Nick Mastronardi
Econometrics III Miravete

Explaining the choice for Measured service as a function of Socio-Economic Variables

- Logit Model
  Household problem: $min_{x\in\{F,M\}}C(x) = \mathbb{I}_F K + \mathbb{I}_M P \cdot y$
  Where $y = f(demographics; \beta) + \varepsilon$
  f specified to be linear. Nonlinearities in relationship accounted for by dummies on month, dummies (piecewise linear bands) on head of hh age, log (income variable).
  Regressor set specified to {doct, dnov (dummies for month, relative to September), dage25, dage35, dage45, dage55, dage65 (dummies for bands on age relative to teens), hhsize, lincval (log of income), moved (dummy if moved), sex}
  Proportion of population choosing measured service should correspond to mass of epsilons st $P \cdot y < K$ should equal (if errors assumed distributed logistically) $\frac{e^k}{e^k + e^{f(x;\beta)}}$
  Therefore, Stata: $max_\beta \log(Likelihood\ of\ observing\ our\ data)$
  Results –

measured	Coef.	Std. Err.
doct	2.457472	.0617489
dnov	2.444679	.0616948
dage25	.2493785	.0959594
dage35	.3511839	.1032649
dage45	.3156313	.1114191
dage55	.2209418	.1124372
dage65	.369765	.1080982
hhsize	-.1690799	.0188653
lincval	.2785012	.029134
moved	.0660751	.021419
sex	-.0647231	.0505181
_cons	-4.230806	.2923375

almost all estimates are significant (not moved)
Probability a HH selects measured is up in October from Sept, but then steady between October and November
Intuitively, the larger the hh size, the less likely the household chooses measured
All other estimates positive, but small

Marginal effects after logit	
y = Pr(measured) (predict)	
0.47958265	
variable dy/dx	Std. Err.
doct* .5369936	0.01055
dnov* .5348868	0.01058
dage25* .0622548	0.02391
dage35* .0875655	0.0256
dage45* .0787221	0.02762
dage55* .0551793	0.02802
dage65* .0921405	0.02672
hhsize -.0421995	0.00471
lincval .0695092	0.00727
moved .0164912	0.00535
sex* -.0161582	0.01261
(*) dy/dx is for discrete change	

- Probit Model

  Similar, now ε assumed to be distributed normally

measured	Coef.	Std. Err.
doct	1.482462	0.035285
dnov	1.474746	0.035265
dage25	0.143187	0.056842
dage35	0.199051	0.061126
dage45	0.177348	0.066339
dage55	0.133436	0.066596
dage65	0.217952	0.063954
hhsize	-0.10096	0.011061
lincval	0.163828	0.017363
moved	0.03812	0.01269
sex	-0.03943	0.029866
_cons	-2.49587	0.17307

Marginal effects after probit
y = Pr(measured) (predict)
0.48427907

ble	dy/dx	S
doct*	.5318427	0.0103
dnov*	.5296406	0.0103
dage25*	.0570732	0.0226
dage35*	.0792688	0.0242
dage45*	.070636	0.0263
dage55*	.0531907	0.0265
dage65*	.0867317	0.0253
hhsize	-.0402446	0.0044
lincval	.0653071	0.0069
moved	.015196	0.0051
sex*	-.0157187	0.0119

- Choice Based Sampling

  Since the sample is not representative of the population, our estimates need correction
  In particular, after I 'cleaned' the data (removed bad observations, observations with missing or problematic variable values over variables I use in my specification of the model), the sample was found to be roughly 49% measured observations, and 51% flat observations.
  In contrast however, the population was reported to be 10% measured and 90% flat.
  Thus, in our sample, measured observations are disproportionately represented.
  To correct for the effects of this misrepresentative sample on my estimates, I chose to 'replicate' observations until my new expanded sample held the proper characteristics. This can be achieved either through the Stata command to replicate, or equivalently to reweight each observation. The results from performing the robust probit regression on this corrected sample along with the marginal effects:

```
                                                mfx compute, dydx at(mean)

                                                Marginal effects after probit
                                                y  = Pr(measured) (predict)
                                                             0.06537169
```

measured	Coef.	Std. Err.	variable	dy/dx	Std. Err.
doct	1.189761	.0229788			
dnov	1.182631	.0229751	doct*	.2354581	0.00521
dage25	.1289945	.0357028	dnov*	.233095	0.00519
dage35	.1913289	.0381255	dage25*	.0171973	0.00488
dage45	.2051146	.0412451	dage35*	.0264439	0.0056
dage55	.1288734	.041772	dage45*	.029246	0.00647
dage65	.201911	.0405668	dage55*	.0176486	0.00601
hhsize	-.0773088	.0071571	dage65*	.0284472	0.00611
lincval	.1533875	.0107355	hhsize	-.0098458	0.00087
moved	.0401479	.0079769	lincval	.0195349	0.00136
sex	-.0104976	.0184615	moved	.0051131	0.00099
_cons	-3.515928	.1123403	sex*	-.0013403	0.00233

- Mixed Logit

  I now allow the regressors 'lincval' (log of income) and 'moved' to have random coefficients. That is, I assert that there exist different types of households, unobserved to the econometrician, whom are effected differently by changes in income and whether or not they moved.

  My justification for different types of households with respect to influence of income on number of calls (and thus probability of choosing measured) is that some households will take advantage of extra income to make less calls (less work related calls dominate) while some households will take advantage of extra income to make more calls (more leisure calls dominate).

  My justification for different types of households with respect to influence of whether they've moved on number of calls made (and thus probability of choosing measured) is that some households may have decided to move to get away from family commitment or because they did not have family commitment, while some households may have had to move for outside reasons and retain strong family connections in a different area. Those households in the first case will have equal or less calls post-move, while those households in the second case will have more calls post-move.

I begin with the default Stata specification that both random coefficients are distributed normally and independently.

measured	Coef.	Std. Err.
doct	2.466394	.0619814
dnov	2.453556	.0619261
dage25	.2482143	.0974948
dage35	.3407629	.1038159
dage45	.2984918	.1111216
dage55	.1938001	.1120612
dage65	.3144931	.1084588
hhsize	-.1698784	.018923
sex	-.0788589	.0508938
_cons	-1.462856	.1376272
Random-effects	Estimate	Std. Err.
lincval: Identity		
sd(_cons)	.246523	.0762592
moved: Identity		
sd(_cons)	.1946099	.0495011

Similar to original logit regression, almost all estimates are significant
Distribution parameter estimates are now significant.
The estimate on the constant is now smaller, presumably absorbed into the random coefficients parameterization.
Other estimates are comparable to before.

I now change the distributional assumption on the coefficients.

Coefficient of log income (lincval) is random and distributed binomial.

measured			coeff	std. error
	doct	|	2.4571	0.061754
	dnov	|	2.4443	0.0617
	dage25	|	0.2835	0.09658
	dage35	|	0.3924	0.10462
	dage45	|	0.3573	0.111676
	dage55	|	0.2472	0.112457
	dage65	|	0.3745	0.108806
	hhsize	|	-0.1636	0.018977
	sex	|	-0.0888	0.050749
	moved	|	0.0646	0.021456
	_cons	|	-1.7077	0.122588
random effects				
	lincval		0.0643	0.014847

Or, if only coefficient of moved is random and distributed binomial (below),

measured		coeff	std. error
	doct |	2.4582	0.061761
	dnov |	2.4454	0.061706
	dage25 |	0.2446	0.096093
	dage35 |	0.3379	0.103463
	dage45 |	0.2935	0.111392
	dage55 |	0.2014	0.112833
	dage65 |	0.3461	0.10843
	hhsize |	-0.1712	0.018882
	sex |	-0.0659	0.050509
	lincval |	0.2774	0.029015
	_cons |	-4.1297	0.288683
random effects			
	moved	0.015	0.010139

Mixed Logit Continued

I now account for the choice-base sampling error in the mixed logit regression.

measured	Coef.	Std. Err.
doct	2.481123	.0508415
dnov	2.469643	.0508569
dage25	.2240095	.068599
dage35	.3601107	.0727557
dage45	.4338075	.0786154
dage55	.2522944	.0795736
dage65	.3677966	.0787582
hhsize	-.1352061	.0130856
sex	.0029419	.0345196
moved	.0745563	.0326936
_cons	-3.932945	.1495689
Random-effects Parameters		
lincval: Identity		
sd(_cons)	.3073629	.08775
moved: Identity		
sd(_cons)	.2379512	.04468

Compared to the previous estimates, correcting the sampling error did not greatly change the sign and values of our estimates.
The standard errors are smaller however.
The greatest change in the estimates from correcting the sampling error occurred in the estimates of the parameters of the distribution over the random coefficients.
Including the replication correction made the distribution parameters such that a higher mass of individuals will now be likely to choose measured service over flat service.
The interpretation is that correcting for the sampling error influenced parameter estimates in a way to suggest that a higher mass of agents will make less calls from having more income or from moving.

- Model of Usage

I now simply regress my measure of calls (number of calls) on my same explanatory variables.

```
regress numcalls doct dnov dage25 dage35 dage45 dage55 dage65 hhsize lincval moved sex
```

numcalls	Coef.	Std. Err.
doct	-62.0687	12.4795
dnov	44.43511	12.4795
dage25	46.08758	21.04258
dage35	161.4222	22.59311
dage45	146.4985	24.40717
dage55	72.10394	24.66315
dage65	-9.829858	23.69293
hhsize	134.9965	4.047997
lincval	-127.2704	6.34255
moved	-9.25959	4.648467
sex	34.00415	11.00311
_cons	1351.459	62.77625

Little confidence should be put into the resulting estimates of the coefficients. The lack of confidence is attributed to the fact that regression includes data on all household's number of calls. The problem is that households who have chosen flat no longer face the same environment when deciding on how many phone calls to make. Once the decision has been made to select flat rates, households are likely to make more phone calls as the marginal cost is no longer present, and thus estimates of the coefficients will result in over-predicting desired phone usage.

- Addressing Sample Selection
  With our understanding of the model, we can use the sample selection information to obtain more efficient estimates. In addition to the initial binary choice regressions explaining probabilities of selecting measured versus flat, we can also use the distribution of the number of calls made by households who chose the measured option. In order to correctly incorporate this information into our regression, I proceed with a Heckit sample selection model.
  The estimates for the coefficients facing the household 'measured vs. flat option' and the household 'number of calls decision' are as follows

	Coef.	Std. Err.
numcalls		
dage25	1.480987	25.34241
dage35	124.4435	27.09594
dage45	123.5587	29.08285
dage55	25.79891	29.57902
dage65	-37.98362	28.36124
hhsize	109.4711	4.95139
lincval	-81.89527	7.91448
moved	-13.79619	5.279244
sex	35.27914	12.61712
_cons	957.8143	79.62146

measured		
doct	1.476687	0.035346
dnov	1.489066	0.035196
dage25	0.14383	0.056852
dage35	0.2011684	0.06114
dage45	0.1786588	0.066353
dage55	0.1317074	0.066595
dage65	0.2158088	0.063947
hhsize	-0.1010326	0.011089
lincval	0.1623141	0.017348
moved	0.0378572	0.012678
sex	-0.039571	0.029876
_cons	-2.484706	0.172918

  I offer caution in accepting these results. According to these results, there appear to be many household that are not optimizing.
  In one case, households who chose flat would have been better off to have chosen measured. Perhaps such household just wanted to simplify their decision and not think about whether they should be making calls.
  Or the opposite, that household who chose measured would have payed less if they had chose flat. Since measured was the default payment option, the reconciliation of this optimization failure is a story that households who chose this option may just have failed to switch their payment plan.
  My justification for lack of confidence in the assumption that households were optimizing in this model is the inconsistency that many households on the flat rate made less calls than households on the measured rate. This inconsistency makes it difficult to identify the threshold level of phone calls; the level such that if a household expects to make less should choose measured, and if expects to make more than should choose flat.
  According to my guess from my 'cleaned' data set, if the equal cost breakpoint on number of calls was around the minimum number of calls made among households choosing flat, then 3000/4500 households who chose measured would have been better off from choosing the flat rate.
  Another justification for this inconsistency could be that households face serious ex-ante uncertainty.

- Using Panel Structure of Data
  Rather than treating households/month as separate observations, we can make use of the fact that the same household was making the decision in 3 months. I employ Stata's xtprobit (time series probit) command for estimating the coefficients in this panel random-effects binary-choice model. The number of nodes indicates the number of points used in performing quadrature approximation of the integral representing the mass of epsilons resulting in the case of 'measured' being preferred to 'flat'.

```
xtprobit measured dage25 dage35 dage45 dage55 dage65 hhsize lincval moved sex, re intpoints(5)
```

measured	Coef.	Std. Err.
doct	3.904146	.1194695
dnov	3.861189	.1172251
dage25	.4803211	.2752552
dage35	.6520209	.2953304
dage45	.5245998	.315062
dage55	.4096317	.3180908
dage65	.7230073	.3098368
hhsize	-.3460554	.0525513
lincval	.5549332	.0798267
moved	.1226667	.0598581
sex	-.1773495	.1491429
_cons	-7.773055	.7887152
/lnsig2u	2.013661	.0671135
sigma_u	2.736913	.0918419
rho	.882224	.0069734

5 integration nodes
Took a matter of seconds using UT remote desktop.

Most estimates are again significant and of the same size. The magnitudes are generally larger. The probability of a household choosing measured increases (or decreases in the case of hh size) more dramatically now that we're considering the time-series aspect of the data, the information that it is the same household making three decisions.

In addition we get estimates for serial correlation. Rho is significant and close to one.

Estimates and standard errors do not change much with more integration nodes, but time of computation increases noticeably.

$$T \sim p^2\{M + M(N_Q)^{q_t}\}$$

Where p is the number of estimable parameters
M is the number of lowest-level (smallest) panels
N_Q is the number of quadrature points
q_t is the total dimension of the random effects (all levels)

Using Panel Structure continued

For 10 integration nodes (seconds);For 15 Integration Node (184 seconds)

measured	Coef.	Std. Err.	measured	Coef.	Std. Err.
doct	3.980788	.1078113	doct	3.970222	.1068785
dnov	3.936062	.1059247	dnov	3.925866	.1050198
dage25	.4213568	.2406041	dage25	.4170989	.2294853
dage35	.5391851	.2571357	dage35	.5333191	.2458764
dage45	.4215151	.2764084	dage45	.4164903	.2646929
dage55	.3649535	.281506	dage55	.361318	.268564
dage65	.6115541	.2711332	dage65	.6046329	.2585837
hhsize	-.3131244	.0471331	hhsize	-.3091587	.0451323
lincval	.4531663	.0738449	lincval	.4469435	.0701539
moved	.0974781	.0516377	moved	.0966222	.0500518
sex	-.1611409	.1230555	sex	-.1603212	.1188824
_cons	-6.706134	.7348831	_cons	-6.643488	.6991284
/lnsig2u	2.016629	.0645616	/lnsig2u	2.008134	.064803
sigma_u	2.740977	.088481	sigma_u	2.729359	.0884353
rho	.882532	.0066931	rho	.8816484	.0067618