MAXIMUM LIKELIHOOD ESTIMATION

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction and Examples
2. Consistency of MLE
3. Asymptotic Distribution
4. MLE Testing
5. Two-Step MLE
6. MLE for Panel Data

1. INTRODUCTION AND EXAMPLES

• Let $(\mathbf{x}_i, \mathbf{y}_i)$ denote a random draw from a population, where $\mathbf{x}_i$ and $\mathbf{y}_i$ can both be vectors. Suppose we are interested in the distribution of $\mathbf{y}_i$ conditional on $\mathbf{x}_i$, $D(\mathbf{y}_i|\mathbf{x}_i)$. If we use a parametric model for a density describing $D(\mathbf{y}_i|\mathbf{x}_i)$, maximimum likelihood estimation is natural.

• Some use the label *conditional maximum likelihood estimation (CMLE)* because we are not modeling the distribution of $\mathbf{x}_i$. But CMLE also has a special meaning in certain panel data contexts.

• For panel data applications, where we have $\{(\mathbf{x}_{it}, \mathbf{y}_{it}) : t = 1, \ldots, T\}$, we can distinguish between modeling the joint distribution of $(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ and just modeling the distributions $D(\mathbf{y}_{it}|\mathbf{x}_{it})$ for each t. (These are sometimes called the "marginal" distributions, but they are really "conditional marginals." The main point is that we are not modeling the joing dependence across time, which can be difficult.)

• As before, with panel data we will use a "small T" setting, so that asymptotic arguments are with independent, identically distributed data.

• First just start with the case where have a density for $D(\mathbf{y}_i|\mathbf{x}_i = \mathbf{x})$ for other possible outcomes $\mathbf{x}$. This fully describes the stochastic behavior of $\mathbf{y}_i$. Contrast with NLS, where we model just the conditional mean. Or LAD, where we model the conditional median. We can recover all of these features, and more, by modeling the entire distribution.

• Let $\boldsymbol{\Theta}$ be the parameter space, as before, as subset of $\mathbb{R}^P$. The model of the density is denoted

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}),\ \mathbf{y} \in \mathcal{Y},\ \mathbf{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

where we assume this is a proper density for each $\boldsymbol{\theta}$. (That is, it is nonnegative and "integrates" to one.)

• We want to allow $\mathbf{y}_i$ to have any characteristic: continuous, discrete, both features. A density can be defined on all cases of interest.

EXAMPLE (Probit): Suppose y_i is a scalar binary response, so it takes just two values, zero and one. Let $\mathbf{x}_i$ be a $1 \times K$ vector with $x_{i1} = 1$ for simplicity. Suppose y_i is generated by a linear latent variable model:

$$\begin{aligned} y_i^* &= \mathbf{x}_i\boldsymbol{\theta} + e_i \\ & \quad e_i|\mathbf{x}_i \sim Normal(0,1) \\ y_i &= 1 \text{ if } y_i^* > 0 \\ &= 0 \text{ if } y_i^* \leq 0. \end{aligned}$$

• A useful shorthand is

$$y_i = 1[y_i^* > 0]$$

where $1[\cdot]$ is the *indicator function*, equal to one if the statement in brackets is true, zero otherwise.

• The data we observe are $(\mathbf{x}_i, y_i)$, and we are usually interesting in the effects of $\mathbf{x}_i$ on y_i.

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(y_i^* > 0|\mathbf{x}_i) = P(\mathbf{x}_i\boldsymbol{\theta} + e_i > 0|\mathbf{x}_i) \\ &= P(e_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = 1 - \Phi(-\mathbf{x}_i\boldsymbol{\theta}) = \boldsymbol{\Phi}(\mathbf{x}_i\boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\Phi}(\mathbf{z}) = \int_{-\infty}^{z} \phi(v)dv$ is the standard normal cdf and

$\phi(v) = (2\pi)^{-1/2} \exp(-v^2/2)$ is the standard normal pdf.

- We have now completely characterized the conditional distribution:

$$f(1|\mathbf{x};\boldsymbol{\theta}) = \Phi(\mathbf{x}\boldsymbol{\theta})$$
$$f(0|\mathbf{x};\boldsymbol{\theta}) = 1 - \Phi(\mathbf{x}\boldsymbol{\theta})$$

or

$$f(y|\mathbf{x};\boldsymbol{\theta}) = [1 - \Phi(\mathbf{x}\boldsymbol{\theta})]^{(1-y)}\Phi(\mathbf{x}\boldsymbol{\theta})^{y}, y = 0, 1$$
$$= 0 \text{ if } y \notin \{0, 1\}.$$

EXAMPLE (Poisson): Let $\mu(\mathbf{x}) = E(y|\mathbf{x})$ where $y \in \{0, 1, 2, \ldots\}$ and $\mu(\cdot) > 0$. Then the conditional distribution is Poisson if the density is

$$f(y|\mathbf{x}) = \exp[-\mu(\mathbf{x})][\mu(\mathbf{x})]^y/y!$$

where $y! = 1 \cdot 2 \cdot \cdots \cdot (y-1) \cdot y$ and $0! = 1$. (The distribution is entirely characterized by the mean.) If $m(\mathbf{x}, \boldsymbol{\theta})$ is the model of the mean, then the model of the density is

$$f(y|\mathbf{x}; \boldsymbol{\theta}) = \exp[-m(\mathbf{x}, \boldsymbol{\theta})][m(\mathbf{x}, \boldsymbol{\theta})]^y/y!$$

or, with $m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta})$,

$$f(y|\mathbf{x}; \boldsymbol{\theta}) = \exp[-\exp(\mathbf{x}\boldsymbol{\theta})]\exp(y\mathbf{x}\boldsymbol{\theta})/y!$$

• The Poisson distribution turns out to have a nice robustness property, that we will use later. Namely, even if the Poisson distribution is incorrect, mean parameters will be consistently estimated if the mean is correctly specified. This leads us to *quasi-maximum likelihood estimation (QMLE)* for the conditional mean.

EXAMPLE (Normal) Let $m(\mathbf{x}, \boldsymbol{\theta})$ be the mean function and $v(\mathbf{x}, \boldsymbol{\theta})$ the variance function. Then

$$f(y|\mathbf{x}; \boldsymbol{\theta}) = [2\pi v(\mathbf{x}, \boldsymbol{\theta})]^{-1/2} \exp\{-[y - m(\mathbf{x}, \boldsymbol{\theta})]^2/[2v(\mathbf{x}, \boldsymbol{\theta})]\}, \ -\infty < y < \infty$$

• Often the mean and variance parameters are entire separate, but not always.

- As it turns out, like the Poisson distribution for estimating parameters of a mean, the normal distribution identifies the first two moments even if the distribution is not normal. That leads us to the notion of QMLE for the conditional mean and conditional variance jointly. For now we assume the density is correctly specified in its entirety.
- If the variance as taken as constant, say σ^2, then the Gaussian (normal) MLE is the NLS estimator, and we know it is consistent if just the conditional mean is correctly specified. Therefore, we already know that the Gaussian MLE is robust to certain kinds of distributional misspecification.

2. CONSISTENCY OF MLE

• The motivation for MLE in introductory statistics is intuitively appealing, but it does not directly lead to a verification of consistency. In fact, we will apply the M-estimation results to the objective function

$$q(\mathbf{w}_i, \boldsymbol{\theta}) = -\log f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta})$$

• $\ell_i(\boldsymbol{\theta}) \equiv \log f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$ called the *log-likelihood function for observation i*. It is random because it depends on $(\mathbf{x}_i, \mathbf{y}_i)$, but we are interested in it as a function of $\boldsymbol{\theta}$.

• Again, for the theory we should be careful about using $\boldsymbol{\theta}_o$ for the "true value" (but drop later for applications).

- So $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_o)$ is the true density of $\mathbf{y}_i$ given $\mathbf{x}_i = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.
- The (Conditional) Maximum Likelihood Estimator of $\boldsymbol{\theta}_o$, $\hat{\boldsymbol{\theta}}$

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}).$$

- Note that this is the starting point. The key is to show that the log likelihood identifies $\boldsymbol{\theta}_o$. This follows by the *Kullback-Leibler Information Inequality*. For our purposes, it implies that

$$E[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_o)|\mathbf{x}_i] \geq E[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})|\mathbf{x}_i], \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

and so

$$E[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_o)] \geq E[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})], \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

• For identification, we assume $\boldsymbol{\theta}_o$ is the unique solution. The key is that it is always a solution. Then, we have to rely on the functional form of the density and the distribution of $\mathbf{x}_i$ to ensure uniqueness. (For example, in the probit and Poisson cases, perfect collinearity in $\mathbf{x}_i$ will violated identification, just as in linear regression.)

• Provided $\ell_i(\boldsymbol{\theta}) \equiv \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and that enough moments of the log likelihood are bouned across $\boldsymbol{\theta}$, the MLE is generally consistent. Just apply the M-estimation consistency result directly.

• Discontinuity in the parameter is rare, but happens if, say, the support of the distribution depends on $\boldsymbol{\theta}$ (as in the case of a $Uniform[0,\theta]$ distribution). The MLE is often consistent, but other arguments are needed. And the large-sample inference changes. In the case of a $Uniform[0,\theta]$ model, $\hat{\theta} = \max(y_1, \ldots y_N)$ does not have a standard normal limiting distribution when properly standardized. The same is true of models of auctions where the unknown parameters include the support of the price offer distribution.

3. ASYMPTOTIC DISTRIBUTION

- Denote the score of the log likelihood as the $P \times 1$ vector

$$\mathbf{s}_i(\boldsymbol{\theta}) = \mathbf{s}(\mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta})' = \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})'$$

Further, the Hessian is still the Jacobian of the score:

$$\mathbf{H}_i(\boldsymbol{\theta}) = \mathbf{H}(\mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta})$$

- A slight notational change from M-estimation:

$$\mathbf{A}_o = -E[\mathbf{H}_i(\boldsymbol{\theta}_o)]$$

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = -E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i]$$

so that $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)$ is postive semi-definite and $\mathbf{A}_o$ is pd.

- As before, let

$$\mathbf{B}_o = E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'].$$

- It is useful to verify certain features of the score. First, because $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})|\mathbf{x}_i],$$

the score generally satisfies

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{0}$$

and so

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)] = \mathbf{0}.$$

• This condition is sometimes referred to as *Fisher consistency*: the MLE solves a population maximization problem, and then we use the sample analog.

• Further, under regularity conditions, the *conditional information matrix equality (CIME)* holds. Namely,

$$-E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i]$$

which implies the *unconditional information matrix equality (UIME)*:

$$-E[\mathbf{H}_i(\boldsymbol{\theta}_o)] = E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'].$$

• In the notation of M-estimation,

$$\mathbf{A}_o = \mathbf{B}_o.$$

• Therefore, for correctly specified (conditional) maximum likelihood problems,

$$Avar[\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] \;=\; \mathbf{A}_o^{-1} \;=\; \mathbf{B}_o^{-1}.$$

• So, generally, one chooses among three estimators of $Avar(\hat{\boldsymbol{\theta}})$:

$$\left(\sum_{i=1}^{N} -\mathbf{H}_i(\hat{\boldsymbol{\theta}})\right)^{-1}, \left(\sum_{i=1}^{N} \mathbf{A}_i(\hat{\boldsymbol{\theta}})\right)^{-1}, \left(\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'\right)^{-1}.$$

- The outer product of the score formulation, while computationally simple, can have severe finite-sample bias; usually the standard errrors are too small on average.
- The Hessian and expected Hessian forms tend to work well. In leading cases, the expected Hessian form depends only on first derivatives.
- If we entertain the possibility that the conditional density is misspecified, then a sandwich form from M-estimation is required. In econometrics, this notion was popularized by White (1982, *Econometrica*).

- EXAMPLE (Probit): The log likelihood for a random draw i is

$$\begin{aligned}
\ell_i(\boldsymbol{\theta}) &= (1-y_i)\log[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})] \; + y_i \log[\Phi(\mathbf{x}_i\boldsymbol{\theta})] \\
\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}) &= -(1-y_i)\mathbf{x}_i\phi(\mathbf{x}_i\boldsymbol{\theta})/[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})] \\
&\quad + y_i\mathbf{x}_i\phi(\mathbf{x}_i\boldsymbol{\theta})/\Phi(\mathbf{x}_i\boldsymbol{\theta}) \\
&= \phi(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i \frac{-(1-y_i)\Phi(\mathbf{x}_i\boldsymbol{\theta}) + y_i[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]} \\
&= \phi(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i \frac{[y_i-\Phi(\mathbf{x}_i\boldsymbol{\theta})]}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]}
\end{aligned}$$

• Therefore, the score is

$$\mathbf{s}_i(\boldsymbol{\theta}) = \phi(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i' \frac{[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta})]}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta})]}$$

and

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \phi(\mathbf{x}_i\boldsymbol{\theta}_o)\mathbf{x}_i' \frac{[E(y_i|\mathbf{x}_i) - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]} = \mathbf{0}$$

because $E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$.

• Note that $E[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{0}$ whenever $E(y_i|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$, even if y_i is not binary. For example, y_i could be a fractional response. (More later with quasi-MLE.)

- The Hessian has the form

$$\mathbf{H}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta}) = \frac{-[\phi(\mathbf{x}_i\boldsymbol{\theta})]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]} + \mathbf{L}(\mathbf{x}_i,\boldsymbol{\theta})[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta})]$$

where $\mathbf{L}(\mathbf{x}_i,\boldsymbol{\theta})$ is the Jacobian of

$$\frac{\phi(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i'}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]}.$$

Under correct specification, we can use

$$\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_o) = -E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \frac{[\phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1-\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]}$$

- Then,

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} \frac{[\phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]^2 \mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1 - \Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]} \xrightarrow{p} \mathbf{A}_o.$$

So the "usual" asymptotic variance estimator is

$$\left(\sum_{i=1}^{N} \frac{[\phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]^2 \mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1 - \Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]} \right)^{-1},$$

which is easily seen to be positive definite when the inverse exists.

• We can verify the conditional information matrix equality. Write $u_i(\theta) = y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta})$ and $u_i = y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$. Then

$$\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})' = [\phi(\mathbf{x}_i\boldsymbol{\theta})]^2\mathbf{x}_i'\mathbf{x}_i\frac{[u_i(\theta)]^2}{\{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1-\Phi(\mathbf{x}_i\boldsymbol{\theta})]\}^2}$$

so

$$\begin{aligned} E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] &= [\phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2\mathbf{x}_i'\mathbf{x}_i\frac{E(u_i^2|\mathbf{x}_i)}{\{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1-\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]\}^2} \\ &= [\phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2\mathbf{x}_i'\mathbf{x}_i\frac{Var(y_i|\mathbf{x}_i)}{\{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1-\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]\}^2} \end{aligned}$$

assuming $E(y_i|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$.

- For a binary response y_i, $Var(y_i|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$, and so

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] = \frac{[\phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]} = -E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i],$$

which is easily seen to be psd for any $\mathbf{x}_i$.

• EXAMPLE (Poisson Regression): The log likelihood for random draw i is

$$\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = -\exp(\mathbf{x}_i\boldsymbol{\theta}) + y_i\mathbf{x}_i\boldsymbol{\theta} - \log(y_i!)$$

and we drop the term $\log(y_i!)$ for computational purposes. (When comparing different models for $D(y_i|\mathbf{x}_i)$ based on the value of the log-likelihood function, this term should be added back in.)

• The score is

$$\begin{aligned}\mathbf{s}_i(\boldsymbol{\theta}) &= -\mathbf{x}_i'\exp(\mathbf{x}_i\boldsymbol{\theta}) + y_i\mathbf{x}_i' \\ &= \mathbf{x}_i'[y_i - \exp(\mathbf{x}_i\boldsymbol{\theta})]\end{aligned}$$

• Therefore,

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{x}_i'[E(y_i|\mathbf{x}_i) - \exp(\mathbf{x}_i\boldsymbol{\theta}_o)] = \mathbf{0}$$

because $E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\theta}_o)$.

• For future reference, note how Fisher consistency holds whenever $E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\theta}_o)$, that is, when the mean is correctly specified. Nothing else about the Poisson distribution needs to be correct.

• The Hessian is particularly easy to compute:

$$\mathbf{H}_i(\boldsymbol{\theta}) - \exp(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i'\mathbf{x}_i,$$

and there is no difference between the Hessian and the expected Hessian given $\mathbf{x}_i$.

• Under the Poisson distributional assumption, the estimated asymptotic variance is

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \exp(\mathbf{x}_i \hat{\boldsymbol{\theta}}) \mathbf{x}_i' \mathbf{x}_i \right)^{-1}.$$

We can verify the CIME:

$$\begin{aligned} E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] &= E(u_i^2|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i \\ &= \exp(\mathbf{x}_i\boldsymbol{\theta}_o)\mathbf{x}_i'\mathbf{x}_i \end{aligned}$$

where $u_i = y_i - \exp(\mathbf{x}_i\boldsymbol{\theta}_o)$.

• All that is used from the Poisson distribution is $Var(y_i|\mathbf{x}_i) = E(y_i|\mathbf{x}_i)$.

• If we allow a general mean function, the formulas are more complicated. But the expected Hessian still has a simple form:

$$-E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \frac{\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)}{m(\mathbf{x}_i, \boldsymbol{\theta}_o)}$$

and a valid estimated asymptotic variance, under the Poisson variance assumption, is

$$\left(\sum_{i=1}^{N} \frac{\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})} \right)^{-1}$$

• Looks like the expression for weighted least squares under correct variance specification when $h(\mathbf{x}_i, \boldsymbol{\gamma}) = m(\mathbf{x}_i, \boldsymbol{\theta})$.

• Aside: When $m(\mathbf{x}_i, \boldsymbol{\theta}) = \exp(\mathbf{x}_i\boldsymbol{\theta})$, A WNLS approach, with $h(\mathbf{x}_i, \boldsymbol{\gamma}) = \exp(\mathbf{x}_i\boldsymbol{\gamma})$ adds more flexibility, as the mean and variance need not be equal. And, can allow for fully robust inference. Can obtain fully robust inference for Poisson regression, too. (This topic is covered explicitly in Chapter 18.)

4. MLE TESTING

- We can apply the same three statistics, Wald, LM (score), and LR statistics.
- Under correct specification of the entire distribution, we do not need a fully robust statistic. The LR statistic is

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) = 2\left[\sum_{i=1}^{N} \ell_i(\hat{\boldsymbol{\theta}}) - \sum_{i=1}^{N} \ell_i(\tilde{\boldsymbol{\theta}})\right]$$

where $\hat{\boldsymbol{\theta}}$ is the unrestricted estimator and $\tilde{\boldsymbol{\theta}}$ is the estimator with Q smooth restrictions imposed. Under H_0,

$$LR \xrightarrow{d} \chi^2_Q.$$

- The Wald statistic has a disadvantage compared with the score and LR statistics in this (and other) nonlinear context: the Wald statistic is not invariant to how the model is parameterized.
- An asymptotic t statistic,

$$t = \frac{(\hat{\theta}_j - a_j)}{se(\hat{\theta}_j)}.$$

is a Wald statistic. Suppose we want to test $H_0 : \theta_o = 1$ against $H_1 : \theta_o > 0$. We can use

$$\frac{(\hat{\theta} - 1)}{se(\hat{\theta})}$$

• Can also define $\gamma = \log(\theta)$ and test $H_0 : \gamma_o = 0$. Then $\hat{\gamma} = \log(\hat{\theta})$.

By the delta method, $\sqrt{N}\,[\log(\hat{\theta}) - \log(\theta_o)] = \theta_o^{-1}\sqrt{N}\,(\hat{\theta} - \theta_o) + o_p(1)$,

or

$$\sqrt{N}\,(\hat{\gamma} - \gamma_o) = \theta_o^{-1}\sqrt{N}\,(\hat{\theta} - \theta_o) + o_p(1).$$

• It follows that

$$Avar(\hat{\gamma}) = Avar(\hat{\theta})/\theta_o^2$$

so

$$se(\hat{\gamma}) = se(\hat{\theta})/\hat{\theta}.$$

- An alternative *t* statistic is

$$\frac{\hat{\gamma}}{se(\hat{\gamma})} = \frac{\hat{\theta}\log(\hat{\theta})}{se(\hat{\theta})} \neq \frac{(\hat{\theta}-1)}{se(\hat{\theta})}.$$

- So we have shown that the Wald test is not invariant to nonlinear transformations of the parameters (and corresponding estimators).

- Can show that the score test, if based on the outer product (not a good idea) or the expected Hessian (a better idea) is invariant to reparameterization. The LR statistic is because it only uses the maximized values of the objective functions.
- In the framework of conditional MLE with conditioning variables, we can apply a version of the *parametric bootstrap*. That is, do not resample $\{\mathbf{x}_i : i = 1, \ldots, N\}$. Simply draw from the distribution described by $f(\cdot|\mathbf{x}_i; \hat{\boldsymbol{\theta}})$ to get data $\{(\mathbf{x}_i, \mathbf{y}_i^{(b)}) : i = 1, \ldots, N\}$. Or, first resample indices to obtain $\{\mathbf{x}_i^{(b)} : i = 1, \ldots, N\}$, and then draw from $f(\cdot|\mathbf{x}_i^{(b)}; \hat{\boldsymbol{\theta}})$ to get $\mathbf{y}_i^{(b)}$.

5. TWO-STEP MLES

- Can apply the results for two-step M-estimation directly. Often, both steps are MLEs, but sometimes the first step is relative easy (say, linear regression) and the second step is nonlinear MLE. We will see this with probit and Tobit models when we apply control function methods.

- When the first-stage estimator is a correctly specified CMLE, there is a “surprsing” efficiency result when the the second-stage estimator satisfies certain assumptions.
- The result is that it is actually more efficient to use a first-stage estimator than if we could use the known value of the parameter in the second stage.

• Assume the first-step estimator, $\hat{\boldsymbol{\gamma}}$, comes from a (conditional) maximum likelihood estimation problem (that satisfies the appropriate regularity conditions):

$$\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{i=1}^{N} \log\, h(\mathbf{v}_i|\mathbf{z}_i;\boldsymbol{\gamma}),$$

where $h(\cdot|\mathbf{z};\boldsymbol{\gamma})$ is a model of the density underlying $D(\mathbf{v}_i|\mathbf{z}_i)$; the population value is $\boldsymbol{\gamma}_o$.

• By the information matrix equality and the usual influence function representation for MLE,

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_o) = \{E[\mathbf{d}_i(\boldsymbol{\gamma}_o)\mathbf{d}_i(\boldsymbol{\gamma}_o)']\}^{-1} N^{-1/2} \sum_{i=1}^{N} \mathbf{d}_i(\boldsymbol{\gamma}_o) + o_p(1),$$

where $\mathbf{d}_i(\boldsymbol{\gamma}) \equiv \nabla_\gamma \log h(\mathbf{v}_i|\mathbf{z}_i;\boldsymbol{\gamma})$ is the $J \times 1$ score of the first-step log-likeliood.

• Assume that the second-step estimator, $\hat{\theta}$, is a two-step M-estimator, solving the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} q(\mathbf{v}_i, \mathbf{w}_i, \mathbf{z}_i, \theta, \hat{\gamma}).$$

• This could be an MLE, but it could be, say, nonlinear least squares.

• Now add the key assumption:

$$D(\mathbf{v}_i|\mathbf{w}_i, \mathbf{z}_i) = D(\mathbf{v}_i|\mathbf{z}_i),$$

which is often called a *conditional independence assumption*: conditional on $\mathbf{z}_i$, $\mathbf{v}_i$ and $\mathbf{w}_i$ are independent.

• In many missing data and treatment effect settings, a conditional independence assumption holds. In those cases, $\mathbf{v}_i$ is usually a binary missing data or treatment indicator. It could also be a multiple set of dummy variables indicator different “treatment” intensities.

• Let $\mathbf{s}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) \equiv \nabla_{\boldsymbol{\theta}} q(\mathbf{v}_i,\mathbf{w}_i,\mathbf{z}_i,\boldsymbol{\theta},\boldsymbol{\gamma})'$ be the $P \times 1$ score of the second-step objective function, but only with respect to $\boldsymbol{\theta}$, and let $\mathbf{F}_o = E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma}_o)]$ (a $P \times J$ matrix).

• Under the generalized conditional information matrix equality, because $\mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma})$ is a function of $(\mathbf{v}_i,\mathbf{w}_i,\mathbf{z}_i)$, we have

$$-E[\nabla_{\boldsymbol{\gamma}} \mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma}_o)|\mathbf{w}_i,\mathbf{z}_i] = E[\mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma}_o)\mathbf{d_i}(\boldsymbol{\gamma}_o)'|\mathbf{w}_i,\mathbf{z}_i].$$

• Using iterated expectations, we conclude that $\mathbf{F}_o = -E[\mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma}_o)\mathbf{d_i}(\boldsymbol{\gamma}_o)']$.

- So, we have shown

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = -\mathbf{A}_o^{-1} N^{-1/2} \sum_{i=1}^{N} \{\mathbf{s}_i^o - E(\mathbf{s}_i^o \mathbf{d}_i^{o\prime})[E(\mathbf{d}_i^o \mathbf{d}_i^{o\prime})]^{-1} \mathbf{d}_i^o\} + o_p(1)$$

$$\equiv -\mathbf{A}_o^{-1} N^{-1/2} \sum_{i=1}^{N} \mathbf{g}_i^o + o_p(1),$$

where $\mathbf{A}_o \equiv E[\nabla_\theta \mathbf{s}_i(\boldsymbol{\theta}_o, \boldsymbol{\gamma}_o)]$ is the $P \times P$ Hessian of the objective function with respect to $\boldsymbol{\theta}$, $\mathbf{g}_i^o \equiv \mathbf{s}_i^o - E(\mathbf{s}_i^o \mathbf{d}_i^{o\prime})[E(\mathbf{d}_i^o \mathbf{d}_i^{o\prime})]^{-1} \mathbf{d}_i^o$ are the population residuals from the population system regression of $\mathbf{s}_i^o$ on $\mathbf{d}_i^{o\prime}$, and the "o" superscript denotes evaluation at $\boldsymbol{\theta}_o$ and $\boldsymbol{\gamma}_o$ or just $\boldsymbol{\gamma}_o$.

• Therefore,

$$Avar\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}$$

where

$$\mathbf{D}_o = E(\mathbf{g}_i^o\mathbf{g}_i^{o\prime}) = Var(\mathbf{g}_i^o).$$

If we knew $\boldsymbol{\gamma}_o$ rather than estimating it by CMLE, the asymptotic variance of the estimator, say $\tilde{\boldsymbol{\theta}}$, would be $Avar\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$ where $\mathbf{B}_o = E(\mathbf{s}_i^o\mathbf{s}_i^{o\prime})$ – the usual expected outer product of the score without accounting for the first-step estimation (because there is none).

- But $\mathbf{B}_o - \mathbf{D}_o$ is positive semi-definite, and so the two-step M-estimator is generally more (asymptotically) efficient than the one-step M-estimator that uses knowledge of $\boldsymbol{\gamma}_o$.
- An immediately implication of the improvement in efficiency in estimating $\boldsymbol{\gamma}_o$ is that if we do use $\hat{\boldsymbol{\gamma}}$ but then ignore the estimation in the second stage, our inference will be conservative. In particular, the standard errors computed from $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$ are larger than they could be.

• Let $\hat{\mathbf{g}}_i'$ be the $1 \times P$ residuals from the multivariate regression of $\hat{\mathbf{s}}_i'$ on $\hat{\mathbf{d}}_i'$, $i = 1, \ldots, N$. Then, we obtain

$$\hat{\mathbf{D}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'.$$

and form the sandwich $\hat{\mathbf{A}}^{-1} \hat{\mathbf{D}} \hat{\mathbf{A}}^{-1} / N$ as $\widehat{Avar}(\hat{\boldsymbol{\theta}})$.

6. PARTIAL (POOLED) MLES FOR PANEL DATA

- With a panel data structure, it is often much easier to specify models for $D(\mathbf{y}_{it}|\mathbf{x}_{it})$ than for $D(\mathbf{y}_i|\mathbf{x}_i)$. For one, the latter usually requires a form of strict exogeneity of (some elements of) $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$.
- For each t, let

$$f_t(\mathbf{y}_t|\mathbf{x}_t;\boldsymbol{\theta})$$

be a model for the density of $D(\mathbf{y}_{it}|\mathbf{x}_{it})$. The vector $\boldsymbol{\theta}$ includes all parameters showing up in any time periods. Usually there is some overlap, if not complete overlap.

- We can easily allow $\mathbf{x}_{it}$ to include lagged dependent variables or other non-strictly exogenous variables.

• Assume that, for each t, the density is correctly specified. As usual, let $\boldsymbol{\theta}_o$ denote the actual population value. Then, by the KLII for each t,

$$E[\log f(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta}_o)] \geq E[\log f(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta})], \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

In some cases, $\boldsymbol{\theta}_o$ will not be the unique solution for each t, but only when we pool across t.

• So, assume that $\boldsymbol{\theta}_o$ is the unique solution to

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=1}^{T} E[\log f(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta})]$$

• The *partial log likelihood* (or *pooled log likelihood*) for cross section observation i is

$$\ell_i(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta}) \equiv \sum_{t=1}^{T} \ell_{it}(\boldsymbol{\theta})$$

• The *partial* or *pooled MLE* solves

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T} \log f(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$$

- The PMLE, $\hat{\boldsymbol{\theta}}$, inherits its large sample properties from the score of $\ell_i(\boldsymbol{\theta})$, the partial log likelihood. Generally, though, we need to account for the panel structure.
- Consistency follows immediately from M-estimation results with $q_i(\boldsymbol{\theta}) = -\ell_i(\boldsymbol{\theta})$. Because we are fixing T, random sampling (in the cross section) is the appropriate framework.
- Define the score for observation i as

$$\mathbf{s}_i(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbf{s}_{it}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbf{s}_{it}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \ell_{it}(\boldsymbol{\theta})'$$

• With smoothness (twice continuously differentiable log likelihood) and $\boldsymbol{\theta}_o$ in the interior of $\boldsymbol{\Theta}$, we can apply the standard M-estimation results:

$$Avar[\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] \;=\; \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$$

where

$$\mathbf{A}_o = -\sum_{t=1}^{T} E[\mathbf{H}_{it}(\boldsymbol{\theta}_o)] = \sum_{t=1}^{T} E[\mathbf{A}_{it}(\boldsymbol{\theta}_o)]$$

$$\mathbf{B}_o = E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'] = E\left[\left(\sum_{t=1}^{T} \mathbf{s}_{it}(\boldsymbol{\theta}_o)\right)\left(\sum_{t=1}^{T} \mathbf{s}_{it}(\boldsymbol{\theta}_o)\right)'\right]$$

$$= \sum_{t=1}^{T} E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{it}(\boldsymbol{\theta}_o)'] + \sum_{t=1}^{T}\sum_{r\neq t}^{T} E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{ir}(\boldsymbol{\theta}_o)'].$$

- For each t, the CIME holds, that is,

$$-E[\mathbf{H}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}] = E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{it}(\boldsymbol{\theta}_o)'|\mathbf{x}_{it}].$$

- But $\mathbf{B}_o$ is generally different from $\mathbf{A}_o$ because of the second term in $\mathbf{B}_o$.
- Generally, we need a sandwich estimator of the form

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$$

where

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})' = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})'$$
$$+ N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r\neq t}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{ir}(\hat{\boldsymbol{\theta}})'$$

- $\hat{\mathbf{A}}$ can be one of three estimators:

$$-N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{H}_{it}(\hat{\boldsymbol{\theta}}),\ N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{A}_{it}(\hat{\boldsymbol{\theta}}),\ \text{or}\ N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})'$$

where the last one (not recommended in general) uses the IME for each t.

• Here, $\mathbf{A}_{it}(\boldsymbol{\theta}_o) = -E[\mathbf{H}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}]$, that is, we condition only on $\mathbf{x}_{it}$ at time t.

• For "canned" applications, the fully robust sandwich form requires a "cluster" option to allow unrestricted serial correlation in the scores $\{\mathbf{s}_{it}(\boldsymbol{\theta}_o) : t = 1, \ldots, T\}$. (A "robust" option usually does not allow serial correlation, but only violation of the information matrix equality for each t.)

• If we assume that the score (evaluated at $\boldsymbol{\theta}_o$, of course) is *serially uncorrelated*, that is,

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{ir}(\boldsymbol{\theta}_o)'] = \mathbf{0}, \text{ all } t \neq r,$$

then $\mathbf{B}_o = \mathbf{A}_o$, and we can use

$$\left(-\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{H}_{it}(\hat{\boldsymbol{\theta}})\right)^{-1}, \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{A}_{it}(\hat{\boldsymbol{\theta}})\right)^{-1}, \text{ or } \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})'\right)^{-1}$$

as $\widehat{Avar}(\hat{\boldsymbol{\theta}})$.

• Aside: For those interested in “large” T, note that we get the same formulas whether we also divide by T when computing averages. The challenge with large T is verifying the law of large numbers and central limit theorem when T increases along with N.

• When will the scores be serially uncorrelated? Suppose the model is *dynamically complete in distribution*, that is,

$$D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{x}_{i1}) = D(\mathbf{y}_{it}|\mathbf{x}_{it}), \; t = 1, \ldots, T.$$

In other words, whatever is included in $\mathbf{x}_{it}$ is sufficient to capture dynamics; no further lags are needed.

• Then

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{x}_{i1}] = E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}] = \mathbf{0},$$

where the first equality follows because $\mathbf{s}_{it}(\boldsymbol{\theta}_o)$ is just a function of $(\mathbf{x}_{it}, \mathbf{y}_{it})$.

• Now take $r < t$, so that $\mathbf{s}_{ir}(\boldsymbol{\theta}_o)$ is necessarily a function of $(\mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1})$. It follows that

$$\begin{aligned} & E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)\mathbf{s}_{ir}(\boldsymbol{\theta}_o)' | \mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{x}_{i1}] \\ = \; & E[\mathbf{s}_{it}(\boldsymbol{\theta}_o) | \mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{x}_{i1}]\mathbf{s}_{ir}(\boldsymbol{\theta}_o)' \\ = \; & \mathbf{0} \cdot \mathbf{s}_{ir}(\boldsymbol{\theta}_o)' = \mathbf{0}. \end{aligned}$$

• So the scores are serially uncorrelated, and we can use the statistics reported from pooled MLE in the usual way, even though we have not necessarily modeled a joint distribution.

EXAMPLE (Pooled Bernoulli): Let $G(\mathbf{x}_{it}\boldsymbol{\theta})$ be the model for $P(y_{it} = 1|\mathbf{x}_{it})$ and assume it is correctly specified, with $0 < G(\cdot) < 1$ and $g(\cdot)$ the derivative of $G(\cdot)$. Then

$$\ell_i(\boldsymbol{\theta}) = \sum_{t=1}^{T}\{(1-y_{it})\log[1-G(\mathbf{x}_{it}\boldsymbol{\theta})] + y_{it}\log[G(\mathbf{x}_{it}\boldsymbol{\theta})]\}$$

$$\mathbf{s}_i(\boldsymbol{\theta}) = \sum_{t=1}^{T}\frac{g(\mathbf{x}_{it}\boldsymbol{\theta})\mathbf{x}_{it}'[y_{it}-G(\mathbf{x}_{it}\boldsymbol{\theta})]}{G(\mathbf{x}_{it}\boldsymbol{\theta})[1-G(\mathbf{x}_{it}\boldsymbol{\theta})]} = \sum_{t=1}^{T}\mathbf{s}_{it}(\boldsymbol{\theta})$$

$$\mathbf{A}_{it}(\boldsymbol{\theta}_o) = \frac{[g(\mathbf{x}_{it}\boldsymbol{\theta}_o)]^2\mathbf{x}_{it}'\mathbf{x}_{it}}{G(\mathbf{x}_{it}\boldsymbol{\theta}_o)[1-G(\mathbf{x}_{it}\boldsymbol{\theta}_o)]}$$

$$\mathbf{A}_o = \sum_{t=1}^{T}E[\mathbf{A}_{it}(\boldsymbol{\theta}_o)].$$

• As long as $P(y_{it} = 1|\mathbf{x}_{it}) = G(\mathbf{x}_{it}\boldsymbol{\theta}_o)$, a valid asymptotic variance estimator is

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{A}_{it}(\hat{\boldsymbol{\theta}})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{A}_{it}(\hat{\boldsymbol{\theta}})\right)^{-1}$$

$$\mathbf{A}_{it}(\hat{\boldsymbol{\theta}}) = \frac{[g(\mathbf{x}_{it}\hat{\boldsymbol{\theta}})]^2\mathbf{x}_{it}'\mathbf{x}_{it}}{G(\mathbf{x}_{it}\hat{\boldsymbol{\theta}})[1 - G(\mathbf{x}_{it}\hat{\boldsymbol{\theta}})]}$$

• The middle term accounts for serial correlation. If

$$P(y_{it} = 1|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1}) = P(y_{it} = 1|\mathbf{x}_{it}),$$

the score is serially uncorrelated, and just one of the outer parts of the sandwich can be used.

• For example, suppose $\mathbf{x}_{it} = (\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1})$ then the condition is

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}, \ldots, y_{i1}, \mathbf{z}_{i1}) = P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}),$$

which says that, in addition to controlling for contemporaneous variables $\mathbf{z}_{it}$, one lag $y_{i,t-1}$ and $\mathbf{z}_{i,t-1}$ are sufficient to capture the dynamics.

• As with linear regression, dynamic completeness would be a strong assumption if we take $\mathbf{x}_{it} = \mathbf{z}_{it}$ or even $\mathbf{x}_{it}$ a subset of $(\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, \ldots, \mathbf{z}_{i1})$. Usually, past outcomes on y_{it} help to predict $y_{i,t-1}$, even after conditioning on current and past other variables.