Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of three methods of analysis

Terry N Flynn[1*], Jordan J Louviere[2], Tim J Peters[3], Joanna Coast[4]

[1]MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol

[2]Centre for the Study of Choice, University of Technology, Sydney

[3]Department of Community Based Medicine, University of Bristol

[4]Health Economics Facility, University of Birmingham

Abstract

Best-worst data from a discrete choice experiment (DCE) can be analysed in many ways. Economists may be tempted to avoid methods that aggregate choice data across respondents (e.g. weighted least squares) in favour of logistic/probit analysis because of concerns about small numbers of observations and implications for utility part-worth estimates. Such concerns are unfounded provided that an orthogonal design is used – issues of multicollinearity should not arise and WLS estimates should be co-linear with those from multinomial/conditional logistic regression. Nevertheless, benefits of model parsimony under WLS should be weighed against the ability of multinomial-based models to estimate the effect of respondent characteristics upon preferences. The study reported here is the first in health care to illustrate this.

A best-worst DCE elicited preferences for aspects of a dermatology consultation. Three analytical methods were compared: two using WLS – the paired and marginal methods, and conditional logistic regression. Regressing one set of results against another always produced a high R-squared whilst standard errors were adversely affected only by the marginal method. Further analysis showed the ability of logistic regression to estimate the effects of respondent-level covariates upon preferences: in particular, the use of effect coding enabled an investigation of whether it was attribute importance and/or level scale values that were affected by sociodemographic factors. This is important because health policies to change the levels of attributes in health care may be very different from those aiming to change the attributes per se.

* Terry N Flynn, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK; Tel: +44 (0)117 928 7375; Fax: +44 (0)117 928 7236; E-mail: terry.flynn@bristol.ac.uk .

This work was supported by the MRC Health Services Research Collaboration.

Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of three methods of analysis

1. Introduction

Stated Preference Discrete Choice Modelling (SPDCM) elicits people's preferences for goods or services based on their intentions expressed in hypothetical situations (Louviere Hensher and Swait, 2000). A traditional discrete choice experiment (DCE) involves choosing the most preferred specification of a good ('alternative' or 'scenario') from a choice set of competing scenarios (Louviere and Timmermans, 1990). When respondents choose their preferred scenario, they are effectively providing information about their preferences *relative* to either a particular scenario or the mean utility in the sample. The utility estimates therefore represent a set of deviations which cannot be used directly to make statements about the *overall* impact of attributes (Flynn et al., 2005).

There are certain issues in Health Services Research (HSR) that do require the analyst to be able to compare the overall impact of attributes. An example would be testing the hypothesis that "waiting time for an appointment is more important to patients than continuity of care" – a statement purely about attributes, with no reference to the associated levels. This issue of separating attribute impact weights and scales – estimating the utility associated with a particular attribute per se (its weight or impact in a utility function) separately from the additional utility gained/taken away by that attribute exhibiting an attractive/unattractive level (the scale value) has been explored and several methods are now available to help address this (Lancsar Louviere and Flynn, 2005). Best-worst scaling (Marley and Louviere, 2004), devised by Finn and Louviere (Finn and Louviere, 1992) and introduced to health care research by McIntosh and Louviere (McIntosh and Louviere, 2002) is one solution. A guide to the use of BWS has been provided (Flynn et al., 2005), but briefly, unlike most traditional DCEs, BWS presents the respondent with each scenario one at a time – in other words the choice set is of size one.

Rather than comparing the utility of entire scenarios, respondents evaluate and compare the utilities of the attributes on offer (or, rather, the particular attribute levels on offer), picking that pair of attributes that maximises the difference in utility between them.

The paper by Flynn *et al* describes various methods of analysis of best-worst choice data but to date there have been no applications in HSR that have compared these methods. Furthermore, there is a need to demonstrate the flexibility of the methods in unbalanced designs (where the number of levels per attribute is not constant) and in estimating the effect of patient-level covariates upon attribute impacts and level scale values. This paper addresses these issues and puts forward ideas for future research in the area. The empirical study will be described in Section 2. Section 3 will summarise the methods of analysis and Section 4 will set out the BWS results. Section 5 will discuss the implications of this work and the final section will conclude.

2. The empirical study

The empirical work was undertaken in the context of a project aiming to quantify preferences for different aspects of access to dermatology secondary care services. The work was conducted alongside a randomised controlled trial, with associated economic evaluation, comparing consultant-led out-patient care with local care provided by a GPSI (General Practitioner with a Special Interest in Dermatology). The development of both attributes and levels was conducted using qualitative work (Coast and Horrocks, 2005). This ensured that the attributes chosen were relevant and grounded in patients' experiences. The four attributes identified were waiting time, degree of expertise of doctor, convenience of attending and degree of individualised care. Waiting time had four levels whilst the other attributes all had two levels.

The process to generate an appropriate design is described elsewhere (Coast et al., 2005) and two versions of the questionnaire were used, a long one utilising 16 scenarios and a short one utilising eight. The analysis reported here relates to the long questionnaire. In each scenario (appointment offered) the respondent was asked to choose which attribute was best and which was worst, based on the levels that the four attributes took.

3. Methods of analysis

Several methods of analysis were performed on the choice data but a common feature was the use of effect coding for the independent variables. The benefits of effect coding over the use of dummy variables have already been illustrated in a traditional DCE (Bech and Gyrd-Hansen, 2005;Lancsar and Savage, 2004). Effect coding is particularly well suited to BWS because the attribute impact is estimated separately from the level scale values (deviations from attribute impact), allowing both comparisons of attribute impact and significance of scale values to be read directly from the results.

As detailed by Flynn *et al*, Best-worst data can be analysed in several ways (Flynn et al., 2005). Choice data can be aggregated (or not) across attribute pairs and/or across respondents, leading to 2x2=4 possible models. Weighted least squares (WLS) is the appropriate method of analysis for the two models that aggregate choices across respondents (models 1 and 2) whilst conditional (multinomial) logistic regression is appropriate when respondent level inference is required (models 3 and 4). The degree of aggregation across choices does not have implications for analysis method.

Models one and three are ‘paired’ analysis models whilst models two and four are ‘marginal’ analysis models (Marley and Louviere, 2004) – paired analysis models treat each unique best-worst pair as an observation whilst marginal analysis models treat each attribute level as an observation (aggregating pairs up to give the marginal frequencies). A full exposition of these methods has been given before

(Flynn et al., 2005) so the next three sections will summarise them briefly with reference to this study in particular.

3.1. Model 1: Paired sample-level analysis

The first method of analysis used in this study was performed at the at the sample level utilising the paired method, which treats each unique best-worst pair as an observation (where order matters). In a design with K attributes where n_k represents the number of levels of attribute k, the number of observations in a main effects design is therefore $2\sum_{i=1}^{K-1}\left[n_i \sum_{k=i+1}^{K} n_k\right]$. A main effects design ensures that across the scenarios in the best-worst exercise every one of these pairings can be estimated. The data were analysed using weighted least squares (weight being the choice totals adjusted to eliminate sampling zeros) with effect codes in order to separate attribute impact weights from level scale values automatically.

3.2. Model 2: Marginal sample-level analysis

The marginal model utilising sample-level choice data was the second to be estimated. It aggregates the choice data to estimate the attribute level utilities using a model that, while simpler for main effects designs, was predicted to suffer from wider confidence intervals around parameters given that the total number of attribute levels to be estimated was relatively small. For main effects designs there are a total of $2\sum_{k=1}^{K} n_k$ observations – each of the attribute levels contributes two observations, a best and a worst total. The data were again analysed by weighted least squares (with weight again being the choice totals adjusted to eliminate zeros).

One aim of the study was to estimate the effect of respondent characteristics (such as age and sex) upon utilities. Therefore covariates were introduced which took the form of respondent-choice interaction

terms and required the use of respondent-level choice data. Conditional (multinomial) regression was therefore used and methods three and four used this as an alternative to the WLS methods described above.

3.3. Models 3 and 4: Conditional (multinomial) logistic analysis

Limited dependent variable models require differences in the probabilities of choice for the various outcomes in a choice set to be associated with differences in the explanatory variables. Since respondent characteristics, such as age, do not vary for potential best-worst pairs in a choice set they cannot affect choice probabilities and cannot be separated out from the overall regression constant term. Thus covariates were interacted with the choice variables (effect codes). Both the paired and marginal regression models above were analysed using conditional logit regression with the *clogit* command in Stata (Stata Corporation, 2005). This required the data to be manipulated to ensure it was in the correct format: each outcome picked had to be expanded out into however many outcomes were available to be picked in that choice set – whether K(K-1) pairs under the paired model (model 3) or K attribute levels under the marginal model (model 4). The dependent variable took a value of one for the outcome picked and zero for the remaining (non-chosen) pairs or attribute levels for that choice set and individual.

3.4. Effects of respondent-level covariates

In analysing the effects of respondent-level covariates upon preferences decisions had to be made regarding the treatment of contradictory signs/significance across the two methods (paired/marginal) and two sample sizes (including respondents who provided some best-worst data versus those who provided complete best-worst data). Given that the original trial was not powered to detect differences between subgroups, it was decided to report only those covariates that were statistically significant at

the 5% level under the paired analysis model for both sample sizes (with any differences under the marginal model highlighted) using a partially adjusted model. Significant covariates were then entered into a fully adjusted model. Given that best-worst scaling distinguishes between two types of preference – the attribute impact and the level scale values – significant effects on either are reported.

4. Results

Ninety-three individuals provided best-worst data that allowed estimation using any of the four methods and 60 individuals provided complete best-worst choice data. The minimum number of appointments answered was five whilst 85 individuals answered 14, 15 or 16 appointments. Stata chooses an attribute impact variable arbitrarily to drop in order to prevent the model being saturated. Therefore, once the least valued attribute was identified, all analyses were performed with this attribute impact omitted to ease interpretation.

4.1. Paired WLS analysis

Table 1 shows the results for the paired method WLS analysis for the 93 respondents who provided at least some choice information in the long questionnaire. Waiting time was the attribute with least impact and its impact weight is therefore omitted – the impact figures for the other three attributes are therefore relative to waiting time (on an interval scale). Doctor expertise is clearly the most highly valued attribute whilst convenience and degree of individualised care are both valued approximately equally. The result of separating overall attribute impact from level scale values is clear: whilst individualised care is not the most important attribute per se, the two levels are very far apart on the utility scale, unlike convenience of attending: there is a difference of 2x1.66=3.32 units between the levels of individualised care but only 2x0.42=0.84 units between the levels of convenience. This illustrates a key advantage of BWS over traditional DCEs: in the latter, only these differences are

estimable. The four levels of waiting time are sensibly ordered but there are clearly increasing returns to scale here.

4.2. Marginal WLS analysis

Table 2 displays the WLS results for the marginal method, again for the 93 individuals who provided some best-worst choice information for the long questionnaire. The estimates are very similar to those of the paired method (although confidence intervals are much wider, reflecting the small number of observations), and an ordinary least squares regression of the nine estimated utilities for the two methods showed a highly linear relationship (See Figure 1: R^2=0.96 with insignificant constant term and slope of 1.04). Repeating this analysis for the 60 individuals who provided complete best-worst choice data gave almost identical results (not shown).

4.3. Comparison of WLS and clogit results

Table 3 shows the conditional logit results using the paired method for the 93 respondents who provided some choice data. Although the pseudo R-squared value of 0.42 is much lower than the adjusted R-squared from the weighted least squares (0.86), the two figures are not strictly speaking comparable. The two sets of parameter estimates are highly linear, with R^2=0.97 (see Figure 2) when regressed using OLS. The results were almost identical when the analysis was repeated using the 60 respondents who provided full best-worst data (R^2=0.97).

Comparison of results for the two estimation methods when using the marginal model gave similar patterns (results not shown). R^2=0.99 was observed for sample sizes of both 93 and 60 when the logistic model estimates were regressed against those for the weighted least squares model.

4.4. Effects of respondent-level covariates

The partially adjusted *clogit* analysis indicated that sex and age did not have strong effects upon mean preferences – there was a statistically significant reduction in the utility scale value associated with a 1 month wait for men but given the lack of effect for the other waiting times (and hence lack of any 'pattern') this was not considered important. However, having attended higher education or being in fulltime employment or having severe self-reported skin problems or being classed as an 'acute' patient did have significant effects for sample size of both 60 and 93. The fully adjusted model therefore included all interactions between attribute impact weights and level scale values and these four covariates. Tables 4 and 5 summarise the paired and marginal results for the 93 respondents who provided some best-worst choice data. Results for the 60 respondents with complete best-worst choice data exhibited very few differences and were qualitatively the same.

There were ten parameters that were significant at the 5% level in both models. The two parameters that were significant only in the marginal model were significant at the 10% in the paired model but the five that were significant only in the paired model did not approach significance at any standard level in the marginal model.

Education

The lower levels of convenience of attending, degree of individualised care and expertise of doctor all provided greater disutility for this group, compared with those without a higher education. There was also an indication of greater range of utility values associated with waiting times. Thus, for those respondents with higher education, the levels of all four attributes lie further apart on the utility scale than those without higher education. All these differences were with respect to the scale level values, not the attribute impact weights, which did not appear to vary by educational status. When analysing using the marginal method there were very similar results. The only difference was that *educ-dr* was significantly different from zero whilst *educ-drpt* was not. A possible explanation for these findings

concerns the correlation often observed between educational attainment and social class and/or income. Thus, it might be expected that greater access to alternative sources of health care (e.g. private health) might mean that those respondents with a higher education experience greater differences in utility associated with levels of the attributes.

Employment

For similar reasons to those given for education, it might be expected that employment would be associated with greater disutility attached to lower levels of attributes. This was observed for convenience and degree of individualised care, across both methods of analysis. However there was a small but significant increase in the impact of individualised care for this group to offset this, under the paired method only. It was interesting that both methods showed an increase in the attribute impact of doctor expertise for those in employment compared with those not.

Recent or severe skin problems

Those with at least one of the ten factors severely disrupted by their skin condition or those with a total score of seven or higher (out of 30) on the scale incorporating these factors might be expected to exhibit smaller differences between levels of attributes and possibly attenuation of differences in attribute impact weights as 'simply getting into secondary care' becomes paramount. These two problems are highly correlated so the second one (total score 7 or above) was used in analysis. Indeed a lack of individualised care was not associated with as much disutility for these groups as with the rest of the sample. Attending a consultation that was hard to get to or seeing a part-time doctor was also not as disliked for this group though these findings, unlike that for individualised care, were not apparent in the marginal method analysis. There was also a suggestion that the impact of the three estimated attributes was less for this group (although only significant in both models for doctor expertise). This would be consistent with the hypothesis that such respondents attach more impact to waiting time than other people.

Being an 'acute' case

Being an assumed 'acute' case (proxied by having first seen their GP within the previous 6 months and only seeing him/her once or twice before referral) does not appear to alter preferences to any great extent: there was greater disutility associated with a lack of individualised care but the only other statistically significant parameter was *acute_1m* and there was no pattern to the waiting time parameters. Indeed, this classification appears to be poor at identifying those respondents who themselves believe their skin problem has been particularly bad – tabulating acute against either of the patient-expressed disease severity factors indicated little agreement (see Table 6).

5. Discussion

This study is the first within HSR to illustrate and compare the results from various methods of analysing best-worst scaling choice data. It demonstrates the flexibility and accuracy of aggregated analysis with as few as twenty observations when factors are manipulated according to a good design matrix. In other words, the problems of multicollinearity and lack of variability in key factors which necessitate large datasets in many econometric studies do not apply. Agreement across paired and marginal methods and across WLS and logistic regressions was extensive and the strengths of effect coding in a best-worst context were illustrated.

The importance of separating attribute impact from level scale values was also apparent when performing patient-level analyses: the study demonstrated that certain factors, most notably higher education but to a lesser extent being in fulltime employment, caused the range of scale values to widen, whilst attribute importance was not changed much (if at all). There was a suggestion that having self-reported recent or severe skin problems increased the importance of waiting time relative to other attributes and narrowed the range of scale values for non-waiting time attributes. Both these sets of findings have intuitive explanations but given the relatively small sample they should be investigated further.

There appeared to be little to distinguish the results from the paired analysis from those of the marginal analysis except larger standard errors under the marginal WLS model, which in some cases made parameters marginally insignificant. The WLS estimates were also highly linear with those from logistic regression and can be reported for any analysis that does not require respondent-level covariates.

5.1. Limitations

An assumption underlying the regression models presented here was that of a constant variance of the random component of utility both within and between individuals. In health economics, the treatment of the variation in utility between respondents within discrete choice experiments has largely been restricted to the use of random effects to model respondent heterogeneity in (usually probit) regression models. However, not only does this particular focus on preference heterogeneity ignore the other factors that might lead to variation in choice behaviour, it is conceptually equivalent to allowing for variation in the fixed component of individuals' utilities (the mean) but not in the variance of the random component. As such it is a partial solution at best and there is evidence to suggest that this simplistic treatment of heterogeneity is not supported empirically (Louviere, 2001). Furthermore, failure to allow for variation in the mean by the inclusion of random effects only leads to incorrect standard errors; failure to recognise variation in the random component of utility leads to incorrect point estimates.

SPDCM studies in HSR have suffered from designs that were more appropriate to old conjoint-analysis studies which desired individual-level (or small group-level) utilities rather than population parameters. In other words, by utilising a common, small, design for all respondents, it is impossible to span the entire response surface (estimate all interactions and thus provide a complete utility function) and

thereby provide population parameters that are protected against unobserved interactions. To an extent, this criticism can be levelled against the present study: splitting the sample into blocks and administering different versions of the questionnaire which together spanned the full factorial (32) appointments would have addressed this issue. However, the study was never intended to provide population parameters and practical constraints precluded such a design. Furthermore, future work will exploit advantages of the current design to investigate individual-level preferences.

A final issue concerns anchoring of the utility estimates. Best-worst choice data from such a task as this provide estimates that are interval-scaled with unknown anchor – when total utilities for each appointment are constructed and ranked, the analyst cannot know at which point utility becomes positive (indicating that the respondents will choose to attend the appointment rather than not attend). This may or may not be a limitation, depending upon the use to which the estimates are put. Planning total service provision to match demand would require unconditional demand information, not the conditional demand information that the results above provide. However, marginal changes in service provision can be addressed using these results by way of calculating marginal rates of substitution as in a traditional DCE. Constructing an outcome or service index based on these results is also possible, but more generally it might be the case that the need for an anchor (for example in a quality of life scale to be used in QALY estimation) necessitates additional information from respondents.

5.2. Future work

One of the aims of this study was to investigate differences in response rates and results between two versions of the questionnaire – one with 8 appointments and one with 16. Differences were found to be minimal (Coast et al., 2005) and future work should compare longer questionnaires, perhaps 16 versus 32 scenarios. If a similarly small design to that here is used, then such a design may permit investigation of interactions and/or any other violations of the IIA assumption.

In addressing the limitation of assuming a constant variance it would seem more logical to exploit the power of best-worst to make individual-level inference than to attempt to introduce random effects into the models detailed above. Indeed, work has begun to utilise the power of best-worst scaling to model individual-level utility functions that require no statistically questionable distributional assumptions surrounding preferences (Louviere et al., 2004). Similar work will be performed for these data.

This study also asked respondents whether they would attend each appointment offered to them. The results from this provide an alternative set of utility estimates (although relative to one appointment or the mean utility). Differences between the two methods will be investigated as will the extent to which the anchor provided by these data can be used to rescale the best-worst data. Future studies will consider utilising qualitative work and simulation studies to ascertain whether the cognitive processes undertaken by respondents provide support for such a data synthesis.

6. Conclusion

This study has shown that aggregated methods provide simple compact datasets yet give results no different from those of individual-level analysis. This study has also illustrated a key advantage of best-worst scaling over traditional DCEs – the ability to separate attribute impacts from level scale values. In so doing it provides additional insights over those from traditional DCEs that should prove attractive in health care research. In particular, this ability to ascertain whether patient subgroups exhibit differences in attribute importance and/or differences in level scale values may have implications for policy.

References

Bech, M. and Gyrd-Hansen, D., 2005. Effects coding in discrete choice experiments. Health Economics. 14, 1079-1083.

Coast, J., Flynn, T. N., Salisbury, C. et al., 2005. Maximising responses to discrete choice experiments: a randomised trial. Health Economics (submitted).

Coast, J. and Horrocks, S., 2005. Developing attributes and levels for discrete choice experiments: a case study using qualitative methods. Journal of Health Services Research and Policy (submitted).

Finn, A. and Louviere, J. J., 1992. Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. Journal of Public Policy & Marketing. 11, 12-25.

Flynn, T. N., Louviere, J. J., Peters, T. J., Coast, J., 2005. Best-Worst Scaling: What it can do for health care and how to do it. Journal of Health Economics (submitted).

Lancsar, E., Louviere, J. J., Flynn, T. N., 2005. Comparing relative attribute impact: Several methods to address the confound between attribute impact and attribute utility scale in stated preference data. Journal of Health Economics (submitted).

Lancsar, E. and Savage, E., 2004. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. Health Economics. 13, 901-907.

Louviere, J. J., 2001. What if consumer experiments impact variances as well as means: Response variability as a behavioural phenomenon. Journal of Consumer Research. 28, 506-511.

Louviere, J. J., Burgess, L., Street, D., Marley, A. A. J., 2004. Modeling the choices of single individuals by combining efficient choice experiment designs with extra preference information. CenSoC working paper series 04-004. Centre for the Study of Choice, Faculty of Business, University of Technology, Sydney.

Louviere, J. J., Hensher, D. A., Swait, J., 2000. Stated choice methods: analysis and application. Cambridge University Press, Cambridge.

Louviere, J. J. and Timmermans, H., 1990. Stated Preference and Choice Models Applied to Recreation Research: A Review. Leisure Sciences. 12, 9-32.

Marley, A. A. J. and Louviere, J. J., 2004. Some probabilistic models of Best, Worst, and Best-Worst choices. CenSoC working paper series 04-005. Centre for the Study of Choice, Faculty of Business, University of Technology, Sydney.

McIntosh, E. and Louviere, J. J. Separating weight and scale value: an exploration of best-attribute scaling in health economics. Health Economics Study Group. January 2002.

Stata Corporation., 2005. Stata Statistical Software. College Station, TX

Table 1: Best-worst utilities (paired WLS method) for sample size of 93

```
      Source |       SS       df       MS              Number of obs =      72
-------------+------------------------------           F(  9,    62) =   49.86
       Model |  62.0060435     9  6.88956038           Prob > F      =  0.0000
    Residual |  8.56642838    62    .1381682           R-squared     =  0.8786
-------------+------------------------------           Adj R-squared =  0.8610
       Total |  70.5724718    71  .993978477           Root MSE      =  .37171

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   1.874463   .0999756    18.75   0.000     1.674614    2.074311
Attribute Impact
Waiting time |          -          -        -       -            -           -
          Dr |   .8436762   .0854399     9.87   0.000     .6728842    1.014468
 Convenience |   .3640331   .0982786     3.70   0.000      .167577    .5604893
   Indivcare |   .3456281   .0645439     5.35   0.000     .2166068    .4746494
Level scale values
      wait3m |  -.7697017   .0894516    -8.60   0.000    -.9485129   -.5908905
      wait2m |  -.4897349   .0971762    -5.04   0.000    -.6839875   -.2954824
      wait1m |   .0597706   .1043847     0.57   0.569    -.1488914    .2684326
      wait0m |   1.199666          -        -       -            -           -
    drpttime |  -.6787816   .0777337    -8.73   0.000     -.834169   -.5233941
  drfulltime |   .6787816          -        -       -            -           -
    convhard |  -.4198611   .1006805    -4.17   0.000    -.6211187   -.2186036
    conveasy |   .4198611          -        -       -            -           -
     indivno |  -1.663849   .0934779   -17.80   0.000    -1.850709    -1.47699
    indivyes |   1.663849          -        -       -            -           -
------------------------------------------------------------------------------
```

Linearity test: i.e. wait3m=3wait2m and wait2m=-wait1m

```
test wait3m=3*wait2m

( 1)  wait3m - 3 wait2m = 0

       F(  1,    62) =    4.81

            Prob > F =    0.0321

test wait2m=-wait1m,accum

 ( 1)  wait3m - 3 wait2m = 0

 ( 2)  wait2m + wait1m = 0

       F(  2,    62) =    7.52

            Prob > F =    0.0012
```

Table 2: Best-worst utilities (Marginal WLS method) for sample size of 93

Source	SS	df	MS	Number of obs =	20
Model	9.40085004	10	.940085004	F(10, 9) =	9.83
Residual	.860535146	9	.095615016	Prob > F =	0.0010
				R-squared =	0.9161
Total	10.2613852	19	.540072904	Adj R-squared =	0.8230
				Root MSE =	.30922

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	4.754318	.116581	40.78	0.000	4.490594	5.018043
bwindic	-.5004844	.1174622	-4.26	0.002	-.7662025	-.2347664
Attribute Impact						
Waiting time	-	-	-	-	-	-
Dr	1.113387	.2205054	5.05	0.001	.6145686	1.612205
Convenience	.5256616	.2339122	2.25	0.051	-.0034846	1.054808
Indivcare	.412651	.1774225	2.33	0.045	.0112934	.8140086
Level scale values						
wait3m	-.8255612	.178787	-4.62	0.001	-1.230005	-.4211169
wait2m	-.5495118	.1962232	-2.80	0.021	-.9933996	-.105624
wait1m	.058626	.2109858	0.28	0.787	-.418657	.5359089
wait0m	1.316447	-	-	-	-	-
drpttime	-.7332097	.17905	-4.10	0.003	-1.138249	-.3281705
drfulltime	.7332097	-	-	-	-	-
convhard	-.5848886	.2236915	-2.61	0.028	-1.090914	-.0788633
conveasy	.5848886	-	-	-	-	-
indivno	-1.423439	.1752517	-8.12	0.000	-1.819886	-1.026992
indivyes	1.423439	-	-	-	-	-

Table 3: Best-worst utilities (Paired clogit method) for sample size of 93

Conditional (fixed-effects) logistic regression — Number of obs = 16908; LR chi2(9) = 2962.84; Prob > chi2 = 0.0000; Log likelihood = -2019.8115; Pseudo R2 = 0.4231

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Attribute Impact						
Waiting time	-	-	-	-	-	-
Dr	1.470231	.0746414	19.70	0.000	1.323936	1.616525
Convenience	.6123097	.0687926	8.90	0.000	.4774786	.7471407
Indivcare	.4184067	.069822	5.99	0.000	.281558	.5552553
Level scale values						
wait3m	-1.547543	.1046755	-14.78	0.000	-1.752703	-1.342382
wait2m	-.8123266	.0977497	-8.31	0.000	-1.003912	-.6207408
wait1m	.2888487	.0949235	3.04	0.002	.102802	.4748953
wait0m	2.071021	-	-	-	-	-
drpttime	-1.282006	.0627516	-20.43	0.000	-1.404997	-1.159016
drfulltime	1.282006	-	-	-	-	-
convhard	-.964847	.0617312	-15.63	0.000	-1.085838	-.843856
conveasy	.964847	-	-	-	-	-
indivno	-2.393477	.0724296	-33.05	0.000	-2.535437	-2.251518
indivyes	2.393477	-	-	-	-	-

Table 4: Best-worst utilities (Paired clogit method adjusting for respondent covariates) for sample size of 93

```
Conditional (fixed-effects) logistic regression   Number of obs   =      15696
                                                  LR chi2(45)     =    2938.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -1780.9401                       Pseudo R2       =     0.4521

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Attribute Impact
Waiting time |          -          -       -       -           -           -
          Dr |   1.342555   .1117852    12.01   0.000      1.12346     1.56165
 Convenience |   .5544422   .1045399     5.30   0.000     .3495477    .7593367
   Indivcare |   .3628801   .1053237     3.45   0.001     .1564495    .5693108
     educ_dr |  -.1317751   .0923188    -1.43   0.153    -.3127166    .0491665
    educ_conv |   .0564573   .0860605     0.66   0.512    -.1122183    .2251328
   educ_indiv |   .0145355   .0862812     0.17   0.866    -.1545725    .1836435
       emp_dr |   .3871814   .0825426     4.69   0.000*     .225401    .5489619
     emp_conv |   .0975757   .0751429     1.30   0.194    -.0497016    .2448531
    emp_indiv |   .1830449   .0763205     2.40   0.016*    .0334595    .3326304
    score7_dr |  -.3202987    .088646    -3.61   0.000*   -.4940416   -.1465559
  score7_conv |  -.1181401   .0826505    -1.43   0.153    -.2801322    .0438519
 score7_indiv |  -.1738758   .0823243    -2.11   0.035*   -.3352284   -.0125232
     acute_dr |    .047303   .0958845     0.49   0.622    -.1406271    .2352332
   acute_conv |  -.0154903   .0864143    -0.18   0.858    -.1848592    .1538785
  acute_indiv |  -.0413346   .0901086    -0.46   0.646    -.2179441     .135275
Level scale values
       wait3m |  -1.958953   .1605818   -12.20   0.000    -2.273687   -1.644218
       wait2m |  -1.117335   .1493553    -7.48   0.000    -1.410066   -.8246039
       wait1m |   .2137621   .1457884     1.47   0.143    -.0719779     .499502
       wait0m |   2.862526          -       -       -           -           -
     drpttime |  -1.470253   .1035633   -14.20   0.000    -1.673234   -1.267273
   drfulltime |   1.470253          -       -       -           -           -
     convhard |  -1.185982    .102335   -11.59   0.000    -1.386555   -.9854091
     conveasy |   1.185982          -       -       -           -           -
      indivno |  -2.843362   .1205684   -23.58   0.000    -3.079671   -2.607052
     indivyes |   2.843362          -       -       -           -           -
      educ_3m |  -.4883798   .1332613    -3.66   0.000*   -.7495672   -.2271924
      educ_2m |  -.2318958   .1232679    -1.88   0.060    -.4734964    .0097049
      educ_1m |   .2727426   .1212335     2.25   0.024*    .0351292    .5103559
    educ_drpt |  -.1920152    .085256    -2.25   0.024*   -.3591139   -.0249165
 educ_convh~d |  -.3173861   .0854444    -3.71   0.000*   -.4848541   -.1499182
 educ_indivno |  -.4161934   .0982534    -4.24   0.000*   -.6087665   -.2236203
       emp_3m |    .049469   .1134641     0.44   0.663    -.1729165    .2718545
       emp_2m |  -.0683019   .1064429    -0.64   0.521    -.2769262    .1403223
       emp_1m |   .0299269   .1039932     0.29   0.774     -.173896    .2337498
     emp_drpt |   .1168757   .0688412     1.70   0.090    -.0180505    .2518019
 emp_convhard |  -.1581264   .0678109    -2.33   0.020*   -.2910334   -.0252194
  emp_indivno |  -.1832685   .0784179    -2.34   0.019*   -.3369647   -.0295723
    score7_3m |  -.1215269   .1246303    -0.98   0.330    -.3657979    .122744
    score7_2m |  -.2264255    .116925    -1.94   0.053    -.4555942    .0027433
    score7_1m |    .022866   .1132425     0.20   0.840    -.1990853    .2448173
  score7_drpt |   .1038593   .0744481     1.40   0.163    -.0420564    .2497749
 score7_con~d |   .1480246   .0727303     2.04   0.042*    .0054759    .2905734
 score7_ind~o |   .2354545   .0826623     2.85   0.004*    .0734394    .3974696
     acute_3m |  -.0789485   .1297866    -0.61   0.543    -.3333256    .1754286
     acute_2m |  -.1220491   .1227597    -0.99   0.320    -.3626537    .1185554
     acute_1m |  -.2721822   .1206255    -2.26   0.024*   -.5086038   -.0357607
   acute_drpt |  -.0977712   .0833049    -1.17   0.241    -.2610458    .0655034
 acute_conv~d |  -.0981763   .0807079    -1.22   0.224     -.256361    .0600084
 acute_indi~o |  -.4416662   .0975084    -4.53   0.000*   -.6327792   -.2505532
------------------------------------------------------------------------------
```

Table 5: Best-worst utilities (Marginal clogit method adjusting for respondent covariates) for sample size of 93

```
Conditional (fixed-effects) logistic regression   Number of obs   =      10464
                                                  LR chi2(45)     =    2975.92
                                                  Prob > chi2     =     0.0000
Log likelihood = -2870.5638                       Pseudo R2       =     0.3414

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Attribute Impact
Waiting time |          -          -        -        -           -           -
          Dr |   .7800891   .0717282    10.88   0.000     .6395045    .9206737
 Convenience |  -.0145932   .0705337    -0.21   0.836    -.1528367    .1236502
   Indivcare |  -.2790265    .082065    -3.40   0.001     -.439871   -.1181821
     educ_dr |   -.182372   .0591949    -3.08   0.002    -.2983919   -.0663522
   educ_conv |   .0550772   .0576892     0.95   0.340    -.0579915    .1681459
  educ_indiv |   .0557139    .066113     0.84   0.399    -.0738652     .185293
      emp_dr |   .2024173   .0544301     3.72   0.000*    .0957363    .3090984
    emp_conv |  -.0747075   .0526635    -1.42   0.156    -.1779261     .028511
   emp_indiv |   .0343838   .0590324     0.58   0.560    -.0813176    .1500853
   score7_dr |  -.1433994   .0606033    -2.37   0.018*   -.2621797   -.0246191
 score7_conv |   .0363472   .0589272     0.62   0.537    -.0791481    .1518424
score7_indiv |  -.0137167   .0637595    -0.22   0.830     -.138683    .1112497
    acute_dr |  -.0016833   .0609309    -0.03   0.978    -.1211056     .117739
  acute_conv |  -.0108402   .0590264    -0.18   0.854    -.1265299    .1048495
 acute_indiv |  -.0731592   .0703533    -1.04   0.298    -.2110492    .0647308
Level scale values
      wait3m |  -1.463668   .1249856   -11.71   0.000    -1.708635   -1.218701
      wait2m |  -.8649398   .1207746    -7.16   0.000    -1.101654   -.6282259
      wait1m |   .0750807   .1203923     0.62   0.533    -.1608838    .3110453
      wait0m |  -2.253527          -        -        -           -           -
    drpttime |  -1.057877   .0732785   -14.44   0.000    -1.201501   -.9142542
  drfulltime |   1.057877          -        -        -           -           -
    convhard |  -.8444142   .0739089   -11.43   0.000    -.9892729   -.6995555
    conveasy |   .8444142          -        -        -           -           -
     indivno |  -2.286102   .0864978   -26.43   0.000    -2.455634   -2.116569
    indivyes |   2.286102          -        -        -           -           -
     educ_3m |  -.2334104   .1028604    -2.27   0.023*   -.4350131   -.0318077
     educ_2m |  -.1007459   .0993255    -1.01   0.310    -.2954203    .0939285
     educ_1m |   .1814796   .0992479     1.83   0.067    -.0130428     .376002
   educ_drpt |  -.0293256   .0605334    -0.48   0.628    -.1479689    .0893177
educ_convh~d |  -.1289342   .0606831    -2.12   0.034*   -.2478709   -.0099975
educ_indivno |   -.198516   .0697814    -2.84   0.004*   -.3352849    -.061747
      emp_3m |   .0558365    .093034     0.60   0.548    -.1265068    .2381797
      emp_2m |  -.0521471   .0901834    -0.58   0.563    -.2289034    .1246091
      emp_1m |   .0007894   .0896923     0.01   0.993    -.1750043    .1765831
    emp_drpt |   .1206734   .0547522     2.20   0.028*     .013361    .2279858
emp_convhard |  -.1141676   .0544945    -2.10   0.036*   -.2209748   -.0073605
 emp_indivno |  -.1278788   .0616273    -2.08   0.038*   -.2486661   -.0070916
   score7_3m |  -.1587383   .1045527    -1.52   0.129    -.3636579    .0461813
   score7_2m |  -.2313603   .1009409    -2.29   0.022*   -.4292008   -.0335197
   score7_1m |    .006774   .0999677     0.07   0.946    -.1891592    .2027071
 score7_drpt |   .0355027     .06114     0.58   0.561    -.0843296    .1553349
score7_con~d |    .079125   .0606697     1.30   0.192    -.0397854    .1980354
score7_ind~o |   .1614977   .0664029     2.43   0.015*    .0313505     .291645
    acute_3m |  -.0080495   .1028504    -0.08   0.938    -.2096325    .1935335
    acute_2m |  -.0661588   .1009716    -0.66   0.512    -.2640596     .131742
    acute_1m |  -.1995659   .1006284    -1.98   0.047*    -.396794   -.0023378
  acute_drpt |   .0085886   .0612935     0.14   0.889    -.1115444    .1287217
acute_conv~d |  -.0325824   .0614052    -0.53   0.596    -.1529344    .0877696
acute_indi~o |  -.3153775   .0735848    -4.29   0.000*   -.4596011   -.1711538
------------------------------------------------------------------------------
```

Table 6: Cross-tabulation of 'acute' respondents versus those with skin score>6

acute	score7 -1	1	Total
-1	39	21	60
1	24	6	30
Total	63	27	90

Figure 1: Graph of Marginal method estimates plotted against Paired method estimates; Sample size = 93

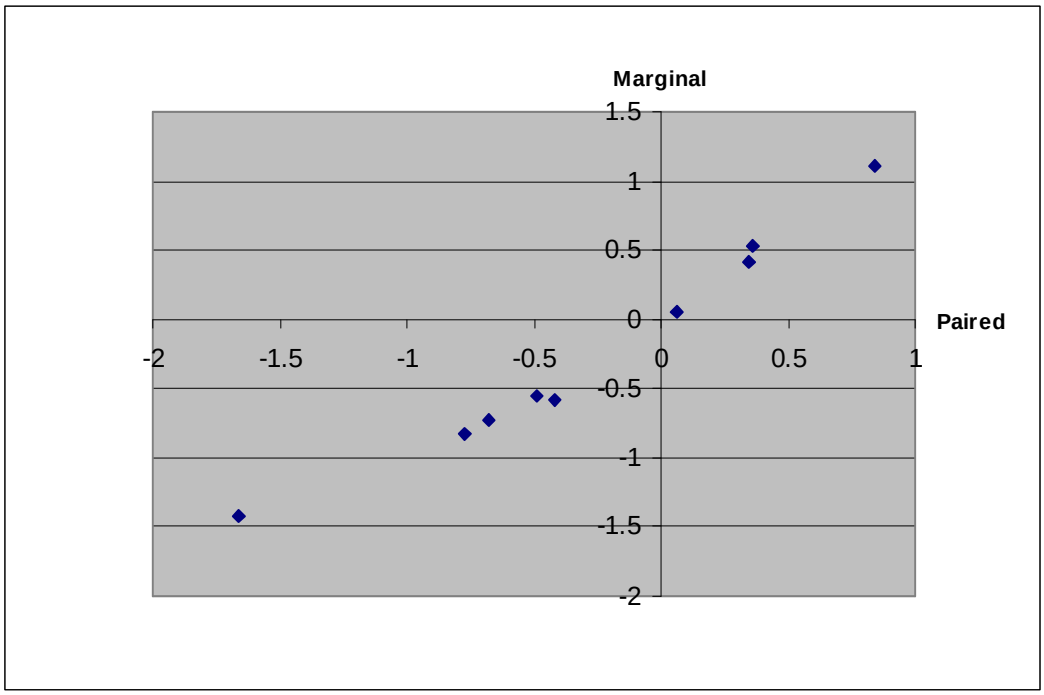

Figure 2: Graph of Paired clogit method estimates plotted against Paired WLS method estimates; Sample size = 93