

# 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 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", which purposes of conservation, and circle "yes" or "no" in a case where there are regional resources but they are not conserved, circle "not conserved"

		Grounds for conservation		Conserving organization		Purpose of conservation												
		1	2	1	2	1	2	3	4	5	6	7	8					
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/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.

		1	2	3	4	5	6	7	8	9	10
Regional products and services	01										
Regional products	02										
Regional services	03										
Regional products and services	04										

2. Facilities using regional resources

Enter number of users and number of users of corresponding facilities that use regional resources, and circle "yes" or "no" corresponding to items.

		Number of facilities	Number of users (persons)	
			1	2
Shops directly selling regional products	01			
Allotment gardens	02			
Regional products	03			
Regional services	04			
Regional products and services	05			
Regional products and services for education	06			
Forest park	07			
Camp ground	08			
Others	09			

	No promotion is provided	Promotion is provided	Promotion is free
10			

[2] State of regional resources

Enter the resource code 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 numbers corresponding to a use of regional resources to undertake interchange projects by item.

**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.

[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.

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

		Not undertaken	Undertaken	Undertaken as an agricultural target
1. Undertaking interchange projects using regional resources	01	⊙	⊙	⊙
2. Undertaking interchange projects using regional resources	02	⊙	⊙	⊙
3. Undertaking interchange projects using regional resources	03	⊙	⊙	⊙

[3] Conservation of regional resources

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

	Grounds for conservation			Regional public organization	Business application	Conservation	Purpose of conservation		
	Private region	Organic region	Agreement				Education	Recreation	Other
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	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙

2. Facilities using regional resources  
Circle numbers corresponding to a use of regional resources, and circle "grounds for the conservation organization" corresponding to the state of promenades in forest parks.

		Number of users (persons)	
		1	2
Shops directly selling regional products	01	⊙	⊙
Allotment gardens	02	⊙	⊙
Forest park	03	⊙	⊙
Forest park	04	⊙	⊙
Forest park for education	05	⊙	⊙
Forest park	06	⊙	⊙
Other	09	⊙	⊙

Passive farm-free	⊙
10	⊙

## 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  $\gamma$
- 2) For each observation in the sample, compute:  $\lambda = \phi(\gamma w_i) / \Phi(\gamma w_i)$
- 3) Estimate  $\beta$  and  $\beta_\lambda$  using a linear regression of  $Y$  on  $X$  and  $\lambda$

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

$\lambda$  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

Shops indirectly selling regional products

Agricultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural parks

Agri-cultural 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  $m_{j^*} > m_{j'} \quad [j' = 1, \dots, j^* - 1, j^* + 1, \dots, J]$

If  $v_j$  has an extreme - value distribution, then 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/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. promoters provided	Promoters provided	Promoters farmer-free
10	⊕	⊕	⊕

[2] State of regional resources

Enter the resource type 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, reservoirs.

**Simple case:**

If  $\tau$  and  $v_j$  are distributed as bivariate normal ...

Then

In the MNL estimator, include

$$\beta_{\lambda,j} * \lambda$$

term(s) in one (or more) of the  $m_j$ , and estimate  $\beta_{\lambda,j}$

Small problem:

What does inclusion of separate  $\beta_{\lambda,j}$  terms (for more than one alternative) imply about the independence of  $v_j$ ?

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

Resource	(ha)	(pond for irrigation) (a)	Paddy	Orchard	Upland
01	1	1	1	1	1
02	2	2	2	2	2
03	3	3	3	3	3
04	4	4	4	4	4
05	5	5	5	5	5
06	6	6	6	6	6
07	7	7	7	7	7
08	8	8	8	8	8

[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	1	2	3
02	1	2	3
03	1	2	3
04	1	2	3

2. Facilities using regional resources

Circle numbers corresponding to a use of regional resources for each of the following facilities that use regional resources. Circle numbers corresponding to the state of provision in forest parks.

	Number of facilities	Number of users (persons)
01	1	2
02	1	2
03	1	2
04	1	2
05	1	2
06	1	2
07	1	2
08	1	2
09	1	2

[3] Conservation of regional resources

If the following regional resources are conserved in the "grounds

or the conservation" area, circle one corresponding item for each resource.

And in a case where there are regional resources not conserved, circle

the number of the resource.

	Groups for conservation			Conserving organization			Purpose of conservation						Not conserved	
	Private region	Cooperative region	Agreement	Regional public body	Private organization	Others	Conservation of soil	Conservation of water resources	Conservation of natural resources	Conservation of forest	Conservation of landscape	Others		
Agricultural land	01	1	2	3	1	2	3	1	2	3	1	2	3	4
Terraced paddy field	02	1	2	3	1	2	3	1	2	3	1	2	3	4
Valley-bottom paddy field	03	1	2	3	1	2	3	1	2	3	1	2	3	4
Forest	04	1	2	3	1	2	3	1	2	3	1	2	3	4
Pond for irrigation/lake	05	1	2	3	1	2	3	1	2	3	1	2	3	4
River/canal	06	1	2	3	1	2	3	1	2	3	1	2	3	4
Agricultural water supply/irrigation system	07	1	2	3	1	2	3	1	2	3	1	2	3	4

	No. provided	Promenade provided	Promenade barrier-free
10	1	2	3

[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<sub>j</sub>*

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 farmer
Access to regional water	01	⊙	⊙	⊙
Access to regional land	03	⊙	⊙	⊙
Access to regional forest	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 (vegetables, fruit)	05	•	•	•
Forestry for education	06	•	•	•
Forest park	07	•	•	•
Others	09	•	•	•

		No. undertaken	Undertaken	Undertaken as an agricultural farmer
Access to regional water	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 ( $\beta_n$ )

The k<sup>th</sup> coefficient in  $\beta_n$  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

$\xi_{nr}$  = randomly generated Kx1 vector of standard normal errors.

R = R different replications (different draw of  $\xi_{nr}$  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

$\tau$  and  $v_j$  are BVN distributed ...

Assume that

$\tau$  and  $\eta_k$  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

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

	Number of users	Number of users (persons)	
		1	2
Specialized facilities	01		
General facilities	02		
Agricultural parks	03		
Facilities for viewing and visiting (forest parks, etc.)	04		
Others	05		
Promenades in forest parks	06		
	07		
	08		
Others	09		

	No. of users (persons)		
	1	2	3
No. of promenades provided			
10			

# An operational version of the 2-stage mixed logit.

Assuming a diagonal  $\Omega$  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...

$\sigma_k, \rho_k$  : estimates of the sd of  $\eta_k$ , and the correlation between  $\eta_k$  and  $\tau$

As described before ...  $\lambda$  and  $\delta$  are observation specific values derived from a first - stage PROBIT :

$$\alpha_z = \gamma w$$

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

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



**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)



[2] State of regional resources

Enter the resource area 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. Enter number of locations for terraced-paddy fields, valley bottom paddy fields and small reservoirs.

	Number of locations	Area (ha)												
		1	2	3	4	5	6	7	8	9	10			
Total land area	01													
Paddy field	02													
	03													
04														
05														
06														
07														
08														

# 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

[3] Conservation of regional resources

If the following regional resources are conserved circle "grounds for the conservation" and "conserving organization", whichever applies, and circle "conservation" for the purposes of conservation. And in a column circle "not conserved".

	Number of locations	Area (ha)												
		1	2	3	4	5	6	7	8	9	10			
Agriculture land	01													
Terraced paddy field	02													
	03													
Forest	04													
Pond for irrigation/lake	05													
River/canal	06													
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. of locations	Number of users (persons)		
		1	2	3
01				
02				
03				

2. Facilities using regional resources

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

	Number of locations	Number of users (persons)		
		1	2	3
01				
02				
03				
04				
05				
06				
07				
08				
09				

	Number of users (persons)		
	No. provided	Provided	Passive
10			