

Correcting for Non-response Bias in Discrete Choice Models: A Two-Stage Mixed Logit Model

Daniel Hellerstein
USDA/Economic Research Service

For the Oct 2, 2006 RFF Workshop...
Sample Representativeness: Implications for Administering and Testing Stated Preference Surveys

[2] State of regional resources

Enter the resources (one of the rural communities). Circle one corresponding item concerning the degree of the inclination of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. For example, 1 for terraced paddy fields, valley bottom fields, etc.

Observable differences

Example:

the average income of the survey is not the same as the general population

In most cases, I contend that ...

differences in measurable factors should not yield biased conclusions.

[3] Conservation of regional resources

If the following regional resources are conserved, circle "grounds for the conservation" and "conserving organization", which best applies, and circle all "purpose" of conservation. And in a case where there are regional resources but they are not conserved, circle "not conserved".

Example:

Although OLS models are most efficient when the X (independent) variables are broadly distributed, a clumpy distribution of X will not cause inconsistent estimates.

Example:

Within-sample data, when combined with observation-specific population weights, can be used to make overall population predictions.

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

		No. undertaken	Undertaken	Undertaken as an agricultural farmer
Agro-ecotourism and agro-ecological parks and agro-ecological gardens	01	☉	☉	☉
Agro-ecological parks and agro-ecological gardens	02	☉	☉	☉
Agro-ecological parks and agro-ecological gardens	04	☉	☉	☉

2. Facilities using regional resources

Enter number of "users" and "number of users" of corresponding facilities that use regional resources, and circle all "number" corresponding to the state of premises in forest park.

		Number of users	Number of users (persons)
Shops directly selling regional products	01	☉	☉
Allotment gardens	02	☉	☉
Shops directly selling regional products	03	☉	☉
Shops directly selling regional products	05	☉	☉
Shops directly selling regional products	06	☉	☉
Camp grounds	08	☉	☉
Others	09	☉	☉

Agriculture land	01
Terraced paddy field	02
Valley bottom paddy field	03
Forest	04
Pond for irrigation/lake	05
River/canal	06
Agricultural water supply channel	07

No. undertaken	Undertaken	Undertaken as an agricultural farmer
☉	☉	☉

[2] State of regional resources

Enter the resource code of the rural communities. Circle one corresponding item concerning the degree of the inclination of land where most of the cultivated land paddy field, orchard land, and upland fields respectively are located for terrace or terraced paddy fields, valley bottom

Unobservable differences

Example:

due to idiosyncratic personality factors, people who return a survey may be the most avid users of a resource

For obvious reasons, direct use of measurable characteristics (of the sample & population) won't help control for unobservable differences.

- However, indirect measures can be used.
- In particular, sample selection models.

		1	2	3	4	5	6	7	8
Total land area	01								
Paddy field	02								
Terraced paddy field	03								
Valley-bottom paddy field	04								
Land under permanent crops	05								
Upland field	06								
Forest and grazing land	07								
Pond for irrigation	08								

[3] Conservation

If the following resources are conserved, circle "grounds for the conservation organization", "business purposes of conservation," and in a case where there are regional resources but they are not conserved, circle "not conserved"

		Grounds for conservation		Conserving		Purpose of								
		Protect region	Other than region	1	2	1	2	3	4	5	6			
Agriculture land	01													
Terraced paddy field	02													
Valley-bottom paddy field	03													
Forest	04													
Pond for irrigation/lake	05													
River/canal	06													
Agriculture water supply/irrigation system	07													

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

		No. undertaken	Undertaken	Undertaken as an agricultural target
Agriculture land and orchard land	01			
Terraced paddy field	02			
Valley-bottom paddy field	04			

2. Facilities using regional resources

Enter number of users and number of users of corresponding facilities that use regional resources, and circle numbers corresponding to items.

		Number of facilities	Number of users (persons)	
			1	2
Shops directly selling regional products	01			
Allotment gardens	02			
Directly selling	03			
Indirectly selling	04			
Home raising and selling	05			
Education	06			
Forest park	07			
Camp ground	08			
Others	09			

	No. promoters provided	Promoters provided	Promoters farmer-free
10			

Sample Selection Models:

Basic Idea:

1. Respondents are “selected into” the sample
2. A 1st-stage model predict who will be selected
3. A 2nd-stage model predicts the variable of interest
4. The random components (RV) in both stages are potentially correlated

Which means ...

1. The first-stage RV of **participants** may *not* have an expected value of zero.
2. Hence, if the two RVs are correlated, the 2nd-stage RV may **not** have an expected value of zero.
3. Hence, the 2nd stage model must control for such inconveniences.

Example: the Heckit Model

Assumptions:

$$(1) Z_i^* = \gamma W_i + \tau_i$$

$$Z_i = 1 \text{ if } Z_i^* > 0$$

$$\text{Prob}(Z_i = 1) = \Phi(\gamma W)$$

$$(2) Y_i = \beta X_i + \varepsilon_i$$

Y_i is observed only if $Z_i = 1$

$$\tau_i \text{ and } \varepsilon_i \sim \text{bivariate normal}(0,0,1, \sigma_\varepsilon, \rho)$$

Model:

- 1) Using data from those in the sample and those not in the sample, use a Probit to estimate γ
- 2) For each observation in the sample, compute: $\lambda = \phi(\gamma w_i) / \Phi(\gamma w_i)$
- 3) Estimate β and β_λ using a linear regression of Y on X and λ

Notes: $\beta_\lambda = \rho \sigma_\varepsilon$

λ is some times called an *inverse - mills ratio*

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects

$$Z_i = 0 \text{ if } Z_i^* < 0$$

$$\text{Prob}(Z_i = 0) = 1 - \Phi(\gamma W)$$

2. Facilities using regional resources

Circle numbers of corresponding facilities that use regional resources, and circle numbers corresponding to the state of premises in forest parks

Number of users (persons)

Shops directly selling regional products

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

Agricultural parks

A digression: the MNL model:

Individual i can choose from $1 \dots J$ possible alternatives.

$$m_j = x_j \beta + v_j$$

Respondent chooses j^* such that $m_{j^*} > m_{j'}$ [$j' = 1, \dots, j^* - 1, j^* + 1, \dots, J$]

If v_j has an extreme - value distribution, than the probability of choosing alternative j is:

$$\frac{\exp(x_j \beta)}{\sum_j \exp(x_j \beta)}$$

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

No. undertaken	Undertaken	Undertaken as an agricultural project
		⊕
		⊕
		⊕
		⊕

[2] State of regional resources

Circle numbers corresponding to a use of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. Enter number of locations for terraced-paddy fields, valley bottom paddy fields and small reservoirs.

Number of locations	Area (ha) (pond for irrigation (a))	Flatland	Semi-slope	Slope-type

Total area

Paddy

Ter

Valley

Upland

Forest

Pond

Irrigat

[3] Co

If the

for the

And

not conc

		Prevent region	Co-oper village region	Agreement	Regional public body	Depend on local gov	Conservation of land	Observation of water resources	Conservation system the system the local	Conservation of landscape	Observation of local resources	Others	Not concerned
		⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Agricultural land	01	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Terraced paddy fields	02	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Valley bottom paddy fields	03	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Forest	04	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Pond for irrigation/lake	05	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
River/canal	06	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Agricultural water supply/irrigation system	07	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕

		04											
Preparation of full set	Agricultural parks	04
	House for training and visiting reference materials on forest	05
	Forestry for education	06
	Forest park	07
	Camp-ground	08
Others	09	

		No promoter is provided	Promoter is provided	Promoter farmer-free
		⊕	⊕	⊕
10		⊕	⊕	⊕

[2] State of regional resources

Enter the resource crop of the rural communities. Circle and corresponding form concerning the degree of the inclination of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. (circle 1 for terrace, 2 for terraced paddy fields, valley bottom)

BIG PROBLEM:

For many models, the “order of alternatives” is arbitrary.

Example:

In a freshwater recreation model, the choice set of accessible waterbodies is different for individuals in different parts of the nation (though the same set of measured variables exist for all waterbodies).

[3] Conservation of regional resources

If the following regional resources are conserved, circle the corresponding number for the conservation policy, and circle all applicable. And in a case where there are regional resources but they are not conserved, circle not conserved.

Hence, across all observations, for any “alternative j”...

there is no obvious reason for the same correlation to exist between u and v_j

A possible workaround is to carefully order alternatives, so that each individuals j'th alternative is somehow similar.

Instead of this, I consider a 2-stage mixed logit model

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

		No. undertaken	Undertaken	Undertaken as an agricultural target
Access to irrigation canals	01	☉	☉	☉
Access to irrigation canals	03	☉	☉	☉
Access to irrigation canals	04	☉	☉	☉

		Number of individuals	Number of users (persons)	
		1	2	
Shops directly selling regional products	01	•	•	•
Allotment gardens	02	•	•	•
Agricultural parks	04	•	•	•
Home growing and eating local products in food	05	•	•	•
Forestry for education	06	•	•	•
Forest park	07	•	•	•
Others	09	•	•	•

		No. undertaken	Undertaken	Undertaken as an agricultural target
Access to irrigation canals	10	☉	☉	☉

A digression: the Mixed MNL model:

Starting with the standard MNL: $prob_j = \frac{\exp(x_j\beta)}{\sum_j \exp(x_j\beta)}$

However, each individual (n) has unique coefficient values (β_n)

The kth coefficient in β_n is modeled as: $\beta_{k,n} = \tilde{\beta}_k + \eta_{n,k}$

where...

$\tilde{\beta}_k$ is the systematic portion of $\beta_{n,k}$

$\eta_{n,k}$ is an observation & coefficient specific random variable drawn from a MVN(0,Ω) random variable

The mixed logit: $prob_j = \frac{\sum_r \left(\frac{\exp(x_j\beta_n^r)}{\sum_j \exp(x_j\beta_n^r)} \right)}{R}$

where

$$\beta_n^r = \tilde{\beta} + \Omega \xi_{nr}$$

$\tilde{\beta}, \Omega$ = estimated values of $\tilde{\beta}$, estimated value of the "square root" of Ω

ξ_{nr} = randomly generated Kx1 vector of standard normal errors.

R = R different replications (different draw of ξ_{nr} for each r = 1...R replications)

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

	No. undertaken	Undertaken	Undertaken as an agricultural market
01	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
02	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
03	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
04	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2. Use of regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

Number of facilities	Number of users (persons)	
	1	2
01	<input type="checkbox"/>	<input type="checkbox"/>
02	<input type="checkbox"/>	<input type="checkbox"/>
03	<input type="checkbox"/>	<input type="checkbox"/>
04	<input type="checkbox"/>	<input type="checkbox"/>
05	<input type="checkbox"/>	<input type="checkbox"/>
06	<input type="checkbox"/>	<input type="checkbox"/>
07	<input type="checkbox"/>	<input type="checkbox"/>
08	<input type="checkbox"/>	<input type="checkbox"/>
09	<input type="checkbox"/>	<input type="checkbox"/>

	No. provided	Provided	Provided as an agricultural market
10	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[2] State of regional resources

Enter the resource area of the rural communities. Circle and corresponding form concerning the degree of the inclination of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. Enter number of locations for terraced paddy fields, valley bottom paddy fields and small reservoirs.

The 2-stage mixed logit.

Basic idea:

Instead of assuming that

τ and v_j are BVN distributed ...

Assume that

τ and η_k are BVN distributed

In other words, the values of one (or several) of the varying parameters will be correlated with the first stage decision.

This assumption removes the need to have special ordering for alternatives!

	Number of locations	Area (ha)																	
		1	2	3	4	5	6	7	8	9	10	11	12						
Total land area	01																		
Paddy field	02																		
	03																		
Valley-bottom paddy field	04																		
Land under permanent crops	05																		
Upland field	06																		
Forest and grazing land	07																		
Pond for irrigation	08																		

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

	No.	No. of users (persons)		
		1	2	3
Accessibility of roads using drainage systems, etc.	01			
Interchange through direct trading between producers and consumers	02			
Interchange through indirect trading through markets	03			
Interchange through direct trading through markets	04			

2. Facilities using regional resources

Circle number of "number of users" and "number of users" corresponding facilities that use regional resources, and circle numbers corresponding to the state of promenades in forest parks.

	No.	Number of users (persons)	
		1	2
Specialized facilities	01		
General facilities	02		
Agricultural parks	03		
Facilities for viewing and visiting (forest parks, etc.)	04		
Forest park	05		
Camp ground	06		
Others	07		

	No.	No. of users (persons)		
		1	2	3
Forest park	08			
Camp ground	09			
Others	10			

An operational version of the 2-stage mixed logit.

Assuming a diagonal Ω covariance matrix

then, where the mixed uses...

$$\beta_{n,k}^r = \beta_k + \sigma_k \xi_{nr,k}$$

the 2 - stage mixed logit uses...

$$\beta_{n,k}^r = \beta_k + \left(\sigma_k \rho_k \lambda \right) + \left(\left(\sigma_k \sqrt{1 - \rho_k^2} \delta \right) \xi_{nr,k} \right)$$

where...

σ_k, ρ_k : estimates of the sd of η_k , and the correlation between η_k and τ

As described before ... λ and δ are observation specific values derived from a first - stage PROBIT :

$$\alpha_z = \gamma w$$

$$\lambda(\alpha_z) = \phi(\alpha_z) / \Phi(\alpha_z)$$

$$\delta(\alpha_z) = \lambda(\alpha_z)(\lambda(\alpha_z) + \alpha_z)$$

[2] State of regional resources

Enter the resource state of the rural communities. Circle one corresponding item concerning the degree of the inclination of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. (1: No inclination, 2: Moderate, 3: Strong) (1: No, 2: Moderate, 3: Strong)

BUT DOES IT WORK?

that is still under investigation.

Some results from artificial data ...

- 1600 observations
- 8 alternatives
- A first stage model determines whether a response is observed.
- The rv in the first stage model is correlated with the rv used to generate the varying parameter.

[3] Conservation of regional resources

If the following regional resources are to be conserved, circle "grounds for the conservation" organization, "business activities", and "types of conservation". And in a case where there are regional resources but they are not conserved, circle "No".

	Agriculture land	Grounds for conservation			Conservation			Purpose of conservation									
		Private region	Open-air region	Agreement	Regional activity	Conservation activity	Conservation of nature	Shops, etc. (recreation)	Forest parks	Forest for education	Others (recreation)	Others (recreation)	Others (recreation)	Others (recreation)			
	01	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	02	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	03	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	04	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	05	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	06	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10
	07	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9	10

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources
Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

		No. undertaken	Undertaken	Undertaken as an agricultural target
Shops directly selling local products and services	01	1	2	3
Shops through direct trading between producers and consumers	02	1	2	3
Local products sold through direct trading	03	1	2	3
Local products sold through indirect trading	04	1	2	3

2. Facilities using regional resources
Enter number of "users" and "number of users" of corresponding facilities that use regional resources, and circle numbers corresponding to the state of promenades in forest parks.

	Number of users (persons)	State of promenades in forest parks	
		1	2
Shops directly selling local products and services	01	1	2
Shops through direct trading between producers and consumers	02	1	2
Local products sold through direct trading	03	1	2
Local products sold through indirect trading	04	1	2
Shops for housing and eating (forest restaurants in forest)	05	1	2
Forestry for education	06	1	2
Forest park	07	1	2
Camp ground	08	1	2
Others	09	1	2

	No. promenades provided	Promenades provided	Promenades free
10	1	2	3

Artificial data 1: correlation=0.5 | 45% rate of response | average WTP=108
MCI, CRsq and WTP computed using all respondents

Model	dβ1	dβ3	dβ3	MCI	CRsq	WTP
Mixed Logit all obs	11%	5%	%3	0.86	0.82	105
MNL, respondents	12%	8%	8%	0.86	0.83	110
Mixed Logit, respondents	50%	3%	2%	0.86	0.83	110
2-stage mixed logit	40%	1%	2%	0.87	0.96	109

dβk : % difference between actual beta (for k'th coefficient)

MCI : McFadden index (1=perfect fit)

CRsq : Cramer R-square (higher values indicate better fit)

WTP : average Estimated willingness to pay for choice opportunity

Actual beta 3.0 -0.2 2.0
 (sd) (2.6)

Artificial data 2: correlation=0.95 | 31% rate of response | average WTP=108
MCI, CRsq and WTP computed using all respondents

Model	dβ1	dβ3	dβ3	MCI	CRsq	WTP
Mixed Logit all obs	7%	5%	6%	0.83	0.78	104
MNL, respondents	53%	70%	70%	0.74	0.77	120
Mixed Logit, respondents	160%	5%	5%	0.78	0.78	125
2-stage mixed logit	3%	5%	10%	0.83	0.78	106

dβk : % difference between actual beta (for k'th coefficient)

MCI : McFadden index (1=perfect fit)

CRsq : Cramer R-square (higher values indicate better fit)

WTP : average Estimated willingness to pay for choice opportunity

Actual beta 3.0 -0.2 2.0
 (sd) (5.6)

Artificial data 3: correlation=0.58 | 43% rate of response | average WTP=23
MCI, CRsq and WTP computed using all respondents

Model	dβ1	dβ3	dβ3	MCI	CRsq	WTP
Mixed Logit all obs	12%	13%	12%	0.79	0.72	21
MNL, respondents	43%	3%	37%	0.74	0.71	24
Mixed Logit, respondents	90%	17%	20%	0.78	0.74	27
2-stage mixed logit	45%	17%	20%	0.79	0.74	24

dβk : % difference between actual beta (for k'th coefficient)

MCI : McFadden index (1=perfect fit)

CRsq : Cramer R-square (higher values indicate better fit)

WTP : average Estimated willingness to pay for choice opportunity

Actual beta 4.0 -0.4 1.0
 (sd) (5.6)

Artificial data 4: corr =0.57 & 0.13 | 46% rate of response | average WTP=33
MCI, CRsq and WTP computed using all respondents

Model	dβ1	dβ3	dβ3	dβ4	MCI	CRsq	WTP
Mixed Logit all obs	5%	12%	16%	20%	0.77	0.70	31
MNL, respondents	22%	39%	37%	35%	0.73	0.69	35
Mixed Logit, respondents	95%	32%	36%	48%	0.76	0.70	37
2-stage mixed logit	20%	32%	37%	77%	0.76	0.69	35

dβk : % difference between actual beta (for k'th coefficient)

MCI : McFadden index (1=perfect fit)

CRsq : Cramer R-square (higher values indicate better fit)

WTP : average Estimated willingness to pay for choice opportunity

Actual beta	4.0	-0.4	1.0	1.2
(sd)	(3.6)			(0.8)

[2] State of regional resources

Enter the resource area of the rural communities. Circle and corresponding form concerning the degree of the inclination of land where most of the cultivated land and paddy field, orchard land, and upland fields respectively are located. Enter number of locations for terraced-paddy fields, valley bottom paddy fields and small reservoirs.

Conclusions

- The 2-stage Mixed Logit offers a somewhat theoretically appealing method of controlling for non-response bias in discrete choice models.

- Under very limited testing, it has shown capacity to improve estimates.

- More work is needed. In particular, the current “heckit like” 2-stage estimator uses limited information when linking correlations to simulated draws . A more ambitious estimator could use a Gibbs sampler to estimate both stages nearly simultaneously

	Number of locations	Area (ha)									
		1	2	3	4	5	6	7	8	9	10
Total land area	01										
Paddy field	02										
	03										
	04										
Land of the permanent crops	05										
Upland field	06										
Forest and grazing land	07										
Pond for irrigation	08										

[4] State of use of regional resources

1. Undertaking interchange projects using regional resources

Circle numbers corresponding to a use of regional resources to undertake interchange projects by item.

	No. of undertakings	Undertaken		Undertaken as an agricultural farmer
		1	2	
Agricultural interchange using agricultural machinery	01			
Interchange through direct trading between producers and consumers	02			
Interchange through intermediaries	03			

2. Facilities using regional resources

Enter number of "small park" and "large park" of corresponding facilities that use regional resources, and circle of members corresponding to forest parks.

	Number of facilities	Number of users (persons)	
		1	2
Small park	01		
Large park	02		
Forest park	03		
Public bath	04		
Public library	05		
Public hall	06		
Public office	07		
Camp ground	08		
Others	09		

	Promenade		
	No. provided	Provided	Passage barrier-free
10			