Integrated likelihoods, Average Partial Effects and Posterior Means

Bryan S. Graham, UC - Berkeley & NBER

March 30, 2021

Let $Y_{it} \in \{0,1\}$ be a binary outcome for unit i in period t. Let X_{it} be a vector of time-varying policy variables of interest to the econometrician and W_{it} a corresponding vector of control variables. We will assume that unit i's period t potential outcome takes the form

$$Y_{it}(x) \stackrel{def}{\equiv} \mathbf{1}(x'\beta_t + A_i + V_{it} \geq 0).$$

If t indexes siblings, then Y_{it} might be an adult outcome for sibling t in family i, X_{it} an indicator for Head Start participation as a small child and W_{it} a vector of pre-Head Start sibling-specific control variables (e.g., birth weight). The latent variable A_i captures unobserved family-specific endowments, while V_{it} captures sibling-specific ones. For empirical context see Garces et al. (2002).

Alternatively Y_{it} might indicate whether farmer i plants hybrid corn in period t, X_{it} might be an indicator of the cost of hybrid adoption (e.g., seed costs), A_i the farmer-specific yield benefit of planing hybrid corn (which may vary due to soil differences across farms, access to irrigation and/or differences in ability/knowledge across farmers), and V_{it} a farmer's time-specific forecast of the benefits of adoption.

Our target estimand is the average structural function (ASF):

$$\begin{aligned} m_t^{\text{ASF}}(x) &= \mathbb{E}[Y_{it}(x)] \\ &= \int \mathbf{1}(x'\beta_t + a + v \geq 0) f_{A,V_t}(a,v)\,\mathrm{d}a\mathrm{d}v. \end{aligned}$$

An average treatment effect (ATE) parameter can be constructed from differences of the ASF evaluated at different values of the policy variable. This particular object was introduced in Chamberlain (1984); but it is, of course, closely related to ATE-type parameters familiar

from the program evaluation literature (cf., Imbens & Wooldridge, 2009). See Blundell & Powell (2003) for additional discussion and references.

Because both A_i and V_{it} may covary with a unit's actual period t input choice, X_{it}, the conditional expectation $\mathbb{E}\left[Y_{it}\middle|\, X_{it}=x\right]$ does not identify $m_t^{\text{ASF}}(x)$; this is selection bias. To identify the ASF we will combine proxy variable assumptions, familiar from the program evaluation literature, with strict exogeneity assumptions, familiar from the panel data literature. Specifically we will assume that

$$V_{it}\middle|\, X_{i1},\ldots,X_{iT},W_{i1},\ldots,W_{iT},A_i \sim \mathcal{N}\left(W_{it}'\gamma_t,1\right). \tag{1}$$

Condition (1) implies that the full sequence of policy variables $X_{i1},\ldots,X_{iT}$ is independent of V_{it} given $W_{i1},\ldots,W_{iT}$ and A_i. We do allow W_{it} and V_{it} to covary so that, for example, birth weight, W_{it}, may covary with the unobserved sibling-specific drivers, V_{it}, of college attendance, Y_{it}. The Gaussian assumption then implies that

$$\Pr\left(Y_{it}=1\middle|\, X_{i1},\ldots,X_{iT},W_{i1},\ldots,W_{iT},A_i\right)=\Phi\left(X_{it}'\beta_t+W_{it}'\gamma_t+A_i\right).$$

We could have just started our analysis with this expression, but the above formulation emphasizes that γ_t has no "causal" or structural significance. Parents may selectively place their children into Head Start. The hope is that conditional on the latent A_i and the observed sibling specific covariates, W_{it}, however, this placement decision is idiosyncratic.

Next we allow for A_i to covary with $X_{i1},\ldots,X_{iT}$ and $W_{i1},\ldots,W_{iT}$:

$$A_i\middle|\, X_{i1},\ldots,X_{iT},W_{i1},\ldots,W_{iT} \sim \mathcal{N}\left(X_i'\delta+W_i'\zeta,\sigma_A^2\right). \tag{2}$$

In the linear panel data model there is no need to explicitly model the conditional distribution of A_i (since projections arguments can be used). This is not possible in the nonlinear setting. While there are various ways to make (2) less parametric – see for example Newey (1994) – some restrictions on the structure of dependence between A_i and $X_{i1},\ldots,X_{iT},W_{i1},\ldots,W_{iT}$ are unavoidable (see Chernozhukov et al. (2013) for some insight as to why this is the case). Equation (2) is a so-called correlated random effects (CRE) specification. It specifies a distribution for the individual-level "parameters" $\{A_i\}_{i=1}^N$, but allows this distribution to depend on observed regressors; hence the "correlated random effects" nomenclature.

Let $A_i^*=A_i-X_i'\delta-W_i'\zeta$ and $V_{it}^*=V_{it}-W_{it}'\gamma_t$; substituting yields

$$Y_{it}=\mathbf{1}\left(X_{it}'\beta_t+W_{it}'\gamma_t+X_i'\delta+W_i'\zeta+A_i^*\geq -V_{it}^*\right)$$

and hence that

$$\Pr\left(Y_{it}=1\middle|\, X_i, W_i, A_i^*\right) = \Phi\left(X_{it}'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^*\right).$$

Let $\beta = (\beta_1', \ldots, \beta_T')'$ and $\eta = (\gamma_1', \ldots, \gamma_T', \delta', \zeta')'$. Integrating out A_i^* yields unit i's contribution to the integrated-likelihood:

$$f\left(Y_i\middle|\, X_i, W_i; \beta, \eta, \sigma_A\right) = \int f\left(Y_i\middle|\, X_i, W_i, A_i^* = a; \beta, \eta\right) \frac{1}{\sigma_A}\phi\left(\frac{a}{\sigma_A}\right) \mathrm{d}a \tag{3}$$

with

$$\begin{aligned} f\left(Y_i\middle|\, X_i, W_i, A_i^*; \beta, \eta\right) = &\prod_{t=1}^{T} \Phi\left(X_{it}'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^*\right)^{Y_{it}} \\ &\times \left[1 - \Phi\left(X_{it}'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^*\right)\right]^{1-Y_{it}} \end{aligned}$$

The integral in (3) is typically computed by Gauss–Hermite quadrature; $\hat{\beta}$ and $\hat{\eta}$ are then chosen to maximize the sample integrated log-likelihood

$$l_N^I\left(\mathbf{Y}\middle|\, \mathbf{X}, \mathbf{W}; \beta, \eta, \sigma_A\right) = \sum_{i=1}^{N} \ln f\left(Y_i\middle|\, X_i, W_i; \beta, \eta, \sigma_A\right).$$

This maximization can be teased out of most statistical software packages (e.g., by using `xtprobit` in STATA).

A less efficient, but somewhat easier, approach to estimation exploits that fact that, by the convolution properties of the normal distribution,

$$\begin{aligned} \Pr\left(Y_{it}=1\middle|\, X_i, W_i\right) &= \mathbb{E}\left[\Pr\left(Y_{it}=1\middle|\, X_i, W_i, A_i^*\right)\middle|\, X_i, W_i\right] \\ &= \Phi\left(\frac{X_{it}'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta}{\sqrt{1+\sigma_A^2}}\right) \\ &= \Phi\left(X_{it}'\beta_t^* + W_{it}'\gamma_t^* + X_i'\delta^* + W_i'\zeta^*\right) \end{aligned}$$

with $\beta_t^* = \beta_t/\sqrt{1+\sigma_A^2}$ and so on. Next form the composite log-likelihood

$$\begin{aligned} l_N^C\left(\mathbf{Y}\middle|\, \mathbf{X}, \mathbf{W}; \beta, \eta\right) = &\sum_{i=1}^{N}\sum_{t=1}^{T} \left[Y_{it} \ln \Phi\left(X_{it}'\beta_t^* + W_{it}'\gamma_t^* + X_i'\delta^* + W_i'\zeta^*\right)\right. \\ &\left. + \left(1 - Y_{it}\right) \ln\left[1 - \Phi\left(X_{it}'\beta_t^* + W_{it}'\gamma_t^* + X_i'\delta^* + W_i'\zeta^*\right)\right]\right] \end{aligned}$$

This is not a true log-likelihood since it fails to account for the dependence across summands belonging to the same cross section unit. Nevertheless maximizing it, which can be done using a basic cross-section `probit` command, will generate consistent estimates of the rescaled parameters β^* and η^*. Accurate inference on these parameters can be conducted using "clustered" standard errors.

Estimation of the average structural function (ASF)

Starting with our initial definition of the ASF we have

$$\begin{aligned} m_t^{\text{ASF}}(x) &= \mathbb{E}\left[\mathbf{1}\left(x'\beta_t + A_i + V_{it} \geq 0\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left(x'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^* \geq -V_{it}^*\right)\right] \\ &= \mathbb{E}\left[\Phi\left(x'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^*\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\Phi\left(x'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta + A_i^*\right)\middle| X_{i1},\ldots,X_{iT},W_{i1},\ldots,W_{iT}\right]\right] \\ &= \mathbb{E}\left[\Phi\left(\frac{x'\beta_t + W_{it}'\gamma_t + X_i'\delta + W_i'\zeta}{\sqrt{1+\sigma_A^2}}\right)\right] \\ &= \mathbb{E}\left[\Phi\left(x'\beta_t^* + W_{it}'\gamma_t^* + X_i'\delta^* + W_i'\zeta^*\right)\right]. \end{aligned}$$

The expressions to the right of the last two equalities are identified. Hence the ASF may be estimated by either

$$\hat{m}_t^{\text{ASF}}(x) = \frac{1}{N}\sum_{i=1}^N \Phi\left(\frac{x'\hat{\beta}_t + W_{it}'\hat{\gamma}_t + X_i'\hat{\delta} + W_i'\hat{\zeta}}{\sqrt{1+\hat{\sigma}_A^2}}\right)$$

if the integrated log-likelihood estimator is used, or

$$\hat{m}_t^{\text{ASF}}(x) = \frac{1}{N}\sum_{i=1}^N \Phi\left(x'\hat{\beta}_t^* + W_{it}'\hat{\gamma}_t^* + X_i'\hat{\delta}^* + W_i'\hat{\zeta}^*\right)$$

if the composite log-likelihood estimator is used instead.

In both cases inference on the ASF can be conducted by formulating the problem as an exercise in sequential GMM. There are two sources of sampling uncertainty: (i) that due to uncertainty about the population distribution of X_i, W_i and (ii) that due to uncertainty about the form the conditional probability function $\Pr\left(Y_{it}=1\middle| X_i, W_i\right)$. If you wish to ignore the first source of uncertainty – see Imbens (2004) for a related discussion – then simply apply the delta method to $\hat{m}_t^{\text{ASF}}(x)$.

Estimation of unit-specific effects

In recent years researchers have shown an interest in forming estimates of unit-specific objects in panel data settings. Perhaps the most famous, and also controversial, such application is the construction of so called teacher valued added measures (VAMs). Since only a finite number of observations is available per cross section unit, consistent estimation of such effects is not possible. However, it is possible to construct a posterior distribution (and hence a posterior expected value) for such effects. This approach requires viewing the correlated random effects distribution (2) as a prior and the integrated likelihood as an empirical Bayes likelihood. We can then compute the posterior mean of, say, A_i^*, as

$$\begin{aligned}\tilde{A}_i^* &= \mathbb{E}\left[A_i^* \middle| \mathbf{W}, \mathbf{X}, \mathbf{Y}; \hat{\beta}, \hat{\eta}, \hat{\sigma}_A\right] \\ &= \frac{\int a f\left(Y_i \middle| X_i, W_i, A_i^* = a; \hat{\beta}, \hat{\eta}\right) \frac{1}{\hat{\sigma}_A}\phi\left(\frac{a}{\hat{\sigma}_A}\right) \mathrm{d}a}{f\left(Y_i \middle| X_i, W_i; \hat{\beta}, \hat{\eta}, \hat{\sigma}_A\right)}.\end{aligned}$$

We can measure the precision of our beliefs by our posterior variance, computed as

$$\mathbb{V}\left(A_i^* \middle| \mathbf{W}, \mathbf{X}, \mathbf{Y}; \hat{\beta}, \hat{\eta}, \hat{\sigma}_A\right) = \frac{\int \left(a - \tilde{A}_i^*\right)^2 f\left(Y_i \middle| X_i, W_i, A_i^* = a; \hat{\beta}, \hat{\eta}\right) \frac{1}{\hat{\sigma}_A}\phi\left(\frac{a}{\hat{\sigma}_A}\right) \mathrm{d}a}{f\left(Y_i \middle| X_i, W_i; \hat{\beta}, \hat{\eta}, \hat{\sigma}_A\right)}.$$

In practice researchers sometimes proceed somewhat differently. First, joint fixed effects maximum likelihood estimates of the common *and* incidental parameters are computed. Second, the incidental parameters are "shrunk" towards a common location using James-Stein type procedures. I prefer the approach based on the integrated/empirical Bayes likelihood with explicit posterior computations.

To further clarify some of the issues involved consider the linear model

$$Y_{it} = X_{it}'\beta + A_i + V_{it}$$

with

$$V_{it} \middle| X_{i1}, \ldots, X_{iT}, A_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

and

$$A_i \middle| X_{i1}, \ldots, X_{it} \sim \mathcal{N}\left(X_i'\delta, \sigma_A^2\right). \tag{4}$$

Assume, for expositional purposes, that β, δ, σ_A^2 and σ^2 are known. In this setting we have

a simple "normal learning model" with $t = 1, \ldots, T$ noisy signals on the target A_i of

$$Y_{it} - X_{it}'\beta = A_i + V_{it}$$

and a prior on A_i of (4). After T periods our posterior mean for $A_i$is

$$\begin{aligned}\tilde{A}_i^* &= \mathbb{E}\left[A_i^* \middle| \mathbf{W}, \mathbf{X}, \mathbf{Y}; \beta, \delta, \sigma_A^2, \sigma^2\right] \\ &= \frac{\frac{1}{\sigma_A^2}}{\frac{1}{\sigma_A^2} + \frac{T}{\sigma^2}}\left[X_i'\delta\right] + \frac{\frac{T}{\sigma^2}}{\frac{1}{\sigma_A^2} + \frac{T}{\sigma^2}}\hat{A}_i\left(\beta\right).\end{aligned}$$

Note that $\hat{A}_i\left(\beta\right) = \frac{1}{T}\sum_{t=1}^{T}\left(Y_{it} - X_{it}'\beta\right)$ is the joint maximum likelihood fixed effects estimate of A_i given β. The posterior mean is a precision weighted average of this estimate and the prior mean. In practice β, δ, σ_A^2 and σ^2 are unknown and replaced with their random effects MLEs.

References

Blundell, R. & Powell, J. L. (2003). *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 2, chapter Endogeneity in nonparametric and semiparametric regression models, (pp. 312 – 357). Cambridge University Press.

Chamberlain, G. (1984). *Handbook of Econometrics*, volume 2, chapter Panel Data, (pp. 1247 – 1318). North-Holland: Amsterdam.

Chernozhukov, V., Fernández-Val, I., Hahn, J., & Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2), 535 – 580.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of head start. *American Economic Review*, 92(4), 999 – 1012.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1), 4 – 29.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5 – 86.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6), 1349 – 1382.

