2B: Multinomial outcomes: Extras

© A. Colin Cameron
U. of Calif. - Davis

.

OeNB Summer School 2010
Microeconometrics
Oesterreichische Nationalbank (OeNB), Vienna, Austria

Based on
A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics: Methods and Applications (MMA), ch.14
Microeconometrics using Stata (MUS), ch.14.
Data examples are from MUS.

Aug 30 - Sep 3, 2010

1. Introduction

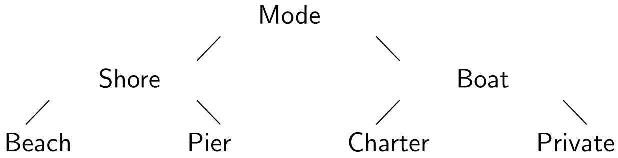
- For unordered data consider models that are richer than multinomial or conditional logit
  - Some do not have a closed form expression for the p_{ij}, so use
    - Maximum simulated likelihood estimation
    - Bayesian methods
- Consider models for more complicated forms of multinomial data: sequential, multivariate.

Outline

1. Introduction
2. Multinomial data: Nested logit model
3. Multinomial data: Random parameters multinomial logit (mixed logit)
4. Maximum simulated likelihood estimation
5. Multinomial data: Multinomial probit model
6. Bayesian methods
7. Multinomial data: Aggregate data
8. Multinomial data: Further Models: sequential, multivariate

2. Nested Logit Model

- Create tree structure for alternatives.
  - Within each branch errors are correlated.
  - Across branches errors are not.
- Fishing mode choice.
  - Assume fundamental distinction is between shore and boat fishing.



- Shore/boat contrast is called level 1 (or a limb).
- Next level is called level 2 (or a branch).
- Here
  - $(\varepsilon_{i,beach}, \varepsilon_{i,pier})$ are a bivariate correlated pair
  - $(\varepsilon_{i,private}, \varepsilon_{i,charter})$ are a bivariate correlated pair
  - the two pairs are independent.
- MNL/CL is special case all errors independent type I extreme value.
- Limitation is that need to specify the nest - not data determined.
- Two different nested logit models exist in the literature.
  - Only one of these (in recent Stata) is consistent with utility maximization.
  - And should have "dissimilarity parameter" in (0,1) interval.

- Nested logit: first define the tree

```
. * Define the tree for nested logit
. nlogitgen type = fishmode(shore: pier | beach, boat: private | charter)
new variable type is generated with 2 groups
label list lb_type
lb_type:
           1 shore
           2 boat

.
. * Check the tree
. nlogittree fishmode type, choice(d)

tree structure specified for the nested logit model

 type    N         fishmode   N      k
─────────────────────────────────────────
 shore 2364 ─┬─ beach     1182   134
             └─ pier      1182   178
 boat  2364 ─┬─ charter   1182   452
             └─ private   1182   418
─────────────────────────────────────────
                     total 4728 1182

k = number of times alternative is chosen
N = number of observations at each level
```

- Nested logit then estimated using following command:
`nlogit d p q || type:, base(shore) || fishmode: income, case(id) nolog`

```
RUM-consistent nested logit regression          Number of obs      =      4728
Case variable: id                               Number of cases    =      1182

Alternative variable: fishmode                  Alts per case: min =         4
                                                               avg =       4.0
                                                               max =         4

                                                   Wald chi2(5)    =    212.37
Log likelihood = -1192.4236                       Prob > chi2     =    0.0000

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
fishmode     |
           p |  -.0267625   .0018937   -14.13   0.000    -.0304741    -.023051
           q |   1.340091   .3080531     4.35   0.000     .7363177    1.943864
------------------------------------------------------------------------------
fishmode equations
------------------------------------------------------------------------------
beach        |
      income |     (base)
       _cons |     (base)
-------------+----------------------------------------------------------------
charter      |
      income |  -8.402017   78.35482    -0.11   0.915    -161.9746    145.1706
       _cons |   69.96998   558.8972     0.13   0.900    -1025.448    1165.388
-------------+----------------------------------------------------------------
pier         |
      income |  -9.458089   80.30189    -0.12   0.906    -166.8469    147.9307
       _cons |   58.94369   500.7358     0.12   0.906    -922.4805    1040.368
-------------+----------------------------------------------------------------
private      |
      income |  -1.634925   8.588643    -0.19   0.849    -18.46836    15.19851
       _cons |   37.52565   230.9065     0.16   0.871    -415.0428     490.094
------------------------------------------------------------------------------
dissimilarity parameters
------------------------------------------------------------------------------
type         |
  /shore_tau |    83.4692   718.5336                     -1324.831    1491.769
   /boat_tau |   52.55949   542.8918                     -1011.489    1116.608
------------------------------------------------------------------------------
LR test for IIA (tau = 1):           chi2(2) =    45.43   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

3. Random Parameters Logit Model

- The random parameters logit model introduces correlation across alternatives through an individual-specific random effect.
- Specifically, for an $m-$choice model we have

$$\begin{aligned} U_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim \text{i.i.d. type I extreme value} \\ \boldsymbol{\beta}_i &\sim \mathcal{N}[\boldsymbol{\beta}, \Sigma] \end{aligned}$$

  - $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i$ induces correlation across alternatives as then $U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + (\mathbf{x}_{ij}'\mathbf{u}_i + \varepsilon_{ij})$ where $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \Sigma]$.
- Conditional on $\boldsymbol{\beta}_i$ the model is easily estimated CL.
  - But additionally need to integrate out $\boldsymbol{\beta}_i$.
  - Use maximum simulated likelihood or Bayesian methods.

- Stata user-written command `mixlogit` has same format as command `clogit`.
  - Here apply for three-choice example (with `charter` dropped).
  - Specify just regressor `p` to have random coefficient.

```
. mixlogit d q d3 d4 d3income d4income, group(id) rand(p)

Iteration 0:   log likelihood = -602.33584  (not concave)
Iteration 1:   log likelihood = -447.46013
Iteration 2:   log likelihood = -435.29806
Iteration 3:   log likelihood = -434.56105
Iteration 4:   log likelihood = -434.52856
Iteration 5:   log likelihood = -434.52844
Iteration 6:   log likelihood = -434.52844

Mixed logit model                                 Number of obs   =       2190
                                                  LR chi2(1)      =      64.57
Log likelihood = -434.52844                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Mean         |
           q |   .7840088   .9147869     0.86   0.391    -1.008941    2.576958
          d3 |   .7742955    .224233     3.45   0.001     .3348069    1.213784
          d4 |   .5617395   .3158082     1.78   0.075    -.0572331    1.180712
    d3income |  -.1199613   .0492249    -2.44   0.015    -.2164404   -.0234822
    d4income |   .0518098   .0721527     0.72   0.473    -.0896068    .1932265
           p |  -.1069866   .0274475    -3.90   0.000    -.1607827   -.0531904
-------------+----------------------------------------------------------------
SD           |
           p |   .0598364   .0191597     3.12   0.002      .022284    .0973888
------------------------------------------------------------------------------
```

4. Maximum Simulated Likelihood Estimation

- Problem: The MLE (with independent data over i) maximizes

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

  - but $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ does not have a closed form solution.
  - e.g. $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \int g(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \alpha)h(\alpha)d\alpha = ?$

- Solution: Maximum simulated likelihood estimator (MSL) maximizes

$$\ln \widehat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

  - where $\widehat{f}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ is a simulated approximation to $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$
  - e.g. $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} g(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \alpha^{(s)})$ where $\alpha^{(s)}$ are draws from the density $h(\alpha)$

- The MSL estimator is consistent and has the usual asymptotic distribution as the MLE if
  - $\widehat{f}(\cdot)$ is an unbiased simulator and satisfies other conditions given below
  - $S \rightarrow \infty$, $N \rightarrow \infty$ and $\sqrt{N}/S \rightarrow 0$ where S is number of simulations.
  - Note that many draws S (to compute $\widehat{f}(\cdot)$) are required.

- Assumed properties of the simulator:
  - $\widehat{f}(\cdot)$ is an unbiased simulator with
$$\mathrm{E}[\widehat{f}(y_i|\mathbf{x}_i,\boldsymbol{\theta})] = f(y_i|\mathbf{x}_i,\boldsymbol{\theta})$$
  - $\widehat{f}(\cdot)$ is differentiable in $\boldsymbol{\theta}$ (or smooth simulator) so gradient methods can be used
  - the underlying draws to compute $\widehat{f}(\cdot)$ are unchanged so no "chatter".
- We need many draws S because simulator is biased for $\ln f(\cdot)$
$$\mathrm{E}[\widehat{f}(\cdot)] = \mathrm{E}[f(\cdot)] \quad \not\Rightarrow \quad \mathrm{E}[\ln \widehat{f}(\cdot)] \neq \mathrm{E}[\ln f(\cdot)].$$
- Binary probit example
  - Density $f_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}$
  - Frequency simulator
$\widehat{f}_i = \frac{1}{S}\sum_{s=1}^{S} 1[\varepsilon_i^{(s)} \leq \mathbf{x}_i'\boldsymbol{\beta}]^{y_i}(1-1[\varepsilon_i^{(s)} \leq \mathbf{x}_i'\boldsymbol{\beta}])^{1-y_i}$
    - $\varepsilon_i^{(s)}$, $s = 1, ..., S$, are random draws from $\mathcal{N}[0,1]$
    - But here not smooth so need to use a different simulator.

MSL Application to Random Parameters Logit

- Recall $U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij}$; $\varepsilon_{ij} \sim$ type I extreme value; $\boldsymbol{\beta}_i \sim \mathcal{N}[\boldsymbol{\beta}, \Sigma]$.
- If $\boldsymbol{\beta}_i$ known then have CL model with $p_{ij} = e^{\mathbf{x}_{ij}'\boldsymbol{\beta}_i} / \sum_{l=1}^{m} e^{\mathbf{x}_{il}'\boldsymbol{\beta}_l}$.
- Instead $\boldsymbol{\beta}_i$ random and needs to be integrated out

$$p_{ij} = \Pr[y_i = j] = \int \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}_i}}{\sum_{l=1}^{m} e^{\mathbf{x}_{il}'\boldsymbol{\beta}_l}} \phi(\boldsymbol{\beta}_i | \boldsymbol{\beta}, \Sigma).$$

- The MSL estimator of $\boldsymbol{\beta}$ and Σ maximizes

$$\begin{aligned} \ln \widehat{L}(\boldsymbol{\beta}, \Sigma) &= \sum_{i=1}^{N} \ln \widehat{f}(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \Sigma) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{m} \ln \left[\frac{1}{S} \sum_{s=1}^{S} \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}_i^{(s)}}}{\sum_{l=1}^{m} e^{\mathbf{x}_{il}'\boldsymbol{\beta}_l^{(s)}}} \right] \end{aligned}$$

  - where $\boldsymbol{\beta}_i^{(s)}$, $s = 1, ..., S$, are random draws from $\phi(\boldsymbol{\beta}_i | \boldsymbol{\beta}, \Sigma)$
  - and at r^{th} round of gradient method draw is from $\phi(\boldsymbol{\beta}_i | \boldsymbol{\beta}^r, \Sigma^r)$.

Method of Simulated Moments

- An alternative less efficient estimator is the method of simulated (MSM) estimator.
- Suppose $\widehat{\boldsymbol{\theta}}$ is a method of moments estimator (MM) that solves

$$\sum_{i=1}^{N} \mathbf{m}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

- Suppose there is unbiased simulator such that $\mathrm{E}[\widehat{\mathbf{m}}(y_i|\mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{m}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$.
- Then the method of simulated (MSM) solves

$$\sum_{i=1}^{N} \widehat{\mathbf{m}}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}$$

  is consistent even if S is small though there is an efficiency loss.
  - When $\widehat{\mathbf{m}}(\cdot)$ is the frequency simulator $\mathrm{V}[\widehat{\boldsymbol{\theta}}_{\mathrm{MSM}}] = (1 + \frac{1}{S})\mathrm{V}[\widehat{\boldsymbol{\theta}}_{\mathrm{MM}}]$.
- In practice the MSL is used much more often even though larger S.

5. Multinomial Probit Model

- Consider three-choice example of the multinomial probit model.
  - ARUM with errors multivariate normal distributed.

$$\begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \right).$$

  - Not all the variance components are identified.
  - Only covariance matrix of differenced errors $\varepsilon_j - \varepsilon_1$, plus one normalization.
  - Here e.g. $\sigma_2^2 = 1$, and σ_{32} and σ_3^2 free.
- Even if error model is technically identified, parameters of the MNP model may be imprecisely estimated (like multicollinearity).
  - Further restrictions are needed in practice.

- Use Stata command asmlogit
  - Uses simulated maximum likelihood
  - With GHK simulator which is a smooth simulator (meaning small change in β changes simulated value of p_{ij} so that objective function is differentiable in β)

```
. * Multinomial probit with case-specific regressors
. drop if fishmode=="charter" | mode == 4
(2538 observations deleted)

. asmprobit d p q, case(id) alternatives(fishmode) casevars(income) ///
>   correlation(unstructured) structural vce(robust) nolog
note: variable p has 106 cases that are not alternative-specific: there is no
      within-case variability

Alternative-specific multinomial probit        Number of obs      =       2190
Case variable: id                              Number of cases    =        730

Alternative variable: fishmode                 Alts per case: min =          3
                                                              avg =        3.0
                                                              max =          3

Integration sequence:         Hammersley
Integration points:                  150           Wald chi2( 4)    =      12.97
Log simulated-pseudolikelihood = -482.30128        Prob > chi2      =     0.0114
```

(Std. Err. adjusted for clustering on id)

d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
fishmode						
p	-.0233627	.0114346	-2.04	0.041	-.0457741	-.0009513
q	1.399925	.5395423	2.59	0.009	.3424418	2.457409
beach	(base alternative)					
pier						
income	-.097985	.0413117	-2.37	0.018	-.1789543	-.0170156
_cons	.7549123	.2013551	3.75	0.000	.3602636	1.149561
private						
income	.0413866	.0739083	0.56	0.575	-.103471	.1862443
_cons	.6602584	.2766473	2.39	0.017	.1180397	1.202477
/lnsigma3	.4051391	.5009809	0.81	0.419	-.5767654	1.387044
/atanhr3_2	.1757361	.2337267	0.75	0.452	-.2823598	.6338319
sigma1	1	(base alternative)				
sigma2	1	(scale alternative)				
sigma3	1.499511	.7512264			.5617123	4.002998
rho3_2	.173949	.2266545			-.2750878	.5606852

(fishmode=beach is the alternative normalizing location)
(fishmode=pier is the alternative normalizing scale)

```
. * Show correlations and covariance
. estat correlation
```

	beach	pier	private
beach	1.0000		
pier	0.0000	1.0000	
private	0.0000	0.1739	1.0000

```
. estat covariance
```

	beach	pier	private
beach	1		
pier	0	1	
private	0	.2608385	2.248533

6. Bayesian Methods

- Bayesian methods begin with
  - Likelihood: $L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$
  - Prior on $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta})$
- This yields the posterior distribution for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{X})}$$

  - where $f(\mathbf{y}|\mathbf{X}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \times \pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is called the marginal likelihood.
  - This uses the result that $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$.
- Bayesian analysis then bases inference on the posterior distribution.
  - e.g. Best point estimate of $\boldsymbol{\theta}$ may be the mean of the posterior distribution.
  - e.g. A 95% confidence interval for $\boldsymbol{\theta}$ is from the 2.5 to 97.5 percentiles of the posterior distribution.

- Bayesian inference is a different inference method
  - treats $\boldsymbol{\theta}$ as intrinsically random
  - whereas classical inference treats $\boldsymbol{\theta}$ as fixed and $\widehat{\boldsymbol{\theta}}$ as random.
- Modern Bayesian methods (Markov chain Monte Carlo)
  - make it much easier to compute the posterior distribution than to maximize the log-likelihood.
- So classical statisticians:
  - use Bayesian methods to compute the posterior
  - use an uninformative prior so $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \simeq L(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$
  - so $\boldsymbol{\theta}$ that maximizes the posterior is also the MLE.
- Or can go all the way and be Bayesian.

Markov chain Monte Carlo (MCMC)

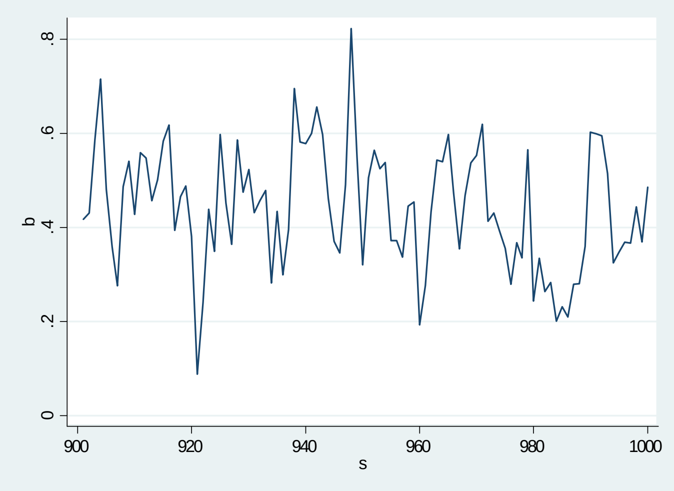
- The challenge is to compute the posterior
  - analytical results are only available in special cases.
  - e.g. If $\mathbf{y}|\mathbf{X}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and known variance and the prior for $\boldsymbol{\beta}$ is normal with specified mean and variance then the posterior for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}$ is also normal.
- Instead use Markov chain Monte Carlo methods:
  - Make sequential random draws $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$.
  - where $\boldsymbol{\theta}^{(s)}$ depends in part on $\boldsymbol{\theta}^{(s-1)}$
  - in such a way that after an initial burn-in (discard these draws)
  - $\boldsymbol{\theta}^{(s)}$ are (correlated) draws from the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$.
- MCMC methods include
  - Gibbs sampler
  - Metropolis and Metropolis-Hastings algorithms
  - Data augmentation

Probit example

- Likelihood: Probit model with single regressor
  - $\ln L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})=\sum_i y_i \ln \Phi(\beta_1+\beta_2 x)+(1-y_i)\ln(1-\Phi(\beta_1+\beta_2 x))$
- Prior: uniform prior (all values equally likely)
  - $\pi(\boldsymbol{\beta})=\pi(\beta_1,\beta_2)=1$
- Posterior: no closed form solution
  - though proper even though the prior was improper
  - instead use Gibbs sampler and data augmentation
- Example: the above with generated data
  - $\beta_1=0$, $\beta_2=1$, $N=100$, $x\sim\mathcal{N}[0,1]$
- Gibbs sampler yields 1,000 correlated draws from the posterior.

Correlated draws

- The last 100 draws from the posterior density of β_2



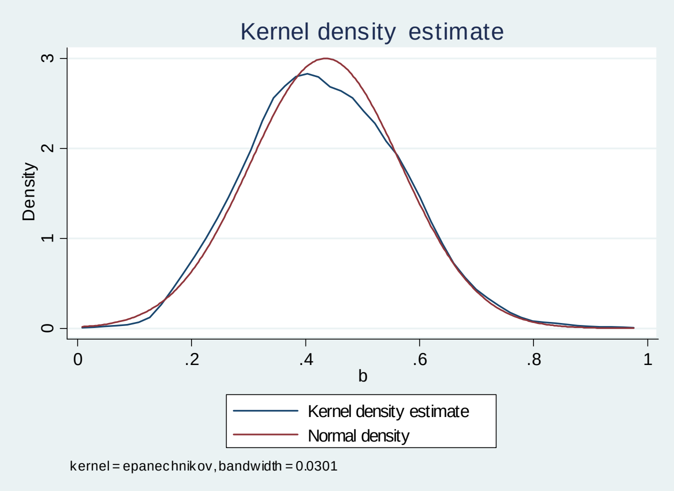
- Correlations of the 1,000 draws of β_2 die out quickly

```
. corrgram b, lags(10)

                                           -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.5127   0.5128   263.7   0.0000
2        0.2581  -0.0068   330.6   0.0000
3        0.1061  -0.0330   341.92  0.0000
4        0.0299  -0.0153   342.82  0.0000
5        0.0137   0.0159   343.01  0.0000
6       -0.0440  -0.0685   344.95  0.0000
7       -0.0330   0.0198   346.05  0.0000
8       -0.0126   0.0144   346.21  0.0000
9       -0.0086  -0.0070   346.29  0.0000
10      -0.0255  -0.0315   346.95  0.0000
```

Posterior density

- Kernel density estimate of the 1,000 draws of β_2
  - centered around 0.4-0.5 with standard deviation of 0.1-0.2.



- More precisely
  - Posterior mean of β_2 is 0.434 and standard deviation is 0.132
  - A 95% percent Bayesian confidence interval for β_2 is $(0.195, 0.701)$.

```
. summarize b

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           b |      1000    .4345774    .1329711    .0379931     .94584

. centile b, centile(2.5, 97.5)

                                                      -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           b |    1000         2.5      .194546        .1848584    .2014523
             |                97.5      .701408        .6852426    .7263849
```

Gibbs Sampler

- Gibbs sampler is simple MCMC method
- used when
  - we can partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$
  - we do not know the posterior $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
  - but we do know the conditional posteriors $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$
- Then make alternating draws from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$
  - Start with $\boldsymbol{\theta}_1^{(1)}$
  - Draw $\boldsymbol{\theta}_2^{(1)}$ from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(1)})$
  - Draw $\boldsymbol{\theta}_1^{(2)}$ from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(1)})$
  - Draw $\boldsymbol{\theta}_2^{(2)}$ from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(2)})$ etc.
- Gibbs eventually gives (correlated) draws from $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ even though

$$\begin{aligned} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \times p(\boldsymbol{\theta}_2) \\ &\neq p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \times p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1). \end{aligned}$$

Data Augmentation

- Consider latent variable model where observed data y are determined completely by y^*.
  - We have data $y_i, \mathbf{x}_i$
  - where $y_i = g(y_i^*)$ with $g(\cdot)$ known
  - and y_i^* depends on $\mathbf{x}_i$ and $\boldsymbol{\theta}$
  - probit is an example.
- Furthermore suppose that Bayesian analysis would be easy if y_i^* was observed
  - so the posterior $p(\boldsymbol{\theta}|y_1^*, ..., y_N^*,\text{data})$ is known.
- Then data augmentation
  - treats the parameters as $\boldsymbol{\theta}$ and $y_1^*, ..., y_N^*$
  - then do Gibbs sampler
    - ★ draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|y_1^*, ..., y_N^*,\text{data})$
    - ★ and draw $y_1^*, ..., y_N^*$ from $p(y_1^*, ..., y_N^*|\boldsymbol{\theta},\text{data})$.

Probit example

- Likelihood: Probit model
  - $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}[0,1]$.
  - $y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$
- Prior: uniform prior (all values equally likely)
  - $\pi(\boldsymbol{\beta}) = 1$
- Known tractable result: for $\mathbf{y}^* \sim \mathcal{N}[\mathbf{X}\boldsymbol{\beta}, \mathbf{I}]$ and uniform prior on $\boldsymbol{\beta}$
  - $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{X})$ is $\mathcal{N}[\widehat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}]$ where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$.
- Data augmentation add $y_1^*, ..., y_N^*$ as parameters.
  - Then $p(\boldsymbol{\beta}|y_1^*, ..., y_N^*, \mathbf{y}, \mathbf{X})$ is $\mathcal{N}[\widehat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}]$
  - And $p(y_1^*, ..., y_N^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ is truncated normal
    - If $y_i = 1$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ left truncated at 0
    - If $y_i = 0$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ right truncated at 0
- So draw $\boldsymbol{\beta}^{(s)}$ from $p(\boldsymbol{\beta}|y_1^{*(s-1)}, ..., y_N^{*(s-1)}, \mathbf{y}, \mathbf{X})$ and draw $y_1^{*(s)}, ..., y_N^{*(s)}$ from $p(y_1^*, ..., y_N^*|\boldsymbol{\beta}^{(s)}, \mathbf{y}, \mathbf{X})$

Multinomial probit example

- Likelihood: Multinomial probit model
  - $U_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}[\mathbf{0}, \Sigma_\varepsilon]$
  - $y_{ij} = 1$ if $U_{ij}^* > U_{ik}^*$ all $k \neq j$
- Prior for $\boldsymbol{\beta}$ and Σ_ε may be normal-Wishart
- Data augmentation
  - Latent utilities $\mathbf{U}_i = (U_{i1}, ..., U_{im})$ are introduced as auxiliary variables
  - Let $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_N)$ and $\mathbf{y} = (y_1, ..., y_N)$
- Gibbs sampler cycles between
  - 1. Conditional posterior for $\boldsymbol{\beta}|\mathbf{U}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$
  - 2. Conditional posterior for $\Sigma_\varepsilon|\boldsymbol{\beta}, \mathbf{U}, \mathbf{y}, \mathbf{X}$, and
  - 3. Conditional posterior for $\mathbf{U}_i|\boldsymbol{\beta}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$.
- Albert and Chib (1993) provide a quite general treatment.
- McCulloch and Rossi (1994) provide a substantive MNP application.

7. Aggregate Data for individual random parameters logit

- Can do regular multinomial logit or NLSUR on aggregated data
  - Here consider harder problem of linking to individual behavior.
- The data available are for brand j in market t :
  - market share s_{jt}, average prices p_{jt}, other product characteristics $\mathbf{w}_{jt}$.
- The underlying model is one of individual behavior
  - utility of individual i for brand j in market t is

$$\begin{aligned} U_{ijt} &= \mathbf{w}_{jt}'\gamma_i - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt} \\ &= \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \xi_{jt} + \varepsilon_{ijt}, \end{aligned}$$

  - where ε_{ijt} is i.i.d. type I extreme value
- Consider the following situations
  - No individual heterogeneity: $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ (only heterogeneity is ε_{ijt})
  - No individual heterogeneity and endogenous $\mathbf{x}_{jt}$ (e.g. prices).
  - Individual heterogeneity: $\boldsymbol{\beta}_i$ is normally distributed.

No individual heterogeneity

- Given ε_{ijt} i.i.d. extreme value, then get usual conditional logit model

$$\Pr[y_{ijt} = 1] = \frac{\exp(\mathbf{x}'_{jt}\boldsymbol{\beta} + \xi_{jt})}{1 + \sum_{k=1}^{m} \exp(\mathbf{x}'_{kt}\boldsymbol{\beta} + \xi_{kt})}$$

- We have aggregate market data so estimate the share

$$s_{jt} = \frac{\exp(\mathbf{x}'_{jt}\boldsymbol{\beta} + \xi_{jt})}{1 + \sum_{k=1}^{m} \exp(\mathbf{x}'_{kt}\boldsymbol{\beta} + \xi_{kt})}.$$

- Introduce an outside good, good 0, normalized so that $\mathbf{x}'_{jt}\boldsymbol{\beta} = 0$.
  - Then $s_{0t} = 1/[1 + \sum_{k=1}^{m} \exp(\mathbf{x}'_{kt}\boldsymbol{\beta} + \xi_{kt})]$
  - So $s_{jt} = \exp(\mathbf{x}'_{jt}\boldsymbol{\beta} + \xi_{jt})/s_{0t}$ and

$$\ln s_{jt} - \ln s_{0t} = \mathbf{x}'_{jt}\boldsymbol{\beta} + \xi_{jt}.$$

  - So can estimate $\boldsymbol{\beta}$ by OLS using market share data.
- Empirical results will depend on the outside good
  - and need to get a share figure for the outside good.

Endogeneity but no individual heterogeneity

- Now suppose the unobserved heterogeneity ξ_{jt} is correlated with prices p_{jt} or other characteristics $\mathbf{x}_{jt}$.
- Then estimate by IV

$$\ln s_{jt} - \ln s_{0t} = \mathbf{x}_{jt}'\boldsymbol{\beta}_{jt} + \xi_{jt},$$

  - where instruments $\mathbf{z}_{jt}$ satisfy $\mathrm{E}[\mathbf{z}_{jt}\xi_{jt}] = 0$
  - e.g. instruments from supply-side if modelling demand.

Individual heterogeneity

- Suppose $U_{ijt} = \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \xi_{jt} + \varepsilon_{ijt}$ where $\boldsymbol{\beta}_i$ is normally distributed
  - then with ε_{ijt} i.i.d. extreme value, get RPL model at individual level.
- But we have only market share data
  - Let $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{u}_i$ and rewrite

$$\begin{aligned} U_{ijt} &= \mathbf{x}_{jt}'\boldsymbol{\beta}_i + \xi_{jt} + \varepsilon_{ijt} \\ &= \mathbf{x}_{jt}'\boldsymbol{\beta} + \xi_{jt} + \mathbf{x}_{jt}'\mathbf{u}_i + \varepsilon_{ijt} \end{aligned}$$

- Integrate out $\mathbf{u}_i$ and ε_{ijt} to leave model depending on $\mathbf{x}_{jt}$ and ξ_{jt}.
  - The set of individuals choosing brand j in market t is

$$A_{jt}(\mathbf{x}_{jt}, \xi_{jt}) = \{\mathbf{u}_i, \varepsilon_{i0t}, ..., \varepsilon_{imt} | U_{ijt} \geq U_{ilt} \text{ for all } l = 0, ..., m\}.$$

  - Integrate out individual heterogeneity to get the market share

$$s_{jt}(\mathbf{x}_{jt}, \xi_{jt} | \boldsymbol{\beta}, \Sigma_{\beta}) = \int_{A_{jt}} df(\mathbf{u}_i, \varepsilon_{i0t}, ..., \varepsilon_{imt})$$

  where $f(\mathbf{u}_i, \varepsilon_{i0t}, ..., \varepsilon_{imt})$ is the joint distribution of the errors
    - ★ iid type 1 extreme value for the ε_{ijt}
    - ★ $\mathcal{N}[\mathbf{0}, \Sigma_{\beta}]$ for $\mathbf{u}_i$

- Now predicted share $s_{jt}(\mathbf{x}_{jt}, \xi_{jt}|\boldsymbol{\beta}, \Sigma_{\beta})$ is very nonlinear
  - the error ξ_{jt} is nonadditive
  - so can't just do NLS of s_{jt} on $s_{jt}(\mathbf{x}_{jt}, \xi_{jt}|\boldsymbol{\beta}, \Sigma_{\beta})$
  - also may be concerend about endogeneity of $\mathbf{x}_{jt}$
- Berry (1984) instead proposed the following (see also Nevo (2000))
  - Solve for ξ_{jt} (viewed as a structural error) as a function of $s_{jt}, \mathbf{x}_{jt}, \boldsymbol{\beta}, \Sigma_{\beta}$.
  - Assume there are instruments $\mathbf{z}_{jt}$ (allows for e.g. endogenous prices)
  - Stack ξ_{jt} and $\mathbf{z}_{jt}$ into $\boldsymbol{\xi}$ and $\mathbf{Z}$ and estimate $\boldsymbol{\beta}$ and Σ_{β} by GMM estimator that minimizes

$$Q(\boldsymbol{\beta}, \Sigma_{\beta}) = [\mathbf{Z}'_{jt}\boldsymbol{\xi}(\boldsymbol{\beta}, \Sigma_{\beta})]'\mathbf{W}[\mathbf{Z}'_{jt}\boldsymbol{\xi}(\boldsymbol{\beta}, \Sigma_{\beta})]$$

- This is computationally challenging
  - Computation of $s_{jt}(\mathbf{x}_{jt}, \xi_{jt}|\boldsymbol{\beta}, \Sigma_{\beta})$ requires numerical methods
  - Inversion to get $\mathbf{x}'_{jt}\boldsymbol{\beta} + \xi_{jt}$ and hence ξ_{jt} requires numerical methods
  - Knittel and Metaxoglou (2008) find problems with many optima that lead to quite different estimated price elasticities.

8. Further Models: Sequential Models

- Example is sequential probit with three alternatives.
  - First choose whether $y = 1$ or $y \neq 1$.
  - Second, if $y \neq 1$ choose whether $y = 2$ or $y = 3$.
- Assume a probit model at each stage, with regressors $\mathbf{x}_2$ at the first stage and regressors $\mathbf{x}_1$ at the second stage.
  - Then

$$\begin{aligned} p_1 &= \Pr[y = 1] = \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1), \\ \frac{p_2}{p_2 + p_2} &= \Pr[y_i = 2 | y_i \neq 1] = \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2), \end{aligned}$$

  - This implies after some algebra

$$\begin{aligned} p_2 &= \Pr[y \neq 1] \times \Pr[y = 2 | y \neq 1] = (1 - \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)) \times \Phi(\mathbf{x}_2'\boldsymbol{\beta}_2) \\ p_3 &= 1 - p_1 - p_2. \end{aligned}$$

  - The likelihood function is then easily obtained and estimation is by ML.

Further Models: Multivariate Models

- Multivariate models have more than one discrete dependent variable.
  - Example: jointly model labor supply and fertility

$$
\begin{aligned}
y_1 &= \begin{cases} 0 & \text{if do not work} \\ 1 & \text{if work} \end{cases} \\
y_2 &= \begin{cases} 0 & \text{if no children} \\ 1 & \text{if children} \end{cases}
\end{aligned}
$$

  - There are four probabilities

$$
\begin{aligned}
p_{00} &= \Pr[y_1 = 0, y_2 = 0] \\
p_{01} &= \Pr[y_1 = 0, y_2 = 1] \\
p_{10} &= \Pr[y_1 = 1, y_2 = 0] \\
p_{11} &= \Pr[y_1 = 1, y_2 = 1].
\end{aligned}
$$

  - These are mutually exclusive and exhaust all possibilities, so that $p_{00} + p_{00} + p_{00} + p_{00} = 1$.

Further Models: Bivariate Probit

- From these probabilities one can form the log-likelihood, and estimate by ML.
  - This is essentially the same as a four-choice multinomial model.
  - All that differs is the story told to derive the functional forms for the probabilities.
- Bivariate probit model is a leading example.
  - Observe $y_1 = 1$ or 0 if $y_1^* >$ or < 0
    and $y_2 = 1$ or 0 if $y_2^* >$ or < 0 where

$$\begin{array}{rcl} y_1^* & = & \mathbf{x}_1'\beta_1 + \varepsilon_1 \\ y_2^* & = & \mathbf{x}_2'\beta_2 + \varepsilon_1 \\ \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \end{bmatrix} & \sim & \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \end{array}$$

Further Models: Bivariate Probit Data Example

- Bivariate probit example: y_1 is health excellent and y_2 is visit doctor.

```
. * Two binary dependent variables: hlthe and dmdvs
. tabulate hlthe dmdu

           |   any MD visit = 1 if
           |        mdu > 0
     hlthe |         0          1 |     Total
-----------+----------------------+----------
         0 |       826      1,731 |     2,557
         1 |     1,006      2,011 |     3,017
-----------+----------------------+----------
     Total |     1,832      3,742 |     5,574

. correlate hlthe dmdu
(obs=5574)

             |    hlthe     dmdu
-------------+------------------
       hlthe |   1.0000
        dmdu |  -0.0110   1.0000
```

- Estimate using Stata command biprobit

```
. * Bivariate probit estimates
. biprobit hlthe dmdu age linc ndisease, nolog

Bivariate probit regression                       Number of obs   =      5574
                                                  Wald chi2(6)    =    770.00
Log likelihood = -6958.0751                       Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
hlthe        |
         age |  -.0178246   .0010827   -16.46   0.000    -.0199466   -.0157025
        linc |    .132468   .0149632     8.85   0.000     .1031406    .1617953
    ndisease |  -.0326656   .0027589   -11.84   0.000    -.0380729   -.0272583
       _cons |  -.2297079   .1334526    -1.72   0.085    -.4912703    .0318545
-------------+----------------------------------------------------------------
dmdu         |
         age |   .0020038   .0010927     1.83   0.067    -.0001379    .0041455
        linc |   .1212519   .0142512     8.51   0.000       .09332    .1491838
    ndisease |   .0347111   .0028908    12.01   0.000     .0290452    .0403771
       _cons |  -1.032527   .1290517    -8.00   0.000    -1.285464   -.7795907
-------------+----------------------------------------------------------------
     /athrho |   .0282258    .022827     1.24   0.216    -.0165142    .0729658
-------------+----------------------------------------------------------------
         rho |   .0282183   .0228088                     -.0165127    .0728366
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =   1.5295    Prob > chi2 = 0.2162
```

Further Models

- Ranked Data
  - With stated preference data we know the second-preferred choice, not just the most-preferred choice.
  - Using this can increase efficiency of estimation
  - e.g. For MNL first preference is MNL with m alternatives, and second preference is MNL with $(m-1)$ alternatives.
- Simultaneous Equations
  - Two binary variables that are simultaneous.
  - Easiest if simultaneity is in latent variables (y_1^*, y_2^*).
    Then work with reduced form in (y_1^*, y_2^*).
  - More difficult if simultaneous in the binary outcomes (y_1, y_2).

9. Some References

- These references are mainly ones that refer to the recent literature.
- For random parameters logit see
  - Hole, A.R. (2007), "Fitting Mixed Logit Models by using Simulated Maximum Likelihood," *Stata Journal*, 7, 388-401.
- For multinomial probit see
  - Liesenfeld, R., and J.-F. Richard (2010), "The dynamic invariant multinomial probit model: Identification, pretesting and estimation," *Journal of Econometrics*, 155, 117-127.
- For maximum simulated likelihood and Bayesian for multinomial data see
  - Train, K. (2004), *Discrete choice methods with simulation*, Cambridge University Press.

- For recent Bayesian multinomial applications see

  - Munkin, M.K., and P.K. Trivedi (2008), "Bayesian analysis of the ordered probit model with endogenous selection," *Journal of Econometrics*, 144, 334-348.
  - Imai, S., N. Jain, and A. Ching (2009), "Bayesian Estimation of Dynamic Discrete Choice Models," *Econometrica*, 1865-1900.

- For individual choice with aggregate market share data see

  - Berry, S.T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242-262.
  - Berry, Steven, Levinsohn, James, Pakes, Ariel, 1995. Automobile prices in market equilibrium. *Econometrica* 63 (4), 841 890.
  - Knittel, C.R., and K. Metaxoglou (2008), "Estimation of Random Coefficient Demand Models: Challenges, Difficulties and Warnings," Manuscript, University of California - Davis.
  - Jiang, R., P. Manchandab and P.E. Rossi (2009), "Bayesian analysis of random coefficient logit models using aggregate data," *Journal of Econometrics*, 149, 136-148.
