Multiple Choice Models: why not the same answer? A comparison among LIMDEP, R, SAS and STATA

Giuseppe Bruno
Bank of Italy

The R User Conference 2011, Warwick, Coventry, U.K.
August 16-18

The views expressed are those of the author only and do not involve the responsibility of the Bank of Italy

OUTLINE

1. Motivation and focus of the paper
2. Four packages at comparison
3. Discrete choice models
4. Logit Models and IIA
5. Multinomial Probit Models
6. Conditional Probit Models
7. Some numerical Examples
8. Concluding remarks

1 . Motivation and focus of the paper

Increasing interest in discrete choice models in the econometrics and behavioural literature.

Statistical and Social research institutes collect and store many microeconomic datasets.

Many statistical and econometric packages provide different algorithms for estimating discrete choice models.

i. Are accurate the estimation routines canned in these packages?
ii. Do we have sound benchmarks to take into account?

2. Four packages at comparison

For answering the previous questions we compare and confront the estimates of some of the most spread packages for discrete choice modeling:

1) LIMDEP (4.0.1)
2) **R (2.13.0)**
3) SAS (9.2)
4) STATA (11.2)

2. Four packages at comparison

Many other statistical packages are available.

Some of them were not available for the scrutiny.

Others do not feature powerful discrete choice estimation procedures.

3. Discrete Choice Models

1) Binary vs multinomial choice models,
2) Nominal vs ordered choices
3) Models with i.i.d. errors vs Models without i.i.d. errors

The simplest models are those where we have a binary decision:

$$y^* = \beta' \mathbf{x} + \epsilon$$

$$y = \begin{cases} 0 & \text{if } y^* \leq 0 \\ 1 & \text{if } y^* > 0 \end{cases}$$

$$Prob(Y = 1 | \mathbf{x}) = \int_{-\infty}^{\beta \mathbf{x}} \phi(t) dt = \Phi(\beta \mathbf{x})$$

$$Prob(Y = 1 | \mathbf{x}) = \frac{exp(\beta \mathbf{x})}{1 + exp(\beta \mathbf{x})}$$

3. Discrete Choice Models

Ordered choices models

$$y_i^* = \beta' \mathbf{x} + \epsilon$$

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq \mu_0 \\ 1 & \text{if } \mu_0 < y_i^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y_i^* \leq \mu_2 \\ \ldots & \\ J & \text{if } y_i^* > \mu_{J-1} \end{cases}$$

$\varepsilon \sim$ normal $\longrightarrow$ ordered probit

$\varepsilon \sim$ logistic $\longrightarrow$ ordered logit

3. Discrete Choice Models

Unordered choice models are motivated by random utility model:

$$U_{ij} = \theta' x_{ij} + \varepsilon_{ij}$$

Index i refers to the decision maker, index j refers to the choice. ε_{ij} is the unobservable residual.

Probability of making choice j is:

$$\text{Prob}(U_{ij} > U_{ik}) \quad \text{for all } k \neq j$$

3. Discrete Choice Models

With IID residuals distributed according to extreme value (Gumbel) we have a closed form expression of the choice probabilities.

$$Pr(y_i = j|\mathbf{Z}_i) = \frac{exp(\boxed{\beta' \cdot \mathbf{z}_{ij}})}{\Sigma_{k=1}^{J} exp(\beta' \cdot \mathbf{z}_{ik})}$$

Variables are choice varying:

<u>Conditional</u> Logit Model

$$Pr(y_i = j|\mathbf{Z}_i) = \frac{exp(\boxed{\beta'_j \cdot \mathbf{z}_i})}{\Sigma_{k=1}^{J} exp(\beta'_j \cdot \mathbf{z}_i)}$$

Variables are choice invariant:

<u>Multinomial</u> Logit Model

3. Discrete Choice Models

With a normal distribution for the residuals we don't have a closed form expression of the choice probabilities.

$$Pr(y_i = j|\mathbf{Z}_i, \Sigma_\varepsilon) = Prob\left[U_{i,j} \geq U_{i,k} \quad \forall k \neq i\right] =$$

$$= \int_{\varepsilon_j=-\infty}^{\infty} \int_{\varepsilon_1=-\infty}^{\varepsilon_j} \cdots \int_{\varepsilon_{j-1}=-\infty}^{\varepsilon_j} \int_{\varepsilon_{j+1}=-\infty}^{\varepsilon_j} \cdots \int_{\varepsilon_J=-\infty}^{\varepsilon_j} \phi\left(\epsilon|\mathbf{Z}_i, \Sigma_\varepsilon\right) d\varepsilon_1 d\varepsilon_2 \cdots d\varepsilon_J$$

This J-multivariate integral can be reduced to a (J-1)-dimensional.
Still a daunting task!

4. Multinomial Logit Models and IIA

Going from binary to multinomial choices brings in the issue of Independence of Irrelevant Alternatives.

Introduction of new choices correlated with the already available choices modifies their log-odds.

Red Bus / Blue Bus example

Modifications of the basic Logit have been developed for taking care of correlations among choices.

Another possibility is provided by …→

5. Multinomial Probit Models

No constraint is put on the covariance structure of the unobserved components of the utility.

With more then 5/6 alternatives the computational complexity gets quite large.

Simulated maximum likelihood or MCMC for Bayesian Analysis are possible avenues.

Some packages do not provide estimation algorithms for Multinomial Probit Models.

5. Multinomial Probit Models

R includes the MNP package which fits the Bayesian Multinomial Probit with Gibbs Sampling.

Stata provides the mprobit commands which imposes **independent standard normal** distribution for the residuals of the utility. No covariances are estimated.

MNP seems the more comprehensive procedure.

5. Multinomial Probit Models

Beliefs	R (10,000 draws)	R (20,000 draws)	R (40,000 draws)	Stata
Somewhat strong				
Intercept	-.22183	-.08894	-.01379	-.44738
Education	-.01124	-.01127	-.00364	-.00152
Income	.00442	.00435	.00345	.01306
Age	.00077	.00060	.00160	.00853
male	-.14141	-.11776	-.09044	-.47198
www	-.13375	-.10476	-.05743	-.26512
Not very strong				
Intercept	.15777	.08092	.03179	.98062
Education	-.01096	-.00683	-.00291	-.02546
Income	.00553	.00670	.00727	.00295
Age	-.00534	-.00543	-.00535	.00209
male	.00704	.01583	.02736	-.27948
www	.04094	.02958	.02569	.01115
Strong				
Intercept	-.13363	-.08183	.02215	-.02200
Education	-.00596	-.00696	.00083	-.00275
Income	-.00363	-.00351	-.00022	-.00083
Age	.01073	.00982	.00690	.02101
male	-.27653	-.23366	-.15502	-.64164
www	.03384	.02082	.01448	.03217

5. Multinomial Probit Models

The numerical results are not very satisfactory.

In a binomial framework the STATA command mprobit computes the same estimate as the probit command.

This test is not allowed in R: MNP refuses to run the estimate with only two categories.

6. Conditional Probit Models

SAS procedure MDC provides a PROBIT estimation with alternative-varying variables,

LIMDEP command MNPROBIT allows PROBIT estimation with alternative-varying & invariant variables.

7. Some Numerical Examples

Conditional Logit Estimates comparisons

Variables	LIMDEP/NLOGIT	R	SAS 9.2	Stata 11
Ground cost	-.0155	-.0155	-.0155	-.0155
Term time	-.0961	-.0961	-.0961	-.0961
Income	.0133	.0133	.0133	.0133
Air const	5.207	-5.207	5.207	5.207
Train const	3.869	-1.338	3.869	3.869
Bus const	3.163	-2.044	3.163	3.163
Car (reference)	-	-	-	-

7. Some Numerical Examples

Conditional Probit Estimates comparisons

Variables	LIMDEP/NLOGIT	R	SAS 9.2	Stata 11
Ground cost	-.0351	-.0116	-.0353	-.0122
Term time	-.0783	-.0345	-.0811	-.0281
Income	.0566	.0148	.0551	.0189
Air const	1.579	1.149	1.792	.8753
Train const	4.304	1.583	4.346	.6329
Bus const	3.634	1.308	3.646	-.6259
Car (reference)	-	-	-	-

7. Some Numerical Examples

Nested Logit Estimates comparisons

Variables	LIMDEP/NLOGIT	R	SAS 9.2	Stata 11
Ground cost	-.0316	-.0316	-.0316	-.03157
Term time	-.1126	-.1126	-.1126	-.11261
Air const	6.0423	6.0423	6.0423	6.0418
Train const	5.0646	5.0646	5.0646	5.0640
Bus const	4.0963	4.0963	4.0963	4.0958
Air included	.0153	-	.0153	.0153
Air tau	.5860	.5860	.5860	.5860
Ground tau	.3889	.3889	.3890	.3890

7. Some Numerical Examples

Mixed Logit Estimates comparisons

Variables	LIMDEP/NLOGIT	R	SAS 9.2	Stata 11
Ground cost	-.0308	-.0310	-.0310	-.0308
Term time	-.1142	-.1141	-.1141	-.1144
Air const	6.1503	6.1436	6.1491	6.1585
Train const	5.0990	5.1057	5.1021	5.1067
Bus const	4.1387	4.1421	4.1401	4.1462
Standard Deviations				
Air sd	2.9351	2.9768	-2.9186	2.9331
Train sd	.01472	.0007	-.0690	.01309
Bus sd	.00638	.0037	.00485	.00066

8. Concluding Remarks

It is relevant to compare canned estimated procedure.

The four examined packages produce quite comparable results in the estimation of multinomial/conditional logit models with different correlation structure among the errors.

Situation changes dramatically once we move to the Multinomial Probit Models: some packages do not provide estimation algorithms for them, others are not so easy to compare.

Development methods for open source statistical software might be improved by a tighter review of the numerical results.

Thank you for your attention.

Giuseppe Bruno
Bank of Italy
Research and International Relations
Head of I.T. Support Unit
giuseppe.bruno@bancaditalia.it