Interpretation issues in heteroscedastic conditional logit models

Michael Burton[a,b,*], Katrina J. Davis[a,c], and Marit E. Kragt[a]

[a]School of Agricultural and Resource Economics, The University of Western Australia, Crawley, WA 6009, Australia

[b]School of Social Sciences, University of Manchester

[c]Centre of Excellence for Environmental Decisions, University of Queensland

*E-mail address: michael.burton@uwa.edu.au

15 May 2016

Working Paper 1603

School of Agricultural and Resource Economics

http://www.are.uwa.edu.au

Citation: Burton, M., Davis, K.J., Kragt, M.E. (2016) *Interpretation issues in heteroscedastic conditional logit models*, Working Paper 1603, School of Agricultural and Resource Economics, University of Western Australia, Crawley, Australia.

Interpretation issues in heteroscedastic conditional logit models

Michael Burton, Katrina J. Davis and Marit E. Kragt

Abstract: This paper identifies an issue with interpretation of significance within heteroscedastic conditional logit models, due to sensitivity of reported results to arbitrary variable-normalization decisions. We advocate the use of willingness-to-pay (WTP) space models to avoid this, as estimates of WTP do not exhibit this effect. However, in cases where error variance may be high, we question whether it is correct to infer that all respondents should be considered to hold a common WTP.

Key words: Scale, Willingness to Pay, Error Variance, Discrete Choice Models, Aggregate Welfare, Part-worths

JEL classifications: C10, C18, C51, Q51

Acknowledgements

The authors would like to acknowledge useful conversations with Arne Hole, and the NERP research team for providing access to the data used here, but they retain responsibility for errors of interpretation and implementation.

Interpretation issues in heteroscedastic conditional logit models

Michael Burton, Katrina J. Davis and Marit E. Kragt

Introduction

It is widely known that error variance/scale issues confound parameter comparisons across discrete choice models (Davis et al. 2016; Louviere and Eagle 2006). It appears to be less well known that these issues can cause problems for interpretation *within* a single model when heterogeneity in error variance is explicitly modelled. Popular approaches to explicitly model scale heterogeneity include generalized multinomial logit models (Fiebig et al. 2010) and heteroscedastic conditional logit (HCL) models (Hensher et al. 1999; Hole 2006a). We base our analysis of scale-derived interpretation problems on HCL models, where interpreting the significance of effects in preference space representation can be misleading because it is confounded by arbitrary choices when normalising variables. We emphasise that this is an issue of interpretation, not model performance. We illustrate this problem of interpretation using an example data set, and conclude that HCL models would be better estimated in WTP space, where issues of interpretation are not present. Our analysis confirms the insight of Davis et al. (2016) that models that accommodate error heterogeneity, such as HCL models, may mask random behaviour, and we conclude with some thoughts on what the implications are when determining aggregate WTP for policy purposes when using these models.

Methods

The conditional logit model of discrete choice can be represented by the probability that person n choses option i from a set J:

$$P_{ni} = \frac{\exp(\lambda_n X_{ni} \beta)}{\sum_{j=1}^{J} \exp(\lambda_n X_{nj} \beta)}$$

Where X_{ni} is a vector of characteristics relating to alternative i and λ_n is an individual-specific scale factor that is inversely proportional to the error variance in the utility function. If homogeneity in the error process is assumed, then the scale factor is constant across all individuals, and normalised to unity. However, if the error variance is different across individuals the error variance can be parameterised as $\exp(Z_n\alpha)$ where Z is a vector of individual specific characteristics (a heteroscedastic error specification). This parametrisation is useful in that it guarantees that the scale is positive for all cases. This model is now widely available and implemented in e.g. Stata[1] and Nlogit. Although specifying a heteroscedastic error term may be important for avoiding misspecification of the model, it introduces the possibility of misinterpretation of model results, an issue that we believe has gone unrecognised in previous literature (but see for example, Davis et al. 2016). In particular, the results for the preference parameters are conditioned by the choice of coding of the individual specific characteristics (the *Z*'s) used in the scale equation.

To illustrate the issue we use a data set derived from a major study of public values for marine ecosystems. Full details of the study, which concerns the South East Commonwealth Marine Reserved Network (SECMRN) in Australia, are reported in Burton et al. (2015). We limit ourselves here to a number of observations about the study: 1) the design featured 5 ecological features of marine reserves (reported here as *x1–x5*), a personal cost attribute (*cost*), and an ASC dummy for the current situation (*SQ*); 2) we use responses from only one of the 3 information treatments employed in the original study (their Treatment 2); 3) after removing protest responses a sample of 274 is available to us; and 4) the ecological system under consideration is remote and generally not well known, leading to a prior expectation that some respondents would find it difficult to make choices. Here we estimate the canonical model only, using attributes but no sociodemographic data to explain preference heterogeneity. We have two variables that explain heterogeneity: knowledge about the marine ecosystems covered by SECMRN (*Know,* coded on a Likert scale 0-3, from knowing nothing to knowing a lot respectively), and certainty with which they gave answers to the choice questions (*Cert*, coded on a Likert scale 0-3, from very certain to not certain at all respectively. The use of the certainty score to explain error variance may seem uninformative but: a) if significant it provides some evidence of internal consistency; b) improving identification of which responses in the sample have been made under higher uncertainty may improve precision of estimation of preference parameters; c) the approach is similar to the use of stated attribute non-attendance to explain marginal values of attributes. Importantly, our interest is in the consequences of recoding these two variables, not the behavioural interpretation of the specific model.

[1] All estimation in this paper used Stata 13.1, and the commands CLOGITHET (Hole, 2006b) and GMNL (Gu et al, 2013).

Results

Problems with interpretation

Results from the first estimation of a HCL model are reported as Model 1 in Table 1, note that all attributes other than *x4* are significant, as are the part-worths (PW). Heterogeneity in scale is explained by both knowledge and certainty: increasing knowledge is associated with a reduction in scale/increase in error variance, while a higher level of reported uncertainty is also associated with a reduction in scale/increase in error variance. Model 2 is statistically and behaviourally identical, but *Know* and *Cert* are now reverse coded (values 0-3 are now mapped 3-0). As expected, the log-likelihood is identical, and the signs of the *Know* and *Cert* variables are reversed. However, inspection of the preference parameters suggests that none are significant (no p-value is smaller than 0.1). Confronted with Model 2 the researcher would be tempted to suggest that the model has failed: attribute levels are not influencing choices. However, that conclusion would be mistaken. The preference parameters reported from HCL models are those associated with a scale parameter of unity. The preference parameters of those who have different scale parameters (i.e. different error variances) have to be retrieved by re-scaling reported parameters. The divergent results reported in Table 1 arise because the preference parameters reported in Model 1 are associated with those with the smallest error variance i.e. those who are most consistent in their choices, while those in Model 2 are associated with those who have the largest error variance (due to the coding associated with *Cert* and *Know*, the implied scale parameter has a maximum value of 1 in Model 1, and a minimum value of 1 in Model 2). This is our central result: *interpretation* of the significance of the preference parameters is confounded by the normalization used for variables explaining scale heterogeneity.

How might the issue of interpretation illustrated in Table 1 be overcome? Firstly, one might conduct a log likelihood test of the model as a whole relative to the null of no significance: in this case $p<0.0000$, suggesting (unsurprisingly, given Model 1) that there is explanatory power in the model, and this is unchanged by normalization.

Table 1. Heteroscedastic conditional logit models: alternative normalizations of scale.

	Model 1					Model 2			
	Coeff.	p>|z|	PW	p>|z|		Coeff.	p>|z|	PW	p>|z|
x1	0.149	0001	13.0	0.001		0.010	0.138	13.0	0.001
x2	0.107	0.001	9.3	0.001		0.007	0.141	9.3	0.001
x3	0.210	0.000	18.3	0.001		0.014	0.138	18.3	0.001
x4	0.195	0.321	17.0	0.363		0.013	0.397	17.0	0.363
x5	0.109	0.012	9.5	0.028		0.007	0.154	9.5	0.028
SQ	-0.982	0.115				-0.066	0.258		
Cost	-0.011	<0.000				-7.8e-4	0.118		
Scale heterogeneity									
Know	-0.655	0.001			KnowRC	0.655	0.001		
Cert	-0.243	0.018			CertRC	0.243	0.018		
LL	-1043.06					-1043.06			

An alternative analysis would be to estimate the model in WTP space, as reported in Table 2. Here there is stability in preference estimates irrespective of coding within the scale equation, as the reported parameters are part-worths, and these are unaffected by scale. What changes within the normalization is only the constant of the scale equation.

Table 2. Heteroscedastic conditional logit models estimated in willingness-to-pay space.

	Model 1			Model 2	
	Coeff.	p>|z|		Coeff.	p>|z|
x1	13.0	0.001		13.0	0.001
x2	9.3	0.001		9.3	0.001
x3	18.3	0.001		18.3	0.001
x4	17.0	0.363		17.0	0.363
x5	9.5	0.028		9.5	0.028
SQ	-85.6	0.048		-85.6	0.048
Cost	-1	fixed		-1	fixed
Scale heterogeneity					
Constant	-4.468	<0.000	Constant	-7.161	<0.000
Know	-0.655	0.001	KnowRC	0.655	0.001
Cert	-0.243	0.018	CertRC	0.243	0.018
LL	-1043.06			-1043.06	

WTP and error heteroscedasticity

Results from Table 2 raise a different question: what is the appropriate measure of aggregate WTP derived from a HCL model? The logic of the model is that there is a single preference function held by all, and differing levels of error variance. The estimate of the part-worth (as reported in either Table 1 or 2) would apply to all members of the sample, and if representative, to the whole relevant population. However, Model 2 in Table 1 suggests that for some portion of the sample the error variance is sufficiently large that behaviour is indistinguishable from random choices. It seems counterintuitive to claim that this group should be included in estimating aggregate WTP[2]. The issue arises because one has two competing models that are statistically equivalent: I) there is a single universal preference function with error heterogeneity, and an error variance that is so large for a subset that their behaviour is indistinguishable from random *versus* II) there is a subset of the population who truly behave randomly, and whose behaviour can be rationalized by any utility function with a sufficiently large error variance. We have no answer to reconciling these two competing world views, but illustrate the problem with our data set. One can retrieve, at the individual level, estimates of the scale parameter implied by Model 1 in Table 1. They range from 1 to 0.13 with a median of 0.62. 15.3% (42 respondents) have a scale coefficient less than 0.32. We estimate a conditional logit model for this subsample (HCL models would not converge) and confirm that their behaviour is not distinguishable from random choices (the p-value associated with a joint test of all parameters being equal to zero is 0.329)[3]. It might be argued that this lack of significance is due to the small sample size. To test for this, 1000 random samples of 42 individuals were drawn from the set of people with scale greater than 0.32, and conditional logit models estimated. 998 out of the 1000 would reject the null hypothesis of random behaviour at a 5% threshold. This suggests that we have a robust result in suggesting that 15% of our sample have behaviour that is indistinguishable from random choices. If this model were to be used for policy analysis to generate aggregate WTP for environmental management changes, should our results apply to 100% of the relevant population, or 85%? And if the latter, what is the appropriate method of identifying those whose behaviour is indistinguishable from random choices?

Conclusion

We have identified an issue of interpretation of HCL models that may lead to erroneous conclusions about significance: reported preference parameters and their significance levels are confounded by the normalization of variables explaining scale heterogeneity. It is now commonplace to recognise that

[2]Simulation results (available from the authors) suggest that the proportion of the sample that may appear to act randomly could be large, but the heteroscedastic conditional logit model still reports significant part-worths.
[3] Note that this was the largest subsample for which this result was true: increasing the cut-off to the next value of scale (0.41), gave significant parameter estimates.

one cannot compare preference parameters across models because of the confounding effect of scale (Louviere and Eagle 2006), but here we show the same effect is manifesting itself within a single model. Partworths avoid this problem, and we recommend that HCL models are estimated in WTP space where measures of preference parameters are invariant to normalization. In WTP space what changes is the intercept of the relationship defining scale/error variance, which is of little empirical interest.

However, our analysis raises a different insight: the HCL model contains a maintained hypothesis that all respondents hold the same preferences, and differ only by their error variance which can be explained by observable characteristics. But for some, the error variance may be sufficiently large to make their choices indistinguishable from random choices. Should this portion of the sample be used when generating aggregate welfare measures?

References

Burton, M., S. Jennings, Fragnol, J. Marre, S. Parades, S. Pascoe, A. Rogers and S. Yamazaki (2015) The south east commonwealth marine reserves network – public knowledge, perceptions and values survey. NERP Marine Biodiversity Hub, June 2015.

Davis, K., M. Burton and M.E. Kragt (2016) *Discrete choice models: scale heterogeneity and why it matters*. Working Paper 1602, University of Western Australia, Perth

Fiebig, D.G., M.P. Keane, J. Louviere and N. Wasi (2010) The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science,* 29: 393-421.

Gu,Y., A.R. Hole and S. Knox (2013) Fitting the generalized multinomial logit model in Stata. *The Stata Journal*, 13(2): 382-397

Hensher, D., J. Louviere and J. Swait (1999) Combining sources of preference data. *Journal of Econometrics,* 89: 197-221.

Hole, A.R. (2006a) Small-sample properties of tests for heteroscedasticity in the conditional logit model. *Economics Bulletin,* 3: 1-14.

Hole A.R. (2006b) CLOGITHET: Stata Module to Estimate Heteroscedastic Conditional Logit Models. Statistical Software Components S456737

Louviere, J. and T. Eagle (2006) Confound it! That pesky little scale constant messes up our convenient assumptions. *Proceedings of the Sawtooth Software Conference*: 211-228.