Model invariance when estimating random parameters with categorical variables

Michael P. Burton
UWA School of Agriculture and Environment, The University of Western Australia,
Crawley, WA 6009, Australia

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

12 May 2018
Working Paper 1804
UWA Agricultural and Resource Economics
http://www.are.uwa.edu.au

Citation: Burton, M.P (2018), Model invariance when estimating random parameters with categorical variables Working Paper 1804, Agricultural and Resource Economics, The University of Western Australia, Crawley, Australia.

Model invariance when estimating random parameters with categorical variables

Michael Burton

Abstract

This paper shows that econometric models that include categorical variables are not invariant to choice of ‘base’ category when random parameters are estimated, unless they are allowed to be correlated. We show that the invariance can lead to significant increases in Type I errors, and distortions in the implied behaviour of respondents. We hypothesise that these biases may influence the economic policy implications of published models that contain this error, but it’s impossible to be sure without re-estimating the model correctly.

Key words: Random parameter, mixed logit, categorical variables

JEL classifications: Q51 Valuation of Environmental effects. C25 Discrete Regression and Qualitative Choice Models

Model invariance when estimating random parameters with categorical variables

Introduction

Typically econometric models that include dummy variables to represent levels of categorical variables are invariant to which level is used as the base. Overall measures of model performance (i.e. LL values) do not change, and the results for any choice of base can be retrieved from any other model that uses a different base. Daley et al (2016) show that there is also invariance across a dummy and effects coding.

What does not seem to be recognised is that this does not always hold when estimating random parameter models: or more precisely, that there are specific requirements for invariance to be achieved. One has to estimate a full covariance model. If one does not have a full covariance representation then different choices about the base category mean one is effectively estimating different models.

To motivate this, consider a simple discrete choice framework, where respondents make choices across 2 alternatives. Each alternative consists of two attributes, a continuous variable, x, and a categorical variable F with 3 levels. The data generating process (DGP) consists of respondents evaluating the deterministic component of utility as

$$V_i = \beta_x X + \beta_{1i}[b_1, \sigma_1^2]F_1 + \beta_{2i}[b_2, \sigma_2^2]F_2 + \beta_{3i}[b_3, \sigma_3^3]F_3 \qquad (1)$$

Subscripts for alternative and choice set number are suppressed. In this general model, there is heterogeneity in preferences surrounding each level of the categorical variable, with the individual value β being drawn from a normal distribution with mean of b and variance of σ^2. For illustration purposes, we assume that in the true data generating process the random parameters are uncorrelated across levels of the categorical variable. Estimation of a model involving this utility function faces the standard problem of linear dependency across the levels of F, and the need to drop one, so that estimates of marginal utility for the remaining levels are relative to a base. If the analyst ignores the heterogeneity and assumes fixed coefficients, the standard invariance results holds: importantly, the invariance result we show in the next section is not due simply to misspecification error (because the fixed parameter model is also misspecified, and yet does not exhibit invariance), but specifically to the incomplete specification of the random parameter representation.

If we take F_1 as the baseline category then $F_1 = 1 - F_2 - F_3$ and hence the estimation specification is:

$$V_i = \beta_x X + \beta_{1i}[b_1, \sigma_1^2] + (\beta_{2i}[b_2, \sigma_2^2] - \beta_{1i}[b_1, \sigma_1^2])F_2 + (\beta_{3i}[b_3, \sigma_3^2] - \beta_{1i}[b_1, \sigma_1^2])F_3 \qquad (2)$$

Because a discrete choice model works in utility differences, the individual specific level of β_{1i} has no effect on choices and hence can be dropped:

$$V_i = \beta_x X + (\beta_{2i}[b_2, \sigma_2^2] - \beta_{1i}[b_1, \sigma_1^2])F_2 + (\beta_{3i}[b_3, \sigma_3^2] - \beta_{1i}[b_1, \sigma_1^2])F_3 \qquad (3)$$

i.e. what one estimates for the categorical variables are estimates of marginal utility relative to the base. What is clear from 3 is that, despite there being no correlation in the underlying marginal utilities, in the normalised estimating model, there will be a correlation between the random parameter distributions for F_2 and F_3 because they both contain the term for β_1. Alternatively, if one estimates a model assuming independence between the random parameters in (3), then that can only be true if the variance σ_1^2 =0 i.e. the base category has no variation in marginal utilities.

This immediately reveals why there is model variance as one changes the baseline category, if one estimates independent random parameters for the remaining category levels: one is implying a different stochastic structure for the marginal utilities in the data generating process each time, which cannot be internally consistent.

Simulations

To illustrate the consequences of this we conduct a simple numerical example. We assume that we have an experiment with a 1000 respondents, each answering 10 choice sets, each with two alternatives. The level of X is drawn randomly from the unit interval, and the level of F is also selected at random from 1,2,3. We employ a large sample and a random design to ensure that our results cannot be the result of any particular design artefact, and a simulated gumbel error term so that a conditional logit model is appropriate for the DGP.

Table 1 gives the parameters for the distributions employed

Table 1 Distributions for marginal utilities used in simulation

	b	σ
β_x	1	0
β_1	1	0
β_2	2	2
β_3	3	3

Note that we assume that the marginal utility for level 1 of F is in fact constant, and has no heterogeneity within it.

Table 2 shows results from estimating a standard fixed coefficient conditional logit model[1], changing the base category for F in each. Obviously results are stable to choice of base category, and one can retrieve any value in a model from any other model.

Table 2 Fixed coefficient conditional logit models

	Model 1	Model 2	Model 3
X	0.602 *(0.501 0.703)*	0.602 *(0.501 0.703)*	0.602 *(0.501 0.703)*
F1		-0.603 *(-0.674 -0.531)*	-0.896 *(-0.969 -0.822)*
F2	0.603 *(0.531 0.674)*		-0.293 *(-0.363 -0.222)*
F3	0.896 *(0.822 0.969)*	0.293 *(0.222 0.363)*	
ll	-6549.2861	-6549.2861	-6549.2861

95% CI in parenthesis

Table 3 repeats the process, but now assumes that the factors included in the model have independent random parameters. Note that the LL values now differ across models (i.e. the model is not invariant to the choice of base level) and estimates from one model cannot be retrieved from the others (i.e. the coefficient for F3 in Model 2 is not the negative of the coefficient for F2 in Model 3).

[1] All estimation was undertaken in Stata 14, using the clogit command for models with fixed coefficients, and the xtmelogit command where random parameters are specified. In the case of choice sets with two alternatives, by defining attributes as differences across alternatives, one can frame a DCE as a standard logit model. This enables one to use xtmelogit, which we find gives identical results to alternative programs (i.e. mixlogit) but is considerably faster, computationally, given the large number of simulations used.

Table 3 Independent, random parameter models

	Model 1	Model 2	Model 3
X	0.866	0.859	0.844
	(0.733 0.999)	*(0.727 0.991)*	*(0.712 0.976)*
F1		-1.056	-1.632
		(-1.221 -0.892)	*(-1.830 -1.434)*
F2	1.055		-0.484
	(0.875 1.236)		*(-0.702 -0.267)*
F3	1.827	0.852	
	(1.580 2.073)	*(0.596 1.108)*	
Estimates of σ			
σ(F1)		1.795	2.165
		(1.602 2.011)	*(1.946 2.408)*
σ(F2)	2.110		2.870
	(1.899 2.345)		*(2.608 3.160)*
σ(F3)	2.919	3.318	
	(2.637 3.231)	*(3.011 3.656)*	
ll	-5707.0781	-5781.2897	-5894.4946

95% CI in parenthesis

Table 4 repeats the process but now allows for correlation in the random parameters. One now has invariance in overall model fit (the LL values are very close) and there is a correspondence between preference parameters across models (e.g. F1 in Model 2 is the negative of F2 in Model 1)

Table 4 Correlated, random parameter models

	Model 1	Model 2	Model 3
X	0.866	0.866	0.866
	(0.733 0.999)	*(0.733 0.999)*	*(0.733 0.999)*
F1		-1.048	-1.822
		(-1.228 -0.867)	*(-2.068 -1.576)*
F2	1.048		-0.775
	(0.867 1.228)		*(-1.049 -0.500)*
F3	1.822	0.775	
	(1.576 2.068)	*(0.500 1.049)*	
Estimates of σ and covariance's			
σ(F1)		2.095	2.905
		(1.880 2.333)	*(2.621 3.220)*
σ(F2)	2.094		3.658
	(1.880 2.332)		*(3.337 4.008)*
σ(F3)	2.905	3.658	
	(2.621 3.220)	*(3.338 4.008)*	
covariance	-0.045	0.609	0.820
	(-0.160 0.071)	*(0.529 0.678)*	*(0.774 0.858)*
ll	-5706.7813	-5706.7792	-5706.7792

95% CI in parenthesis

What this example illustrates is that one can achieve model invariance when including categorical variables with random parameters, but one has to allow for correlations in random parameters, and not assume independence. The majority of papers do not do this, and one can find examples of this practice in a number of journals (e.g. Zhang et al 2018; Zhu et al 2015; Oehlmann et al, 2017; Thiene et al 2017, Ozdemir et al 2016)

A different question is the consequence of the invariance. Does this cause any meaningful bias in inferences about preferences?

To test this a second set of simulations were undertaken where the standard deviation of the distribution of F3 was changed (from 2 to 6), and Model 3 alone estimated (i.e. F3 used as the base category). 500 simulated data sets were drawn for each value of the standard deviation of F3. The results in Table 5 show the mean value for the estimated parameter, divided by the true parameter. The values in parenthesis show the proportion of times the null hypothesis of no difference between the estimated value and the true value is rejected (i.e. the rate at which one commits a Type 1 error). As check on estimation precision, the first column reports the results when σ(F3)=2, and the correct, correlated model is estimated. In that case, normalised parameter estimates are all close to unity, and the error rate is close to 5% in all cases.

When the uncorrelated random parameter model is estimated, the estimates for the variable with a fixed parameter (X) are relatively stable: close to the true value and with a Type I error rate close to 0.05, although it is notable that in all cases the mean effect is less than 1 (implying some attenuation) and the Type I error rate grows substantially as the standard deviation gets higher. However, for all other variables there is evidence of considerable attenuation in estimates of both the mean and standard deviation of the marginal utilities, and these effects get larger the larger standard deviation of F3. For a value of the standard deviation of F3=3 (i.e. the standard deviation equals the mean of the variable) one is seeing estimates of the mean of the distribution of F1 and F2 at 89 and 66% of their true values, while the estimated heterogeneity in those preferences is only 75/77% of the true value.

Table 5 Estimates of average normalized coefficient (relative to true value) and proportion of times null hypothesis of equality with true parameter rejected (500 simulations). F3 as base, estimation assumes independent random distributions for parameters on F1 and F2.

		σ(F3)				
	$2^{\$}$	2	3	4	5	6
X	0.99 (0.046)	0.98 (0.044)	0.98 (0.058)	0.98 (0.070)	0.98 (0.064)	0.98 (0.078)
F1	1.00 (0.052)	0.92 (0.436)	0.89 (0.520)	0.86 (0.606)	0.83 (0.660)	0.81 (0.714)
F2	1.00 (0.036)	0.78 (0.602)	0.66 (0.846)	0.54 (0.944)	0.42 (0.976)	0.32 (0.984)
σ(F1)	1.00 (0.044)	0.79 (0.970)	0.75 (1.000)	0.71 (1.000)	0.66 (1.000)	0.62 (1.000)
σ(F2)	1.00 (0.052)	0.82 (0.982)	0.77 (1.000)	0.73 (1.000)	0.68 (1.000)	0.64 (1.000)

$ model estimated with correlated random effects as a comparison

Evaluation using real data

As a final test, we investigate the implications of misspecification using a real data set. In a forthcoming paper (Iftekhar et al 2018) we report results from a discrete choice model evaluating public preferences for land allocation in water treatment plant buffer zones. 4 land uses are possible (nature conservation, industry, commercial and agriculture) and as % measures they have to add to 100, and hence one has to be dropped from estimation for identification. This is analogous to the model outlined above. In that paper we estimated a model in WTP space, with correlation of the random effects of the three included land types and the status quo. In Table 6 we report a similar model estimated on that data, but in preference space, and without a random status quo effect (so the model matches that above, with random effects only on the variables of interest, and changes in model outcomes only due to the use of uncorrelated random effects and changing baselines). We re-estimate the model 4 times, with correlated (models A1-A4) and uncorrelated (B1-B4) random parameters, and changing the base in each case. Note that the interest here is in the implications of changing the base group when not allowing for correlated random parameters, not wider specification issues.

The LL values for A1-A4 are all close, implying invariance, but those for B1-B4 are not, confirming that base category choice is changing the fundamentals of the model.

Comparing the pairs of models, there is a tendency for the estimates of standard deviations of the distributions to be reduced in the models without correlated errors, as seen in Table 5 above, but this is not universal, although there are no large increases in estimates, but some substantial reductions (i.e. σ(Ag) halves from 0.046 to 0.023 in models A1 and B1). The estimates of the mean of the distribution also change, and of most importance are not symmetric i.e. the value of recreational land use relative to commercial in model B2 is 0.005, and not significant, while that of commercial land use relative to recreational in B3 is -0.026, and is significant. The estimate of population average relative values of these land uses is significantly changed, simply by changing the base category.

Table 6 Mixlogit results for alternative choices of base category: A1-A4 using correlated random parameters, B1-B4 uncorrelated random parameters

	Base=Nat		Base =Com		Base=Rec		Base=Ag	
	A1	B1	A2	B2	A3	B3	A4	B4
Cost	-0.021	-0.020	-0.021	-0.021	-0.021	-0.022	-0.021	-0.021
	[-0.026 -0.016]	[-0.025 -0.016]	[-0.026 -0.016]	[-0.026 -0.017]	[-0.056 -0.016]	[-0.027 -0.017]	[-0.026 -0.016]	[-0.026 -0.017]
Sq	0.798	0.846	0.794	0.714	0.789	1.156	0.790	0.811
	[0.487 1.108]	[0.546 1.145]	[0.484 1.103]	[0.430 0.999]	[0.481 1.097]	[0.836 1.476]	[0.481 1.099]	[0.516 1.106]
Nat			0.051	0.052	0.040	0.029	0.042	0.052
			[0.038 0.064]	[0.042 0.063]	[0.025 0.055]	[0.016 0.043]	[0.030 0.054]	[0.041 0.062]
Com	-0.052	-0.054			-0.011	-0.026	-0.009	-0.003
	[-0.065 -0.039]	[-0.065 -0.043]			[-0.027 0.005]	[-0.040 -0.013]	[-0.019 0.001]	[-0.013 0.007]
Rec	-0.040	-0.043	0.011	0.005			0.002	0.006
	[-0.055 -0.025]	[-0.058 -0.028]	[-0.004 0.027]	[-0.009 0.020]			[-0.014 0.019]	[-0.009 0.021]
Ag	-0.043	-0.050	0.009	0.002	-0.002	-0.021		
	[-0.054 -0.031]	[-0.060 -0.040]	[-0.001 0.019]	[-0.007 0.011]	[-0.019 0.014]	[-0.035 -0.007]		
σ(Nat)			0.070	0.046	0.078	0.067	0.046	0.033
			[0.055 0.086]	[0.032 0.060]	[0.063 0.093]	[0.054 0.078]	[0.027 0.064]	[0.017 0.049]
σ(Com)	0.070	0.050			0.090	0.065	0.047	0.048
	[0.055 0.086]	[0.038 0.062]			[0.074 0.105]	[0.052 0.077]	[0.033 0.062]	[0.036 0.060]
σ(Rec)	0.079	0.081	0.089	0.080			0.099	0.083
	[0.064 0.094]	[0.068 0.094]	[0.074 0.105]	[0.067 0.093]			[0.082 0.116]	[0.069 0.096]
σ(Ag)	0.046	0.023	0.047	0.033	0.098	0.067		
	[0.028 0.063]	[0.005 0.041]	[0.033 0.061]	[0.021 0.046]	[0.081 0.115]	[0.055 0.080]		
LL	1360.2251	-1375.36	1360.4543	-1385.157	1360.3235	1458.3070	1360.4086	-1373.0679

Nat=nature and conservation, Com=commercial use, Rec=recreational use, Ag=Agricultural use

95% CI in parenthesis

Table 7 Estimates of mean WTP for uncorrelated models, for different baselines

	Baseline category			
	Nat	Com	Rec	Ag
Nat		2.48 [1.81 3.14]	1.33 [0.78 1.89]	2.41 [1.72 3.10]
Com	-2.66 [-3.39 -1.93		-1.21 [-1.94 -0.47]	-0.15 [-0.60 0.31]
Rec	-2.10 [-2.79 -1.41]	0.26 [-0.43 0.94]		0.29 [-0.43 1.01]
Ag	-2.47 [-3.18 -1.75]	0.09 [-0.33 0.50]	-0.95 [-1.69 -0.21]	

95% CI in parenthesis

This effect is highlighted in Table 7, which reports the estimate of mean WTP for the uncorrelated models (-1 times ratio of mean parameter estimate to cost parameter estimate). The lack of symmetry is clear, and there are 3 cases where one has large changes in inference about relative land use values: recreational land use relative to nature land use (-$2.10 v $1.33) recreational land use relative to commercial land use ($0.26 v -$1.21) and agriculture v recreational use (-$0.95 v $0.29). The fact that all three cases involve recreational use suggest that possibly preferences towards recreational use have the greatest heterogeneity, and hence the greatest impact when induced correlation is ignored. That is confirmed by inspection of the estimates of standard deviations is Table 6, for models A1-A4, where the highest 6 of 12 estimates involve the recreational attribute. These changes in estimates of relative WTP for land use types could be sufficient to change policy inferences: from the model with recreation as the base, agriculture is clearly inferior to recreation (at the mean), while when agriculture is the base, one infers that respondents are indifferent between recreation and agriculture as a land use. The correlated models suggest that the latter result is correct.

Conclusions

Bottom line: mis-specifying the random parameter distribution influences both the estimated mean and variance of the marginal utilities. Whether this matters or not depends on any particular application: does it change the policy conclusion? One could argue that this is a general issue: choices about how to model marginal utilities are likely to influence outcomes. But normally an author will have made a justification of why the choice was made, and possibly reflect on the consequences of that. No-one expects to have to justify why a particular level of a categorical variable is used as the base, as typically that choice is innocuous. However, this note shows that this is not the case if independence in random effects is assumed. A minimum requirement for any model (mis-specified or not) is that it should be invariant to an arbitrary choice about how categorical variables are included in the model, and for many published papers, that's not true.

References

Daly, A., Dekker, T. and Hess, S. (2016) Dummy coding vs effects coding for categorical variables: Clarifications and extensions *Journal of Choice Modelling* 21: 36–41

Iftekhar, M. S., Burton, M., Zhang, F., Kininmonth, I., Fogarty, J., 2018. Understanding social preferences for land use in wastewater treatment plant buffer zones. *Landscape and Urban Planning*. Revised paper submitted.

Oehlmann,M, Meyerhoff,J., Mariel,P., Weller,P. (2017) Uncovering context-induced status quo effects in choice experiments *Journal of Environmental Economics and Management* 81: 59-73

Ozdemir,S., F. Reed Johnson,F.R., and Whittington,D. (2016) Ideology, public goods and welfare valuation: An experiment on allocating government budgets, *Journal of Choice Modelling,* 20:61-72

Theine,M., Swait,J. and Scarpa,R. (2017) Choice set formation for outdoor preferences: the role of motivations and preference discrimination in site selection for the management of pubic expenditures on protected areas *Journal of Environmental Economics and Management* 81:152-173

Zhang,W and Sohngen,B. (2018) Do US anglers care about harmful algal blooms? A discrete choice experiment of Lake Erie recreational anglers *American Journal of Agricultural Economics* 100(3): 868-888

Zhu,C., Lopez,R., and Liu,X. (2015) Information cost and consumers choices of healthy foods *American Journal of Agricultural Economics* 98(1) 41-53