Nonlinear Models in Panel Data

Whitney K. Newey
MIT
September 2007

Great thing about panel data is that it allows us to control for individual effects that are correlated with regressors. Well known how to do this in linear models. These notes are about what can be done in nonlinear models.

Likelihoods with Individual Effects

Data: $Y_i = [Y_{i1}, ..., Y_{iT}]', X_i = [X_{i1}, ..., X_{iT}]', (i = 1, ..., n)$.

To motivate the model we consider the linear model as a starting point:

$$Y_{it} = X_{it}'\beta + \alpha_i + \eta_{it}, E[\eta_{it}|X_i, \alpha_i] = 0.$$

Alternative, equivalent formulation:

$$E\left[Y_{it}|X_i, \alpha_i\right] = X_{it}'\beta + \alpha_i.$$

The model specifies the conditional mean of Y_i given X_i, α_i. A likelihood version of this specifies the conditional pdf $f(y|x, \alpha, \theta)$ of Y_i given X_i, α_i and a parameter vector θ.

Ex: Normal linear model: For e_T a $T \times 1$ vector of $1'$s,

$$Y_i|(X_i, \alpha_i) \sim N(X_i\beta + \alpha_i e_T, \sigma^2 I_T).$$

This is distributional version of a linear model.

Ex: Binary choice model: $Y_{it} \in \{0, 1\}$; such as labor force participation.

$$Y_{it}, (t = 1, ..., T) \text{ independent}, \text{Prob}(Y_{it} = 1|X_i, \alpha_i) = G(X_{it}'\beta + \alpha_i).$$

Ex: Count data: $Y_{i1}, ..., Y_{iT}$ indep, $Y_{it}|X_i, \alpha_i$ Poisson with mean $\exp(X_{it}'\beta + \alpha_i)$.

The central question in nonlinear panel data models is how to control for presence of the individual effect α_i. Most methods that control for α_i in linear models do not extend to nonlinear ones. For example, differencing does not work. In the linear conditional expectation model, we have

$$E\left[Y_{it} - Y_{it-1}|X_i\right] = X_{it}'\beta + E[\alpha_i|X_i] - (X_{i,t-1}'\beta + E[\alpha_i|X_i]) = (X_{it} - X_{i,t-1})'\beta,$$

so can regress difference in Y on difference in X to consistently estimate β. In nonlinear model, α_i does not drop out when we difference. For example in binary choice model,

$$E\left[Y_{it} - Y_{it-1}|X_i\right] = E[G(X_{it}'\beta + \alpha_i) - G(X_{it-1}'\beta + \alpha_i)|X_i].$$

Here the α_i does not get differenced out, due to the nonlinearity of $G(\bullet)$. Discussion question: Does using the linear probability model fix this problem?

Fixed Effects and the Incidental Parameters Problem

Fixed effects is generally inconsistent in a nonlinear model as n grows with T fixed. Here by fixed effects we mean maximizing the log-likelihood over each α_i as well as θ. In a linear model, when we do least squares treating α_i as a parameter to be estimated we do get consistency. When we do maximum likelihood treating α_i as a parameter to be estimated we generally do not. This is known as the *incidental parameters problem.* It is caused by only having T observations to estimate each α_i, so that as n grows the estimate of α_i remains random. In linear models this randomnes gets "averaged out." In nonlinear models it does not.

To be more precise we can derive an expression for the limit of the fixed effects estimator as n grows with T fixed. The estimator is

$$\hat{\theta} = \arg \max_{\theta, \alpha_1, \ldots, \alpha_n} \frac{1}{n} \sum_{i=1}^{n} \ln f(Y_i|X_i, \theta, \alpha_i).$$

Alternatively, if we concentrate out each α_i, for a fixed θ each fixed effect is given by

$$\hat{\alpha}_i(\theta) = \max_{\alpha} \ln f(Y_i|X_i, \theta, \alpha_i).$$

Substituting in and maximize over θ to get $\hat{\theta}$,

$$\hat{\theta} = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ln f(Y_i|X_i, \theta, \hat{\alpha}_i(\theta)).$$

By the usual extremum estimator, as n grows for fixed T the estimator $\hat{\theta}$ has plim

$$\theta_T = \arg \max_{\theta} E[\ln f(Y_i|X_i, \theta, \hat{\alpha}_i(\theta))].$$

Randomness in $\hat{\alpha}_i(\theta)$ leads to inconsistnecy of $\hat{\theta}$. If $\hat{\alpha}_i(\theta)$ were replaced by

$$\bar{\alpha}_i(\theta) = \arg \max_{\alpha} E[\ln f(Y|X, \theta, \alpha)],$$

would get consistency. So, the problem is a kind of a measurement error in this nonlinear model.

Ex: Binary logit, $Y_{it} \in \{0,1\}$,

$$\Pr(Y_{it} = 1|X_i, \alpha_i) = \exp(\beta_0 X_{it} + \alpha_i)/[1 + \exp(\beta_0 X_{it} + \alpha_i)].$$

It is known that the fixed effects estimator $\hat{\beta}_{FE}$ satisfies

$$\hat{\beta}_{FE} \xrightarrow{p} 2\beta_0$$

Thus, bias can be severe.

Conditional Maximum Likelihood

Sometimes there is a statistic S_i such that α_i drops out of the conditional likelihood of Y_i given X_i and S_i. In such a case,

$$\begin{aligned} f(Y_i|X_i, S_i, \beta, \alpha_i) &= f(Y_i|X_i, S_i, \beta), \\ \hat{\beta} &= \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ln f(Y_i|X_i, S_i, \beta). \end{aligned}$$

This estimator is consistent and asymptotically normal, as usual for a conditional MLE. Also, it is asymptotically efficient when the distribution of α_i conditional on X_i is unknown. Thus, conditioning on S_i provides an excellent solution. The problem is that such an S_i only exists in a few cases. These include the Gaussian linear model, binary choice logit, the Poisson model for count data, and the proportional hazards model. In most other models there is no such S_i. Thus, the conditional MLE has limited usefulness.

Correlated Random Effects:

An approach that does apply generally is to model the distribution of α_i conditional on X_i. In a likelihood setting, such a model corresponds to a p.d.f. of α_i given X_i, which we denote by $g(\alpha|X, \gamma)$, where γ are the parameters of this model. The the conditional likelihood of Y given X is then obtained by integrating out α, as

$$f(Y|X, \beta, \gamma) = \int f(Y|X, \beta, \alpha) g(\alpha|X, \gamma) d\alpha.$$

The MLE is given by

$$\hat{\beta}, \hat{\gamma} = \arg\max_{\beta,\alpha} \frac{1}{n} \sum_{i=1}^{n} \ln f(Y_i|X_i, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \ln \int f(Y_i|X_i, \beta, \alpha) g(\alpha|X_i, \gamma) d\alpha$$

This approach is very general, but the consistency of $\hat{\beta}$ depends on the $g(a|X, \gamma)$ being correctly specified. Also, it may be difficult to calculate the integral. More fundamentally, these models may not be time consistent, in that the form of them changes if more time periods are included.

Ex: Correlated random effects probit. Suppose that conditional on (X_i, α_i), the latent variables $Y_{i1}^*, ..., Y_{iT}^*$ are independent and Y_{it}^* has distribution $N(X_{it}'\beta + \alpha_i, \sigma_t^2)$. Let $x_i = vec(X_i')$ be the vector of all observations across t on the regressors. Suppose also that the conditional distribution of α_i given X_i is $N(x_i'\lambda, \sigma_\alpha^2)$. Assume that the observed binary variables Y_{it} satisfy $Y_{it} = 1(Y_{it}^* > 0)$. Then for $\theta = (\beta', \lambda', \sigma_1^2, ..., \sigma_T^2, \sigma_\alpha^2)$,

$$\begin{aligned}
\Pr(Y_{it} &= 1|X_i, \theta) = \int \Phi((X_{it}'\beta + \alpha)/\sigma_t)\sigma_\alpha^{-1}\phi((\alpha - x_i'\lambda)/\sigma_\alpha^2)d\alpha \\
&= \int\int_{-\infty}^{(X_{it}'\beta+\alpha)/\sigma_t} \phi(r)\sigma_\alpha^{-1}\phi((\alpha - x_i'\lambda)/\sigma_\alpha^2)d\alpha dr \\
&= \int\int_{-\infty}^{X_{it}'\beta} \sigma_t^{-1}\phi((u+\alpha)/\sigma_t)\sigma_\alpha^{-1}\phi((\alpha - x_i'\lambda)/\sigma_\alpha^2)d\alpha du \\
&= \int\int_{-\infty}^{X_{it}'\beta} \phi((u - x_i'\lambda)/\sqrt{\sigma_t^2 + \sigma_\alpha^2})d\alpha du \\
&= \Phi\left(\frac{X_{it}'\beta + x_i'\lambda}{\sqrt{\sigma_t^2 + \sigma_\alpha^2}}\right);
\end{aligned}$$

where the second equality follows by Φ being the standard normal CDF, the third by the change of variables $u = \sigma_t r - \alpha$, and the fourth by the fact that the integral over α corresponds to a mixture of $N(0, \sigma_t^2)$ and $N(x_i'\lambda, \sigma_\alpha^2)$.

One could also derive a joint probability, but it is complicated because of correlation across time periods. That would be needed for for MLE, but can estimate just from marginal probabilities for each time period. Idea is to estimate the probability given x, and then do minimum distance to estimate β and other parameters. Let e_t denote the t^{th} $T \times 1$ unit vector and

$$\pi_t = \frac{e_t \otimes \beta + \lambda}{\sqrt{\sigma_t^2 + \sigma_\alpha^2}},\ e_t = t^{th}.$$

Then we have

$$\Pr(Y_{it} = 1|X_i, \theta) = \Phi(x_i'\pi_t).$$

Thus, we can do probit on each time period separately to obtain $\hat{\pi}_1, ..., \hat{\pi}_T$. Let $\delta_t = 1/\sqrt{\sigma_t^2 + \sigma_\alpha^2}$, $(t = 1, ..., T)$, where we normalize $\delta_1 = 1$. Reparameterize so that $\theta = (\beta', \lambda', \delta_2, ..., \delta_T)'$ and for $\pi = (\pi_1', ..., \pi_T')'$ let

$$h(\pi, \theta) = \begin{pmatrix} \delta_1\pi_1 - e_1 \otimes \beta - \lambda \\ \vdots \\ \delta_T\pi_T - e_T \otimes \beta - \lambda \end{pmatrix}.$$

We can then do minimum distance, using the individual probit $\hat{\pi}$ mentioned above.

Here is an empirical example from Chamberlain's (1984) Handbook of Econometrics Chapter. It is a labor force participation example, with 924 women, for 1968, 70, 72, 74.. The two regressors are number of children under 6 and number of children. Here are the results:

Probit	-.121	-.058	Logit	.-573	-.336
	(.046)	(.029)		(.115)	(.120)

Quite different estimates; ratios are similar.

The Chamberlain (correlated random effects) estimator is troubling in that it depends on T in an essential way. Also, there are many coefficients in π. A more parsimonius model, less sensitive to time specification is to assume that $\alpha_i \sim N(\lambda'\bar{x}, \sigma_\alpha^2)$ conditional on X_i.

Important question is what the parameter of interest is. In some contexts it is β, which might be parameters of utility function. However, in binary choice we might want to consider "average sructural function"

$$\mu(X) = \int \Phi((X'\beta + \alpha)/\sigma_t) f(\alpha) d\alpha$$

By iterated expectations, holding X fixed,

$$\begin{aligned} \mu(X) &= E[E[\Phi((X'\beta + \alpha_i)/\sigma_t)|X_i]] \\ &= E[\Phi(\delta_t(X'\beta + x_i'\lambda))] \end{aligned}$$

This object can be estimated by $\hat{\mu}(X) = \sum_{i=1}^n \Phi(\hat{\delta}_t(X'\hat{\beta} + x_i'\hat{\lambda}))/n$.

Some Semiparametric Results

There are some distribution free results that are useful. An example is Poisson model, where conditional on X_i and α_i the variable Y_{it} is independent over time and Poisson with mean $e^{X_{it}'\beta + \alpha_i}$. Good model for count data with patents. Woodridge showed that consistency of CMLE only requires

$$E\left[Y_{it}|X_i, \alpha_i\right] = e^{X_{it}\beta + \alpha_i}$$

This is a good exercise.

Honore has results for Tobit. See Handbook of Econometrics chapter by Arellano and Honore on Honore's website. Manski had a maximum score estimator for binary choice model with fixed effect. Weakness of both of these is require homoskedasticity over time, an assumption almost never satisfied.

Fixed Effects Again

The difficulty of finding consistent estimators for these models has led to reexamination of fixed effects. Recently been found in Monte Carlo studies that in spite of the inconsistency, bias not large in applications. Also, large T bias corrections have been derived.

One can use a simple expansion to consider how bad fixed effects bias is and how to correct. Intuitively, as T grows the randomness in the estimated fixed effects should go away and hence $\lim_{T\longrightarrow\infty}\theta_T = \theta_0$. One can show more under certain smoothness conditions, that

$$\theta_T = \theta_0 + \frac{B}{T} + O(\frac{1}{T^2}).$$

Assume also that as n and T both grow, the fixed effects estimator is asymptotically normal when centered at its plim, so that

$$(nT)^{1/2}\left(\widehat{\theta} - \theta_T\right) \xrightarrow{d} N(0, \Omega).$$

Consider then what happens when n and T grow at the same rate, i.e. $n/T \to \rho$. We have

$$\begin{aligned}
(nT)^{1/2}\left(\widehat{\theta} - \theta_0\right) &= (nT)^{1/2}\left(\widehat{\theta} - \theta_T\right) + (nT)^{1/2}(\theta_T - \theta_0) \\
&= (nT)^{1/2}\left(\widehat{\theta} - \theta_T\right) + (nT)^{1/2}\frac{B}{T} + O((nT)^{1/2}/T^2) \\
&\xrightarrow{d} N\left(B\rho^{1/2}, \Omega\right).
\end{aligned}$$

Here there is asymptotic bias even when T grows at the same rate as n. Consequently, asymptotic confidence intervals for the fixed effects esimator will be asymptotically incorrect even when T grows at the same rate as n.

A bias corrected estimator could be formed using at estimator $\hat{B}$ of B,

$$\hat{\theta}_1 = \hat{\theta} - \hat{B}/T.$$

Suppose that the bias correction $\hat{B}$ is well estimated in the sense that

$$(nT)^{1/2}(\hat{B} - B)/T \xrightarrow{p} 0.$$

Assume that $n/T^3 \longrightarrow 0$, i.e. T grows faster than the cube root of n. Plugging in as before we get,

$$\begin{aligned}
(nT)^{1/2}\left(\widehat{\theta}_1 - \theta_0\right) &= (nT)^{1/2}\left(\widehat{\theta} - \theta_T\right) + (nT)^{1/2}(\theta_T - \theta_0 - \hat{B}/T) \\
&= (nT)^{1/2}\left(\widehat{\theta} - \theta_T\right) + (nT)^{1/2}(B - \hat{B})/T + O((nT)^{1/2}/T^2) \\
&\xrightarrow{d} N\left(0, \Omega\right).
\end{aligned}$$

The condition $n/T^3 \longrightarrow 0$ suggests this may lead to decent estimators in sample sizes typical in econometrics, e.g. $n = 1000$, $T > 10$.

Formulas for $\hat{B}$ complicated: See Hahn and Newey (2004) Econometrica.

Monte Carlo Example: Like Heckman (1981). Design is:

$$\begin{aligned} y_{it} &= 1(x_{it}\theta_0 + \alpha_i + \varepsilon_{it} > 0), \\ \alpha_i &\sim N(0,1), \varepsilon_{it} \sim N(0,1), \\ x_{it} &= t/10 + x_{i,t-1}/2 + u_{it}, \\ x_{i0} &= u_{i0},\, u_{it} = U(-1/2, 1/2). \\ N &= 100, T = 8; \beta = 1, -1. \end{aligned}$$

Results for estimators of θ_0. Also, estimators of average of the derivative of the choice probability $\Phi(x'\theta + \alpha)$ with respect to x at a particular $x = w$, which is

$$\mu = \theta_0 \bar{E}[\phi(w'\theta_0 + \alpha_i)].$$

The fixed effects estimator of this object is

$$\hat{\mu} = \hat{\theta} \sum_{i=1}^{n} \phi\left(w'\hat{\theta} + \hat{\alpha}_i\right)/n.$$

Table Three: Properties of $\hat{\theta}$, $T = 8$.					
Estimator of θ_0	Mean	Med.	SD	$\hat{p}; .05$	$\hat{p}; .10$
MLE	1.18	1.17	.151	.267	.370
Jackknife	.953	.950	.119	.056	.102
Analytic	1.05	1.05	.134	.062	.135
Analytic-M	1.05	1.05	.132	.060	.126

Table Five: Properties of $\hat{\theta}$, $T = 4$					
Estimator of θ_0	Mean	Med.	SD	$\hat{p}; .05$	$\hat{p}; .10$
MLE	1.42	1.41	.397	.269	.373
Jackknife	.752	.743	.262	.100	.177
Analytic	1.12	1.11	.306	.055	.101
Analytic-M	1.21	1.20	.335	.102	.172

Table Four: Properties of $\hat{\mu}$, $T = 8$.					
Estimator of μ/μ_0	Mean	Med.	SD	$\hat{p}$;.05	$\hat{p}$;.10
MLE	1.02	1.02	.131	.078	.140
Jackknife	1.00	.992	.130	.086	.159
Analytic	1.02	1.02	.133	.090	.153
Analytic-M	1.02	1.02	.131	.087	.154

Table Six: Properties of $\hat{\mu}$, $T = 4$.					
Estimator of μ/μ_0	Mean	Med.	SD	$\hat{p}$; .05	$\hat{p}$; .10
MLE	1.00	1.00	.257	.103	.168
Jackknife	1.06	1.05	.307	.159	.224
Analytic	.996	.994	.265	.113	.178
Analytic-M	1.05	1.05	.266	.117	.185