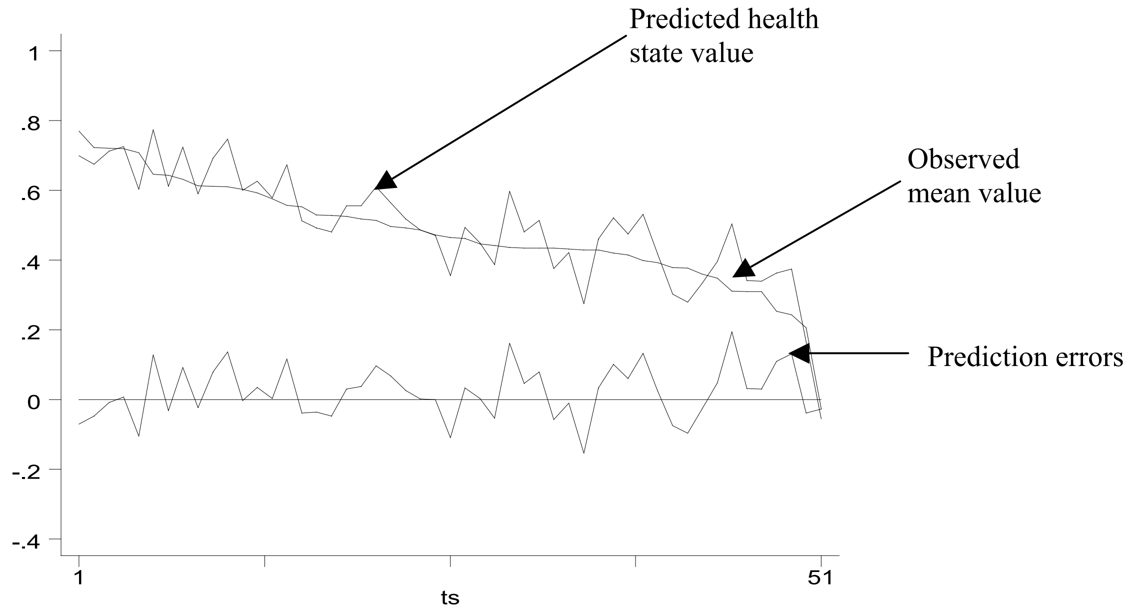


Figure 2: Observed and predictive health state values for HUI2 Standard Gamble model

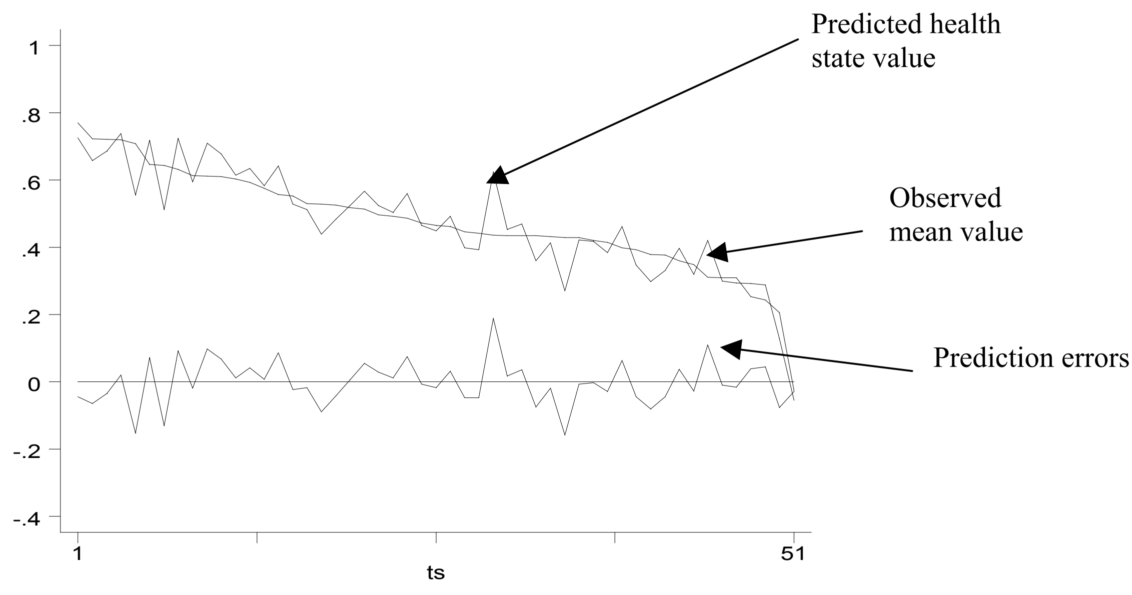


Figure 3: Observed and predicted health state values for the SF-6D rank model

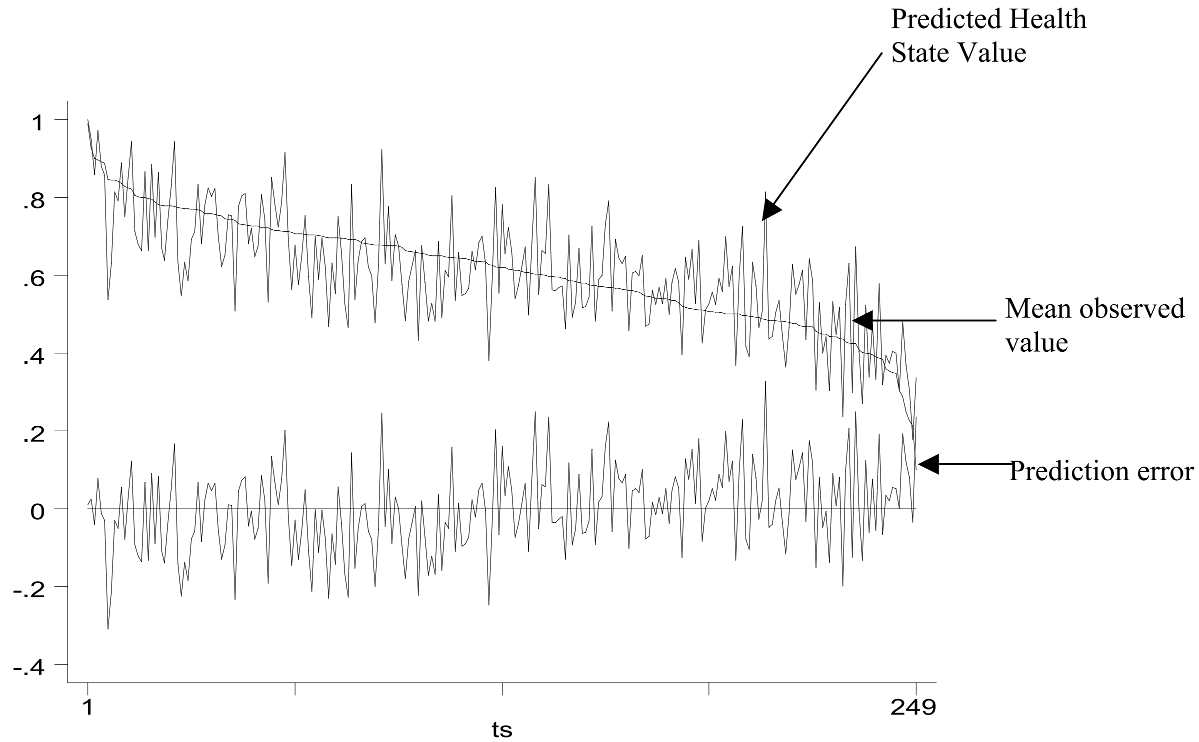
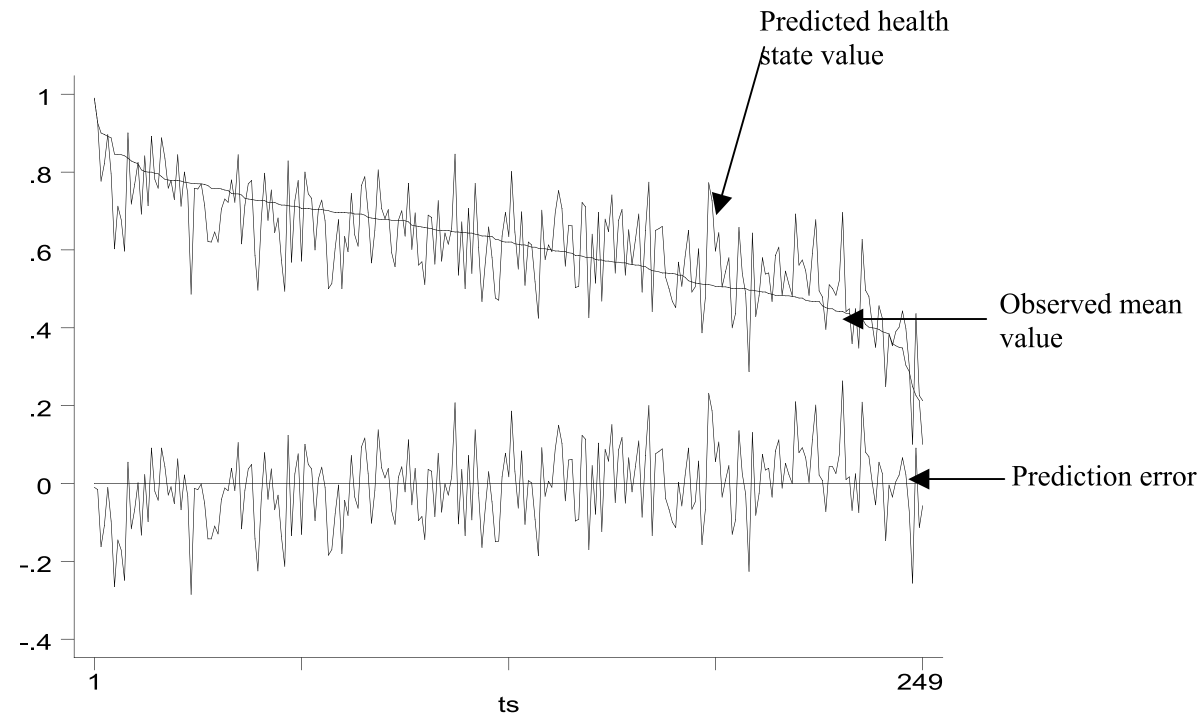


Figure 4: Observed and predicted health state values for the SF-6D Standard Gamble model



References

[1] Torrance G.W. Measurement of health state utilities for economic appraisal: a review Journal of Health Economics 1986;5:1-30
[2] Brazier J.E. Deverill M. Green C. Harper R. Booth A. A review of the use of health status measures in economic evaluation Health Technology Assessment 1999;3(9)
[3] Bleichrodt H. Johanneson M. An experimental test of a theoretical foundation for rating-scale valuations Med Dec Making 1997;17:208-216
[4] Dolan P. Sutton M. Mapping visual analogue scale scores on to time trade off and standard gamble utilities Social Science and Medicine 1997; 44(10):1519-1530
[5] National Institute for Clinical Excellence (NICE) Guide to the methods Technology Appraisal. Draft for consultation London: NICE 2003
[6] Bleichrodt H. Johannesson M. An experimental test of a theoretical foundation for rating scale valuations Medical Decision Making 17:208-216
[7] Torrance G.W. Feeny D. Furlong W. Visual Analogue Scales: Do they have a role in the measurement of preferences for health states? Med. Dec. Making 2001;21:329-334
[8] Kind P. Turning lead into gold – the alchemist returns Paper presented to HESG Birmingham 1999
[9] Kind P. Deriving cardinal scales from ordinal preference data: the analysis of time trade-off data using pairwise judgement models, Paper presented to HESG Brunel University 1996
[10] Furlong W. Feeny D. Torrance G.W. Barr R. Horsman J. Guide to design and development of health state utility instrumentation CHEPA Working Paper #90-9 McMaster University Hamilton, Ontario 1990
[11] McFadden D. Conditional Logit analysis of qualitative choice behaviour. In Zarembka P (ed) Frontiers in Econometrics New York Academic Press;1974 105-142
[12] Salomon J A The use of ordinal ranks in health state valuations IHEA Conference San Francisco June 2004
[13] Dolan P. Modelling valuations for EuroQol health states. Medical Care 1997;35(11):1095-1108
[14] Torrance G.W. Feeny D.H. Furlong W.J. Barr R.D. Zhang Y. Wang Q. A multi-attribute utility function for a comprehensive health status classification system: Health Utilities Mark 2 Medical Care 1996;34(7):702-722
[15] Brazier J.E. Roberts J. Deverill M. The estimation of a preference based measure of health from the SF-36 Journal of Health Economics, 2002; 21(2):271-292
[16] McCabe C. Stevens K. Roberts J. Brazier J.E. Health state values for the HUI2 descriptive system: results from a UK survey Sheffield Health Economics Group Discussion paper 03/3 University of Sheffield 2003 Sheffield
[17] Ljung G. and Box G. On a measure of lack of fit in time series models Biometrika 1979;66:265-270
[18] Sturgis P. Thomas R. Deriving a preference based utility score for the SF-6D: technical report. Survey Methods Centre at SCPR 1998
[19] Dyer J.S. Sarin R.K. Relative risk aversion. Management Science 1982;28(8):857-886
[20] Broome J. QALYs Journal of Public Economics 1994;50(2):149-167
[21] Hausman J.A. Ruud P.A. Specifying and testing econometric models for rank-ordered data Journal of Econometrics 1987; 34:83-104
[22] Saigal S. Feeny D. Rosenbaum P Furlong W Burrows E Stoskopf B Self-perceived health status and health related quality of life of extremely low birth weight infants at adolescence JAMA 1996;276(6):453-459