The Incidental Parameters Problem

Bryan S. Graham, UC - Berkeley & NBER

March 26, 2021

In 1948 Jerzy Neyman and Elizabeth Scott published a remarkable paper in *Econometrica.* In fact it was the only paper appearing the January 1948 issues of *Econometrica.* It also represents the only time Neyman, who along with Sir Ronald Fisher, is *the* giant of 20th century statistics, published in an economics journal. Neyman's coauthor, Elizabeth Scott, was a Ph.D. student in Astronomy when the paper was written (and the questions posed in the paper appears to have arisen from questions related to measurement error in astronomical data). Although Scott was a trained astronomer, she was unable to pursue a career in that field because women were prevented from using the telescope at Berkeley at the time. She instead pursed a (distinguished) career in statistics – obtaining an appointment in the Berkeley Math Department in 1951. An award in her honor is presented every two years at the Joint Statistical Meetings.

The following example appears in the paper. Let (Y_{i1}, Y_{i2}) be independent $\mathcal{N}(\alpha_i, \sigma^2)$ random variables for $i = 1, \ldots, N$. The location parameter for each i – α_i– is different, while the scale parameter, σ^2, is common to all units. We will call $\{\alpha_i\}_{i=1}^N$ the *incidental parameters* and σ^2 the common parameter. It is a good exercise to verify that the MLEs for this problem are

$$\begin{aligned}\hat{\alpha}_i &= \frac{Y_{i1} + Y_{i2}}{2},\ i = 1, \ldots, N \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^N \frac{(Y_{i1} - \hat{\alpha}_i)^2 + (Y_{i2} - \hat{\alpha}_i)^2}{2}. \\ &= \frac{1}{4N}\sum_{i=1}^N (Y_{i1} - Y_{i2})^2 .\end{aligned}$$

First observe that $\hat{\alpha}_i \sim \mathcal{N}\left(\alpha_i, \frac{\sigma^2}{2}\right)$. The MLEs of the incidental parameters, while in this case unbiased, are not consistent (for $N \to \infty$). Second, consider the common parameter σ^2. Since, $\hat{\alpha}_i = \frac{Y_{i1}+Y_{i2}}{2}$ we have $\hat{\sigma}^2 = \frac{1}{4N}\sum_{i=1}^N (Y_{i1} - Y_{i2})^2$. Next observe that $\mathbb{E}\left[(Y_{i1} - Y_{i2})^2\right] =$

$2\sigma^2$ and hence $\hat{\sigma}^2 \xrightarrow{p} \frac{\sigma^2}{2}$. The estimate of the common parameter is also inconsistent. Like James & Stein (1961) over a decade later, Neyman & Scott (1948) provided an example where MLE performs poorly. This was very much against prevailing intuitions at the time.

This example is non-standard because the dimension of the nuisance (incidental) parameter $\{\alpha_i\}_{i=1}^N$ grows with N. Classical theorems on the asymptotic properties of MLE do not apply. Settings like this arise naturally in the context of panel data. In that context the incidental parameters model heterogeneity across cross sectional units, while the common parameters are of primary interest (although sometimes we may also be interested in the incidental parameters, or averages of (functions of) them). In this note we'll present a classic example of incidental parameters bias in a non-linear panel data model. The example is both interesting and historically important, but also helpful for developing intuitions about how MLE might behave in more complicated settings with incidental parameters. We will then consider two classic "solutions" to incidental parameters bias: (i) conditioning on a sufficient statistic for them, and (ii) integrated likelihood. The latter approach has natural connections to Bayesian statistics. The survey paper by Arellano (2003) is worth reading. Lancaster (2000) provides a historical survey of the incidental parameters problem. A classic reference in economics is Chamberlain (1980).

Joint fixed effects maximum likelihood

For each of $i = 1, \ldots, N$ randomly sampled agents we observe a binary choice, $Y_{it} \in \{0, 1\}$, and $K \times 1$ regressor vector, X_{it}, in each of $t = 1, \ldots, T$ periods. Choice in each period is determined according to the rule

$$Y_{it} = \mathbf{1}\left(X_{it}'\beta_0 + A_i - U_{it} \geq 0\right) \tag{1}$$

with U_{it} an unobserved (standard) logistic random variable; independently and identically distributed across agents and over time. Here A_i is time-invariant, agent specific heterogeneity that is unobserved by the econometrician.

Let $X_i = (X_{i1}', \ldots, X_{iT}')'$, $\mathbf{X} = (X_1, \ldots, X_N)'$ and $\mathbf{A} = (A_1, \ldots, A_N)'$. We will begin by considering the properties of the maximum likelihood estimates of β and $\mathbf{A}$. We'll call this the joint fixed effects MLE.

The log-likelihood function conditional on $\mathbf{X}$is

$$l_N\left(\beta, \mathbf{A}\right) = \sum_{i=1}^{N} l_i\left(\beta, A_i\right),$$

where

$$l_i(\beta, A_i) = \sum_{t=1}^{T} Y_{it} \ln F(X_{it}'\beta + A_i) + (1 - Y_{it}) \ln [1 - F(X_{it}'\beta + A_i)] \tag{2}$$

with

$$F(X_{it}'\beta + A_i) = \frac{\exp(X_{it}'\beta + A_i)}{1 + \exp(X_{it}'\beta + A_i)}.$$

The unit i's contribution to the score vectors for the common parameter, β, and the N incidental parameters, $\{A_i\}_{i=1}^N$, are

$$\begin{aligned} S_{\beta i}(\beta, A_i) &= \sum_{t=1}^{T} (Y_{it} - F(X_{it}'\beta + A_i)) X_{it} \\ S_{Ai}(\beta, A_i) &= \sum_{t=1}^{T} (Y_{it} - F(X_{it}'\beta + A_i)), \ .i = 1, \ldots, N. \end{aligned}$$

Conditional on the common parameter, β, the MLE of A_i is

$$\hat{A}_i(\beta) = \arg\max_{a \in \mathbb{R}^1} l_i(\beta, a)$$

or the solution to

$$S_{Ai}\left(\beta, \hat{A}_i(\beta)\right) = \sum_{t=1}^{T} \left(Y_{it} - F\left(X_{it}'\beta + \hat{A}_i(\beta)\right)\right) = 0. \tag{3}$$

An important feature of (3) is that, conditional on β, only unit i's observations contribute to our estimate of A_i.

A special case

In what follows we will consider the case with $T = 2$, $X_{i1} = (1, 0)'$ and $X_{i2} = (1, 1)$ for all $i = 1, \ldots, N$ and $\beta = (\alpha, \gamma)'$. That is the model consists of an "intercept" and a period 2 time dummy. This is a classic special case considered by Chamberlain (2010) and others. For a given value of β, there will only be four possible solutions to (3); one for each of the four possible sequences for the outcome across the two periods (note that there is no cross-sectional variation in X_i).

(Y_1, Y_2)	$l_i(\beta, A_i)$	$\hat{A}_i(\beta)$
$(0,0)$	$\ln[F(-\alpha - A_i)] + \ln[F(-\alpha - \gamma - A_i)]$	$-\infty$
$(0,1)$	$\ln[F(-\alpha - A_i)] + \ln[F(\alpha + \gamma + A_i)]$	$-\alpha - \frac{\gamma}{2}$
$(1,0)$	$\ln[F(\alpha + A_i)] + \ln[F(-\alpha - \gamma - A_i)]$	$-\alpha - \frac{\gamma}{2}$
$(1,1)$	$\ln[F(\alpha + A_i)] + \ln[F(\alpha + \gamma + A_i)]$	∞

Column 2 of the above table reports the form of the log-likelihood contribution for each of the four possible outcome sequences given in Column 1. You should verify that you can re-construct Column 2 on your own using (2) above. Column 3 reports the MLE of A_i (conditional on a given value of the common parameter β).

Consider the $(0,1)$ case first. Differentiating $l_i(\beta, A_i)$ with respect to A_i yields, for the $(0,1)$ case:

$$\begin{aligned} -\frac{F'(-\alpha - A_i)}{F(-\alpha - A_i)} + \frac{F'(\alpha + \gamma + A_i)}{F(\alpha + \gamma + A_i)} &= -(1 - F(-\alpha - A_i)) + 1 - F(\alpha + \gamma + A_i) \\ &= F(-\alpha - A_i) - F(\alpha + \gamma + A_i). \end{aligned}$$

Therefore $\hat{A}_i(\beta)$ satisfies

$$\begin{aligned} F\left(-\alpha - \hat{A}_i(\beta)\right) &= F\left(\alpha + \gamma + \hat{A}_i(\beta)\right) \\ -\alpha - \hat{A}_i(\beta) &= \alpha + \gamma + \hat{A}_i(\beta) \end{aligned}$$

and hence equals

$$\hat{A}_i(\beta) = -\alpha - \frac{\gamma}{2}.$$

For the $(1,0)$ case we get the same solution. For the $(0,0)$, the MLE of A_i is $-\infty$; whereas for the $(1,1)$ case it is ∞.

With these four possible values of $\hat{A}_i(\beta)$ we can form the concentrated log-likelihood

$$l_N^c(\beta) = \sum_{i=1}^{N} l_i\left(\beta, \hat{A}_i(\beta)\right).$$

The first thing to note is that the $(0,0)$ and $(1,1)$ cases make zero contribution to the concentrated log-likelihood. In contrast, each $(0,1)$ unit contributes

$$2\ln\left[F\left(\frac{\gamma}{2}\right)\right]$$

and each $(1,0)$ unit contributes

$$2\ln\left[1-F\left(\frac{\gamma}{2}\right)\right].$$

Hence the concentrated log-likelihood equals

$$l_N^c(\beta)=2\sum_{i=1}^{N}\left\{(1-Y_{i1})\,Y_{i2}\ln\left[F\left(\frac{\gamma}{2}\right)\right]+Y_{i1}\,(1-Y_{i2})\ln\left[1-F\left(\frac{\gamma}{2}\right)\right]\right\}.$$

Let $p_0=\Pr\left(Y_{i1}=0,Y_{i2}=1\middle|\,Y_{i1}+Y_{i2}=1\right)$. The MLE of p_0 is

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^{N}(1-Y_{i1})\,Y_{i2}}{\sum_{i=1}^{N}\left\{(1-Y_{i1})\,Y_{i2}+Y_{i1}\,(1-Y_{i2})\right\}}\\ &= \frac{\sum_{i=1}^{N}\mathbf{1}\left(Y_{i1}=0\right)\mathbf{1}\left(Y_{i2}=1\right)}{\sum_{i=1}^{N}\mathbf{1}\left(Y_{i1}+Y_{i2}=1\right)}\end{aligned}$$

By the invariance property of MLE we have that

$$\begin{aligned}\hat{\gamma} &= 2F^{-1}(\hat{p})\\ &\xrightarrow{p} 2F^{-1}(p_0).\end{aligned}$$

Next observe that, using the form of the logit choice model,

$$\begin{aligned}&\Pr\left(Y_{i1}=0,Y_{i2}=1\middle|\,Y_{i1}+Y_{i2}=1,A=a\right)\\ =\;&\frac{\left[1-F\left(\alpha+a\right)\right]F\left(\alpha_0+\gamma_0+a\right)}{\left[1-F\left(\alpha_0+a\right)\right]F\left(\alpha_0+\gamma_0+a\right)+F\left(\alpha_0+a\right)\left[1-F\left(\alpha_0+\gamma_0+a\right)\right]}\\ =\;&\frac{\exp\left(\alpha_0+\gamma_0+a\right)}{\exp\left(\alpha_0+\gamma_0+a\right)+\exp\left(\alpha_0+a\right)}\\ =\;&\frac{\exp\left(\gamma_0\right)}{1+\exp\left(\gamma_0\right)}\\ =\;&F\left(\gamma_0\right),\end{aligned}$$

The law-of-iterated expectations then gives

$$\begin{aligned}p_0 &= \Pr\left(Y_{i1}=0,Y_{i2}=1\middle|\,Y_{i1}+Y_{i2}=1\right)\\ &= \mathbb{E}\left[\Pr\left(Y_{i1}=0,Y_{i2}=1\middle|\,Y_{i1}+Y_{i2}=1,A_i\right)\right]\\ &= F\left(\gamma_0\right),\end{aligned}$$

and hence that

$$\hat{\gamma}\xrightarrow{p}2F^{-1}\left(F\left(\gamma_0\right)\right)=2\gamma_0.$$

The joint fixed MLE of the common parameter γ is inconsistent. This is another example of incidental parameters bias.

Conditional fixed effects maximum likelihood

One intuition for the poor properties of the joint fixed effects MLEs is that the badly estimated incidental parameters "contaminate" our estimates of the common parameter. Conditional fixed effects approaches search for implications of the model which are invariant to the incidental parameters, but which still contain identifying information about β. For a certain class of models with exponential family structure, such estimates are available as conditional maximum likelihood estimators. The conditional fixed effect logit estimator is one such estimator.

Re-arranging the log-likelihood yields

$$\begin{aligned} L(\beta, \mathbf{A}) &= \Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X}, \mathbf{A}) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{\exp(X_{it}'\beta + A_i)}{1 + \exp(X_{it}'\beta + A_i)} \right]^{Y_{it}} \left[\frac{1}{1 + \exp(X_{it}'\beta + A_i)} \right]^{1 - Y_{it}} \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{\exp(X_{it}'\beta + A_i)^{Y_{it}}}{1 + \exp(X_{it}'\beta + A_i)} \right] \\ &= c(\mathbf{X}; \beta, \mathbf{A}) \exp\left(\left[\sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it} X_{it}' \right] \beta + \left[\sum_{t=1}^{T} Y_{1t} \right] A_1 + \cdots + \left[\sum_{t=1}^{T} Y_{Nt} \right] A_N \right) \\ &= c(\mathbf{X}; \beta, \mathbf{A}) \prod_{i=1}^{N} \exp\left(\sum_{t=1}^{T} Y_{it} X_{it}' \beta + \left[\sum_{t=1}^{T} Y_{it} \right] A_i \right). \end{aligned}$$

Hence $\sum_{t=1}^{T} Y_{it}$ is a sufficient statistic for A_i and we have that

$$\begin{aligned} \Pr\left(Y_i = y | X_i, A_i, \sum_{t=1}^{T} Y_{it} \right) &= \frac{\exp\left(\sum_{t=1}^{T} Y_{it} X_{it}' \beta \right)}{\sum_{v \in \mathbb{V}_i} \exp\left(\sum_{t=1}^{T} v_t X_{it}' \beta \right)} \\ &= \Pr\left(Y_i = y | X_i, \sum_{t=1}^{T} Y_{it} \right) \end{aligned}$$

where

$$\mathbb{V}_i = \left\{ v : v \in \{0,1\}^T, \sum_{t=1}^{T} v_t = \sum_{t=1}^{T} Y_{it} \right\}.$$

Conditional on $\sum_{t=1}^{T} Y_{it}$, the likelihood of the event $Y_{i1} = y_1, \ldots, Y_{iT} = y_T$ does not vary with A_i. Hence we can choose $\hat{\beta}$ to maximize

$$l_N^{\text{cond}}(\beta) = \sum_{i=1}^{N} \left\{ \left[\sum_{t=1}^{T} Y_{it} X_{it}' \right] \beta - \ln \left[\sum_{v \in \mathbb{V}_i} \exp \left(\sum_{t=1}^{T} v_t X_{it}' \beta \right) \right] \right\}.$$

This yields the conditional fixed effects maximum likelihood estimate of β. The main challenge in computing $\hat{\beta}$ involves efficiently evaluating $\sum_{v \in \mathbb{V}_i} \exp\left(\sum_{t=1}^{T} v_t X_{it}' \beta\right)$. Historically this was actually a serious problem. Fortunately at some point it was realized that this denominator could be computed using a recursive algorithm.

Let$S_i = \sum_{t=1}^{T} Y_{it}$be the number of times unit i's outcome equals 1. The set $\mathbb{V}_i$ includes $\binom{T}{S_i}$ elements. This set consists of all possible sequences $Y_{i1}, \ldots, Y_{iT}$ such that $S_i = \sum_{t=1}^{T} Y_{it}$ (think in terms of Bernoulli trials). Next define

$$\begin{aligned} w_i(T, S_i) &= \sum_{v \in \mathbb{V}_i} \exp \left(\sum_{t=1}^{T} v_t X_{it}' \beta \right) \\ &= \sum_{v \in \mathbb{V}_i} \exp \left(\sum_{t=1}^{T-1} v_t X_{it}' \beta \right) \exp\left(v_T X_{iT}' \beta\right) \end{aligned}$$

which equals

$$w_i(T-1, S_i)$$

if $v_T = 0$ and equals

$$w_i(T-1, S_i - 1) \exp\left(X_{iT}' \beta\right)$$

if $v_T = 1$. We therefore have the recursive relationship

$$w_i(T, S_i) = w_i(T-1, S_i) + w_i(T-1, S_i - 1) \exp\left(X_{iT}' \beta\right).$$

The first term above captures the contribution of all sequences in $\mathbb{V}_i$ which end with $v_T = 0$, whereas the second term is the contribution of all those sequences with $v_T = 1$. We set $w_i(T, s) = 0$ if $T < s$ and $w_i(T, 0) = 1$. The denominator of the conditional likelihood can be built up recursively using relationships like the one sketched above.

When $T = 2$ the conditional logit estimator takes a particular simple form. In that case we have

$$\begin{aligned}
\Pr\left(Y_{i1}=1\middle|\, X_i, A_i, Y_{i1}+Y_{i2}=1\right) &= \frac{\Pr\left(Y_{i1}=1, Y_{i2}=0\middle|\, X_i, A_i\right)}{\Pr\left(Y_{i1}=1, Y_{i2}=0\middle|\, X_i, A_i\right)+\Pr\left(Y_{i1}=0, Y_{i2}=1\middle|\, X_i, A_i\right)} \\
&= \frac{\frac{\exp\left(X_{i1}'\beta+A_i\right)}{1+\exp\left(X_{i1}'\beta+A_i\right)}\frac{1}{1+\exp\left(X_{i2}'\beta+A_i\right)}}{\frac{\exp\left(X_{i1}'\beta+A_i\right)}{1+\exp\left(X_{i1}'\beta+A_i\right)}\frac{1}{1+\exp\left(X_{i2}'\beta+A_i\right)}+\frac{1}{1+\exp\left(X_{i1}'\beta+A_i\right)}\frac{\exp\left(X_{i2}'\beta+A_i\right)}{1+\exp\left(X_{i2}'\beta+A_i\right)}} \\
&= \frac{\exp\left(X_{i1}'\beta+A_i\right)}{\exp\left(X_{i1}'\beta+A_i\right)+\exp\left(X_{i2}'\beta+A_i\right)} \\
&= \frac{1}{1+\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}.
\end{aligned}$$

Similar calculations give $\Pr\left(Y_{i1}=0\middle|\, X_i, A_i, Y_{i1}+Y_{i2}=1\right)=\frac{\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}{1+\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}$. Hence, in the $T=2$ case, the conditional fixed effects MLE maximizes

$$l_N^{\text{cond}}(\beta)=2\sum_{i=1}^{N}\left\{\left(1-Y_{i1}\right)Y_{i2}\ln\left[\frac{\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}{1+\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}\right]+Y_{i1}\left(1-Y_{i2}\right)\ln\left[\frac{1}{1+\exp\left(\left(X_{i2}-X_{i1}\right)'\beta\right)}\right]\right\}$$

Integrated likelihood or random effects approaches

An alternative, and indeed quite natural, approach to large numbers of incidental parameters is to assign a prior distribution to them. One can then integrate these parameters out of the likelihood. The resultant criterion function is an integrated, random effects or empirical Bayes likelihood. I will have more to say about this approach in the next lecture. A quick sketch starts with the same choice model

$$Y_{it}=\mathbf{1}\left(X_{it}'\beta+A_i-V_{it}\geq 0\right).$$

However now it is convenient to make a normal assumption on the time-varying shocks

$$V_{it}\middle|\, X_{i1},\ldots,X_{it},A_i\sim\mathcal{N}\left(0,1\right).$$

Our earlier approaches left the joint distribution of A_i and $X_{i1},\ldots,X_{it}$ unrestricted. In (correlated) random effects approaches we will restrict this relationship. Here we will assume that

$$A_i\middle|\, X_{i1},\ldots,X_{iT}\sim\mathcal{N}\left(X_i'\pi,\sigma_A^2\right).$$

Substitution yields

$$Y_{it}=\mathbf{1}\left(X_{it}'\beta+X_i'\pi+A_i^*\geq V_{it}\right)$$

such that

$$\Pr\left(Y_{it}=1|\,X_i,A_i^*\right)=\Phi\left(X_{it}'\beta+X_i'\pi+A_i^*\right).$$

The correlated random effects estimator choose β and $\eta=(\pi',\sigma_A)'$ to maximize

$$l^I\left(Y_i|\,X_i;\beta,\eta\right)=\ln\left[\int\prod_{t=1}^{T}\Phi\left(X_{it}'\beta+X_i'\pi+a\right)^{Y_{it}}\left[1-\Phi\left(X_{it}'\beta+X_i'\pi+a\right)\right]^{1-Y_{it}}\frac{1}{\sigma_A}\phi\left(\frac{a}{\sigma_A}\right)\mathrm{d}a\right].$$

References

Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Económicas*, 27(3), 423 – 458.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47(1), 225 – 238.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78(1), 159 – 168.

James, W. & Stein, C. M. (1961). Estimation with quadratic loss. *Berkeley Symposium on Mathematical Statistics and Probability*, 4(1), 361 – 379.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2), 391 – 413.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1 – 32.

