Efficient and Practical Econometric Methods for the SLID, NLSCY, NPHS

Philip Merrigan
ESG-UQAM, CIRPÉE

Using Big Data to Study
Development and Social Change, Concordia University,
November 2103

Intro

- Longitudinal Survey of Income and Earnings (SLID)
- National Longitudinal Survey of Children and Youth (NLSCY)
- National Population Health Survey (NPHS)
- All short panels (3 to 8 periods)
- Only available in *Stats Can* Research Data Centres

Data Formatting

- *SLIDRETS RDC software will sample SLID for you cross-sectionally and longitudinally.*
- *NPHS - Each wave includes data for all waves, one record per individual must be transformed in panel data form, one record per individual per wave.*
- *NLSCY - One data set per wave, different names in each wave, must be transformed as in the NPHS.*

  (Programs soon available at CIQSS to perform this task.)

Non-Linear Models for Longitudinal Data

- Mixed models for Longitudinal Data are difficult to work with if data sets are large and specifications rich.
- Advances in computing power have greatly increased the speed with which parameters of these models can be estimated.
- Major Statistical Software packages now include mixed model procedures for non-linear models.

Non-Linear Models for Longitudinal Data

- STATA has offered for many years efficient routines for the panel probit, logit, poisson and has recently added the ordered logit and ordered probit procedures. They admit individual heterogeneity for the intercept.

Non-Linear Models for Longitudinal Data

- GLAMM, a STATA package written by Rabe-Hesketh, S., Skrondal, A. and Pickles, A. performs the same procedures as the STATA package but adds several others including Survival models, and the multinomial logit.
- GLAMM admits random slopes in the models as well as random intercepts, greatly increasing the number of numerical computations for estimation.

Non-Linear Models for Longitudinal Data

- Given the size of the panels, fixed effects as implemented with dummy variables for each individual produce inconsistent estimates in all non-linear methods (probit, logit, poisson, multinomial logit, etc.).
- However, in most cases, with non-experimental data, spurious correlation is a major problem and the introduction of individual fixed effects is crucial to enhance the credibility of estimates as causal.

Non-linear models

- For the logit binary form, the conditional logit of Chamberlain (stata xtlogit, fe) will handle fixed effects at the individual level.
- Unfortunately, average marginal effects cannot be recovered because the method provides no estimates of the fixed effect distribution.
- If estimation must be conditioned on fixed effects and if marginal effects are necessary to achieve the goals of the paper, (as is generally the case), Wooldridge (2011) provides an easily implementable procedure to include fixed effects in the estimation.

Simple method to include fixed individual effects in Non-linear models

- Say y*(it) is the latent dependent variable, x(it) is a regressor, c(i) is the individual effect, and u(it) the error term. Hence,
- y*(it) = b(0) + b(1)*x(it) + c(i) + u(it),

  i=1…N, t=1…T.
- Woolridge's suggestion is to write:

  c(i) = a(0) + g(1)*mx(i) + v(i)

where mx(i) is the mean of x(i) over T.

Simple method to include fixed individual effects in Non-linear models

- We get

  $y^*(it) = c(0) + b(1)*x(it) + g(1)*mx(i) + v(i) + u(it)$, $i=1\ldots N$, $t=1\ldots T$.
- We assume that all the correlation between x(it) and c(i) is captured by the presence of mx(i) that we include in the regression.
- Given that most non-linear models have a linear index, or several of these linear indexes (multinomial models for example), this « trick » can be used in a large number of cases.

Simple method to include fixed individual effects in Non-linear models

- A simple test of individual heterogeneity is the null that the coefficients on the means are equal to 0 which, in most cases, will be rejected.
- If there are several time varying regressors in the model, there will be a large number of coefficients to estimate - even more so, if one includes interaction terms for the means. But with advances in computing power, all this becomes feasible.

Simple method to include fixed individual effects in Non-linear models

- Caveat: this method adds credibility to a causal interpretation for variables that change value from one period to the next, but does little to address spurious correlation for variables that are constant during the sample period although they must be included in the model.
- Very useful, when one is trying to estimate the effects of income on discrete outcomes such as good health or outcomes on child development.
- The idea is to obtain credible estimates without using instrumental variable procedures.

Simple method to include fixed individual effects in Non-linear models

- Unfortunately, the STATA xt methods with the « random effect » option will not provide average marginal effects, but average marginal effects for a specific value of v(i), i.e. 0.
- However, the « pa » option in the xt commands with the « exchangeable » option for the correlation structure of the error term will provide estimates very similar to the « re » option, but will also provide estimates of the average marginal effects with standard errors.

Simple method to include fixed individual effects in Non-linear models

- Must be careful when estimating marginal effects not to confuse effects computed and different values of v(i) with average marginal effects. (Good discussion in Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata, Third Edition*. College Station, TX: Stata Press.)

Simple method to include fixed individual effects in Non-linear models

- For models with random slopes as well as random intercepts, GLAMM does provide estimates of marginal probabilities as well as estimated probabilities for particular values of the individual heterogeneity term. However, there are no procedures for estimating marginal effects.
- Because predicted probabilities are available, it is quite simple to program estimates of marginal effects. However, standard errors or marginal effects will take longer to compute if one uses bootstrap methods, even with very powerful computers.

Longitudinal estimation with bootstrap weights

- *Stats Can* data sets provide bootstrap weights permitting inference that takes into account the complex sampling design for these surveys.
- With longitudinal data, in general, weights are provided for individuals who are sampled each wave. For the NLSCY and NPHS, this means much smaller data sets if one uses the 8 waves (they remain large).

Missing values

- With high attrition, the issue of missing values becomes more important.
- SAS and STATA now offer superb packages (PROC MI and mi) to perform multiple imputation for almost any type of continuous or discrete dependent or explanatory missing value using bayesian markov chain methods.
- Again, when data sets are large, this involves a very large number of computation. For example, STATA suggests 100 burn-ins.

Missing values

- Suggestions: first, perform regressions without missing values or with dummy variables indicating missing values. If a preferred specification is found, then, re-run with 10 imputed data sets to measure sensitivity of results to missing values.
- The same can be said of inference with bootstrap « weights ».

Conclusion

- Computing time is an important challenge for Big Data, in particular, for non-linear panel data models.
- The most recent generation of computers with parallel processing permits the estimation of complex models in reasonable amounts of time.
- The addition of fixed individual effects in models which seek to find causal links is problematic.

  When spurious correlation is an evident problem, we detailed a fixed effect procedure that can be easily implemented. (example: impact of income on health or child development outcomes).
- Multiple imputations at a minimum should be used to address missing value problems.