Classical Non-Linear Methods: Maximum Likelihood

Bertel Schjerning

University of Copenhagen, Department of Economics

January 29, 2026

Plan for Classical Non-Linear Methods

Lecture 4: M-estimation, Intro, Non-linear LS (W.12)

Lecture 5: Asymptotic properties of M-estimators (W.12)

Lecture 6: M-estimator inference, Variance estimation (W.12)

Lecture 7: Maximum likelihood estimation (W.13)

Next: Specific non-linear models. (Bertel)

Maximum Likelihood Estimation: Aim

Previously: Modelled feature(s) of distribution $\mathrm{D}(y \mid \mathbf{x})$.

- E.g. $\mathrm{E}[y \mid \mathbf{x}]$ and $\mathrm{var}(y \mid \mathbf{x})$.

Maximum likelihood estimation (MLE) more ambitious.

Model for *entire distribution* $\mathrm{D}(y \mid \mathbf{x})$.

Why MLE? Advantages

Efficiency

- MLE uses entire $\mathrm{D}(y \mid \mathbf{x})$.
- (Correct) Structure $\implies$ Information.

May estimate any feature

- Conditional moments: $\mathrm{E}[y \mid \mathbf{x}]$, $\mathrm{var}(y \mid \mathbf{x})$, ...
- Conditional prob's: $\mathrm{P}(y = 1 \mid \mathbf{x})$, $\mathrm{P}(y \in [a, b] \mid \mathbf{x})$, ...
- Conditional density.
- Derivatives (wrt. $\mathbf{x}$) thereof ...

Why Not MLE? Drawbacks

Structure incorrect $\Longrightarrow$ Non-robustness

- MLE uses entire $\mathrm{D}(y \mid \mathbf{x})\ldots$
- Inconsistent (in general) if misspecified.
  - Exceptions exist...

Outline

Framework

Truth vs. Model

Object of interest: "True" density $p_o(\mathbf{y}|\mathbf{x})$ of $\mathrm{D}(\mathbf{y}_i|\mathbf{x}_i)$.

- Possible values $(\mathbf{y}, \mathbf{x}) \in \mathscr{Y} \times \mathscr{X}$ for $(\mathbf{y}_i, \mathbf{x}_i)$.
- Density understood in broad sense.
  - Discrete and/or continuous elements allowed.
  - Only discrete: Swap integrals for sums.

Parametric model: Family of cond'l densities

$$\mathcal{F} := \{(\mathbf{y}, \mathbf{x}) \mapsto f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$$

with parameter space $\Theta \subseteq \mathbb{R}^P$ (fixed!)

Model Assumptions

Parametric model: $\mathcal{F} = \{(\mathbf{y}, \mathbf{x}) \mapsto f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$.

Assumptions

1. Legitimate densities:

$$f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \geqslant 0, \text{ for } (\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \in \mathscr{Y} \times \mathscr{X} \times \Theta,$$

$$\int_{\mathscr{Y}} f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \nu(\mathrm{d}y) = 1, \text{ for } (\mathbf{x}, \boldsymbol{\theta}) \in \mathscr{X} \times \Theta.$$

2. Correct specification: $p_o(\cdot \mid \cdot) \in \mathcal{F}$.

I.e. for some $\boldsymbol{\theta}_o \in \Theta$,

$$p_o(\mathbf{y} \mid \mathbf{x}) = f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}_o), \text{ for } (\mathbf{y}, \mathbf{x}) \in \mathscr{Y} \times \mathscr{X}.$$

Call $\boldsymbol{\theta}_o$ "true theta."

Identification in Maximum Likelihood Context

Definition: $\boldsymbol{\theta}_o$ identified if and only if for all $\boldsymbol{\theta} \in \Theta \setminus \{\boldsymbol{\theta}_o\}$ s.t.

$$f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \neq f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}_o) \text{ for } \textit{some } (\mathbf{y}, \mathbf{x}) \in \mathscr{Y} \times \mathscr{X}.$$

Conversely: if $\boldsymbol{\theta}_o$ ***not*** identified, then some $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$ yields

$$f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}_o) \text{ for } \textit{all } (\mathbf{y}, \mathbf{x}).$$

- I.e., $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_o$ are observationally equivalent.
- Can't tell if data generated from one or the other.

Discussion

Suppose that for some $(\alpha_o, \mu_o) \in \mathbb{R}^2$,

$$y_i = \alpha_o + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(\mu_o, 1).$$

Q: Are (α_o, μ_o) identifiable?

Estimand

Identification implies: $\boldsymbol{\theta}_o$ uniquely solves population problem

$$\max_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right]. \tag{PP}$$

(To be shown.)

Equivalently, $\boldsymbol{\theta}_o$ (uniquely) solves

$$\min_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[-\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right].$$

Taking $q\left(\mathbf{w}, \boldsymbol{\theta}\right) = -\ln f\left(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}\right) \implies \boldsymbol{\theta}_o$ an M-estimand.

Estimator

Analogy principle suggests sample problem

$$\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ell_i \left(\boldsymbol{\theta} \right), \tag{SP}$$

with (conditional) likelihood contribution

$$\ell_i \left(\boldsymbol{\theta} \right) := \ln f \left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta} \right).$$

Maximum likelihood estimator (MLE): Any solution $\widehat{\boldsymbol{\theta}}_N$ to SP.

- Every MLE an M-estimator!

Example: Probit

Example: Probit

Binary outcome y_i, i.e. $\mathscr{Y} = \{0, 1\}$, has cond'd density

$$p_o(y \mid \mathbf{x}) = p_o(1 \mid \mathbf{x})^y \left[1 - p_o(1 \mid \mathbf{x})\right]^{1-y}, \quad y \in \{0, 1\},$$
$$p_o(1 \mid \mathbf{x}) := \mathrm{P}(y_i = 1 \mid \mathbf{x}_i = \mathbf{x}) \quad \text{("success prob")}$$

Probit model $\{\mathbf{x} \mapsto \Phi(\mathbf{x}\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ for $p_o(1|\mathbf{x})$.

- $\Phi : \mathbb{R} \to (0, 1)$: standard normal CDF.
- $\Theta \subseteq \mathbb{R}^P$.

Correctly specified if for some $\boldsymbol{\theta}_o \in \Theta$,

$$p_o(1 \mid \mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\theta}_o), \text{ all } \mathbf{x}.$$

Example: Probit

Probit model for density of $\mathrm{D}(y_i \mid \mathbf{x}_i)$:

$$f\left(y \mid \mathbf{x}; \boldsymbol{\theta}\right) = \Phi\left(\mathbf{x}\boldsymbol{\theta}\right)^{y} \left[1 - \Phi\left(\mathbf{x}\boldsymbol{\theta}\right)\right]^{1-y}, \quad y \in \{0, 1\}.$$

Probit log-likelihood contribution

$$\ell_i\left(\boldsymbol{\theta}\right) = y_i \ln \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right) + \left(1 - y_i\right) \ln\left[1 - \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right)\right].$$

Probit estimator: Any solution to

$$\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left\{y_i \ln \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right) + \left(1 - y_i\right) \ln\left[1 - \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right)\right]\right\}.$$

Identification and Solution Uniqueness

Identification and Uniqueness

Claim: $\boldsymbol{\theta}_o$ identified $\Rightarrow$ uniquely solves PP,

$$\boldsymbol{\theta}_o = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \, \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right].$$

Will invoke **Jensen's inequality:** g *concave* + Z random

$$\Rightarrow \mathrm{E}\left[g\left(Z\right)\right] \leqslant g\left(\mathrm{E}\left[Z\right]\right).$$

Inequality *strict* provided g *strictly* concave + Z *non-constant*.

Identification and Uniqueness

- Fix $\boldsymbol{\theta} \in \Theta$.
- $g := \ln(\cdot)$
- $Z := f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) / f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_o)$
- Condition on $\mathbf{x}_i$ . . . [BB]

Identification and Uniqueness

Correct specification + legitimate density $\Rightarrow$

$$\mathrm{E}\left[\frac{f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}\right)}{f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}_o\right)}\middle|\mathbf{x}_i\right] = \int_{\mathscr{Y}} \frac{f\left(\mathbf{y} \mid \mathbf{x}_i;\boldsymbol{\theta}\right)}{f\left(\mathbf{y} \mid \mathbf{x}_i;\boldsymbol{\theta}_o\right)} p_o\left(\mathbf{y} \mid \mathbf{x}_i\right) \nu\left(\mathrm{d}\mathbf{y}\right)$$

$$=$$

Hence

$$\mathrm{E}\left[\ln\left(\frac{f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}\right)}{f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}_o\right)}\right)\middle|\mathbf{x}_i\right] \leqslant$$

Rearranging,

$$\mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}_o\right)|\mathbf{x}_i\right] \geqslant \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i;\boldsymbol{\theta}\right)|\mathbf{x}_i\right].$$

Identification and Uniqueness

Have shown: No matter $\mathbf{x}_i$,

$$\mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}_o\right)|\mathbf{x}_i\right] \geqslant \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)|\mathbf{x}_i\right].$$

Taking expectations,

$$\mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}_o\right)\right] \geqslant \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right].$$

$\boldsymbol{\theta} \in \Theta$ arbitrary $\implies$ $\boldsymbol{\theta}_o$ solves PP.

Identification and Uniqueness

Have shown: $\boldsymbol{\theta}_o$ solves PP, i.e.

$$\mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}_o\right)\right] \geqslant \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right] \text{ for all } \boldsymbol{\theta} \in \Theta.$$

$\boldsymbol{\theta}_o$ identified $\implies$ $Z = f\left(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta}\right) / f\left(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta}_o\right)$ *non-constant.*

$\ln\left(\cdot\right)$ *strictly* concave, so Jensen $\implies$

$$\mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}_o\right)\right] > \mathrm{E}\left[\ln f\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right)\right] \text{ for all } \boldsymbol{\theta} \in \Theta \backslash \left\{\boldsymbol{\theta}_o\right\}.$$

Hence, identification implies unique maximizer.

Asymptotic Properties

Asymptotic Properties of MLE

Recall: Every MLE an M-estimator,

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \{ - \ln f \left(\mathbf{y}_i \mid \mathbf{x}_i ; \boldsymbol{\theta} \right) \}.$$

May appeal to general results:

- Consistency (W. Thm. 12.2)
- Asymptotic normality (W. Thm. 12.3).

Will verify/reduce relevant conditions.

Consistency

M-Estimator Consistency with Compactness

Theorem (W. Theorem 12.2)

If

1. $\boldsymbol{\theta}_o$ *uniquely minimizes* $\boldsymbol{\theta} \mapsto \mathrm{E}\left[q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)\right]$ *("identification"),*
2. $\Theta \subseteq \mathbb{R}^P$ *compact* *(i.e. closed + bounded),*
3. $q\left(\mathbf{w}, \cdot\right)$ *continuous (in* $\boldsymbol{\theta}$*),*

(+ technical conditions), then

1. *SP has a solution* $\left(\widehat{\boldsymbol{\theta}}_N\right)$*, and*
2. *any selection* $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$ *of minimizers is consistent for* $\boldsymbol{\theta}_o$*,*

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_o.$$

ML-Estimator Consistency with Compactness

Q: Conditions verified?

1. Unique PP solution? Follows from $\boldsymbol{\theta}_o$ identified,

$$\boldsymbol{\theta} \neq \boldsymbol{\theta}_o \implies f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \neq f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_o) \text{ some } (\mathbf{y},\mathbf{x}).$$

2. Θ compact? (Nothing new...)

3. $q(\mathbf{w},\cdot) = -\ln f(\mathbf{y}|\mathbf{x};\cdot)$ continuous?

   - Follows if $f(\mathbf{y}|\mathbf{x};\cdot)$ continuous (and is $\neq 0$).
   - **Probit:** Φ *is* cont's [and takes values in (0,1)].

(ML) ID'n + compactness + log-L cont' $\implies$ MLE consistency.

M-Estimator Consistency without Compactness

Theorem (Newey and McFadden, 1994)

Let

1. $Q : \mathbb{R}^P \to \mathbb{R}$ *be uniquely minimized at* $\boldsymbol{\theta}_o$;
2. *each (random)* $\{\widehat{Q}_N : \mathbb{R}^P \to \mathbb{R}\}_{N=1}^{\infty}$ *convex; and,*
3. $\widehat{Q}_N(\boldsymbol{\theta}) \to_p Q(\boldsymbol{\theta})$ *for each* $\boldsymbol{\theta} \in \mathbb{R}^P$.

Then

1. *a minimizer* $\widehat{\boldsymbol{\theta}}_N$ *of* $\widehat{Q}_N$ *exists with probability* $\to 1$; *and*
2. *for any minimizer selection,* $\widehat{\boldsymbol{\theta}}_N \to_p \boldsymbol{\theta}_o$.

If $q(\mathbf{w}, \boldsymbol{\theta})$ convex in $\boldsymbol{\theta}$, so is $N^{-1} \sum_i q(\mathbf{w}_i, \boldsymbol{\theta}) [= \widehat{Q}_N(\boldsymbol{\theta})]$.

ML-Estimator Consistency without Compactness

Q: Conditions verified?

1. If $\boldsymbol{\theta}_o$ identified (in ML sense), then $\boldsymbol{\theta}_o$ uniquely minimizes

$$\mathbb{R}^P \ni \boldsymbol{\theta} \mapsto \mathrm{E}[-\ln f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})]. \qquad (= Q(\boldsymbol{\theta}))$$

2. If $-\ln f(\mathbf{y}|\mathbf{x};\cdot)$ convex (in $\boldsymbol{\theta}$), then so is

$$\mathbb{R}^P \ni \boldsymbol{\theta} \mapsto \frac{1}{N}\sum_{i=1}^{N}\{-\ln f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})\}. \qquad (= \widehat{Q}_N(\boldsymbol{\theta}))$$

   - Check $f(\mathbf{y}|\mathbf{x};\cdot)$ *log-concave*.
   - **Probit:** $\Phi(\cdot)$ and $1-\Phi(\cdot)$ *are* log-concave.

3. Now $\widehat{Q}_N(\boldsymbol{\theta}) \to_p Q(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \mathbb{R}^P$ follows from LLN.

(ML) identification + log-concavity $\implies$ MLE consistency.

Asymptotic Normality

Asymptotic Normality of M-Estimators

Theorem (W. Theorem 12.3)

Provided

- $\boldsymbol{\theta}_o$ *unique PP solution + interior to* Θ *compact,*
- $q(\mathbf{w}, \cdot)$ *cont' + twice cont' diff' on* $\operatorname{int}\Theta$*,*
- $\mathrm{E}[\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = \mathbf{0}$*, and* $\mathrm{E}[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)]$ *positive definite,*
- *(+ technical),*

any selection $\{\widehat{\boldsymbol{\theta}}_N\}_{N=1}^{\infty}$ *of minimizers satisfies*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_o) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\right),$$
$$\mathbf{A}_o := \mathrm{E}[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)],$$
$$\mathbf{B}_o := \mathrm{E}\left[\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)\,\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)'\right].$$

Oh, that Pesky Minus...

Thm. 12.3 designed for *min*imization.

Turn max'n into min'n:

$$q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) = -\ln f\left(\mathbf{y}_i \middle| \mathbf{x}_i; \boldsymbol{\theta}\right) = -\ell_i\left(\boldsymbol{\theta}\right).$$

Hence above score $(\mathbf{s})$/Hessian $(\mathbf{H})$ of $-\ln f$ (wrt. $\boldsymbol{\theta}$).

In what follows,

$$\begin{aligned}
\mathbf{s}_i\left(\boldsymbol{\theta}\right) &:= \nabla_\theta \ell_i\left(\boldsymbol{\theta}\right)', && \text{(no minus)} \\
\mathbf{H}_i\left(\boldsymbol{\theta}\right) &:= \nabla_\theta \mathbf{s}_i\left(\boldsymbol{\theta}\right) = \nabla_\theta^2 \ell_i\left(\boldsymbol{\theta}\right), && \text{(no minus)} \\
\Rightarrow \mathbf{A}_o &= -\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right)\right].
\end{aligned}$$

Information Matrix Equalities, I

Additionally assuming...

- $\boldsymbol{\theta}_o$ interior to Θ,
- $\ln f\left(\mathbf{y}|\mathbf{x};\cdot\right)$ twice cont' diff' on int Θ,
- (+ technical)

May now apply Thm. 12.3.

But further structure available...

Information Matrix Equalities, II

Under quite mild (smoothness) conditions,

$$-\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right)\middle|\,\mathbf{x}_i\right] = \mathrm{E}\left[\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)'\middle|\,\mathbf{x}_i\right], \qquad \text{(CIME)}$$

$$\Rightarrow -\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right)\right] = \mathrm{E}\left[\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)'\right]. \qquad \text{(UIME)}$$

(Un)Conditional Information Matrix Equality.

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

Asymptotic variance simplifies for MLE.

Asymptotic Normality of ML-Estimators

Theorem (W. Theorem 13.2)

Provided

- $\boldsymbol{\theta}_o$ *identified + interior to* Θ *compact,*
- $\ln f\left(\mathbf{y}|\mathbf{x};\cdot\right)$ *cont' + twice cont' diff' on* $\operatorname{int}\Theta$*,*
- *+ technical (including CIME justification),*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_o) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{A}_o^{-1}\right),$$
$$\mathbf{A}_o := -\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right)\right] \qquad (= \mathbf{B}_o)$$

Hence $\operatorname{Avar}(\widehat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1}/N$.

Asymptotic Variance Estimation

Asymptotic Variance Estimators

Three candidates for $\widehat{\mathbf{A}}$:

$$= -\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i(\widehat{\boldsymbol{\theta}}), \qquad \text{(least structural)}$$

$$\text{or} \quad = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i(\widehat{\boldsymbol{\theta}})\mathbf{s}_i(\widehat{\boldsymbol{\theta}})', \qquad \text{(per UIME)}$$

$$\text{or} \quad = \frac{1}{N}\sum_{i=1}^{N}\mathbf{A}(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}), \qquad \text{(semi-structural)}$$

$$\text{where} \quad \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) := -\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right)|\mathbf{x}_i\right].$$

Each $\widehat{\mathbf{A}} \to_p \mathbf{A}_o\,(= \mathbf{B}_o)$ under mild (add'l) cond's.

Avar Estimation: Discussion

$$\widehat{\text{Avar}}(\widehat{\boldsymbol{\theta}}) =$$

$$\underbrace{\left(-\sum_{i=1}^{N}\mathbf{H}_i(\widehat{\boldsymbol{\theta}})\right)^{-1}}_{(1)}, \quad \underbrace{\left(\sum_{i=1}^{N}\mathbf{s}_i(\widehat{\boldsymbol{\theta}})\mathbf{s}_i(\widehat{\boldsymbol{\theta}})'\right)^{-1}}_{(2)}, \text{ or } \underbrace{\left(\sum_{i=1}^{N}\mathbf{A}(\mathbf{x}_i,\widehat{\boldsymbol{\theta}})\right)^{-1}}_{(3)} ?$$

Pros/Cons:

1. Always available, 2^{nd}-order diff', p.s.d.
2. Easy to compute, 1^{st}-order diff', p.s.d.
3. Harder to derive, often p.d. + good in small(ish) samples.

Example: Probit Avar Estimation

Probit Avar Estimation

Recall: Cond'l probit density

$$f\left(y \mid \mathbf{x}; \boldsymbol{\theta}\right) = \Phi\left(\mathbf{x}\boldsymbol{\theta}\right)^{y} \left[1 - \Phi\left(\mathbf{x}\boldsymbol{\theta}\right)\right]^{1-y}, \quad y \in \{0, 1\}.$$

Cond'l probit log-likelihood contribution:

$$\ell_i\left(\boldsymbol{\theta}\right) = y_i \ln \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right) + \left(1 - y_i\right) \ln\left[1 - \Phi\left(\mathbf{x}_i\boldsymbol{\theta}\right)\right].$$

We'll illustrate option (3) ("cond'l Hessian"):

1. Derive score, $\mathbf{s}_i\left(\boldsymbol{\theta}\right)$.
2. Derive $\mathbf{A}\left(\mathbf{x}_i, \boldsymbol{\theta}_o\right) = -\mathrm{E}\left[\mathbf{H}_i\left(\boldsymbol{\theta}_o\right) | \mathbf{x}_i\right]$.
3. Sum up + insert $\widehat{\boldsymbol{\theta}}$ + invert.

Probit Avar Estimation

Step 1: Derive score.

Chain rule + gather $\Rightarrow$

$$\mathbf{s}_i(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) = \frac{[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta})]\,\varphi(\mathbf{x}_i\boldsymbol{\theta})}{\Phi(\mathbf{x}_i\boldsymbol{\theta})[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta})]}\mathbf{x}_i'.$$

Step 2: Derive $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)$.

$$\begin{aligned}
\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) &= -\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] \\
&= \mathrm{E}\left[\mathbf{s}_i(\boldsymbol{\theta}_o)\,\mathbf{s}_i(\boldsymbol{\theta}_o)'\middle|\mathbf{x}_i\right]. \qquad \text{(CIME)} \\
&= \mathrm{E}\left[\frac{[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2\,\varphi(\mathbf{x}_i\boldsymbol{\theta}_o)^2}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)^2[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)]^2}\mathbf{x}_i'\mathbf{x}_i\middle|\mathbf{x}_i\right].
\end{aligned}$$

Probit Avar Estimation

Step 2: Derive $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)$ (ctnd).

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = \mathrm{E}\left[\left[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\right]^2 \middle| \mathbf{x}_i\right] \frac{\varphi(\mathbf{x}_i\boldsymbol{\theta}_o)^2}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)^2\left[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\right]^2}\mathbf{x}_i'\mathbf{x}_i.$$

y_i binary + $p_o(1|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)$

$$\Rightarrow \mathrm{E}\left[\left[y_i - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\right]^2 \middle| \mathbf{x}_i\right] = \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\left[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\right].$$

Hence

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = \frac{\varphi(\mathbf{x}_i\boldsymbol{\theta}_o)^2}{\Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\left[1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}_o)\right]}\mathbf{x}_i'\mathbf{x}_i$$

Probit Avar Estimation

Step 3: Sum + insert $\widehat{\boldsymbol{\theta}}$ + invert.

$\Rightarrow$ estimator of probit Avar:

$$\widehat{\text{Avar}}(\widehat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}) \right)^{-1},$$

$$\mathbf{A}(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}) := \frac{\varphi(\mathbf{x}_i \widehat{\boldsymbol{\theta}})^2}{\Phi(\mathbf{x}_i \widehat{\boldsymbol{\theta}})[1 - \Phi(\mathbf{x}_i \widehat{\boldsymbol{\theta}})]} \mathbf{x}_i' \mathbf{x}_i.$$

- $\mathbf{A}(\mathbf{x}_i, \widehat{\boldsymbol{\theta}})$ at least p.s.d.
- Can get same from 2^{nd}-order diff'n. (Check!)