

Classical Non-Linear Methods: Maximum Likelihood

Bertel Schjerning

University of Copenhagen, Department of Economics

February 11, 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 $D(y | \mathbf{x})$.

- ▶ E.g. $E[y | \mathbf{x}]$ and $\text{var}(y | \mathbf{x})$.

Maximum likelihood estimation (MLE) more ambitious.

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

Why MLE? Advantages

Efficiency

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

May estimate any feature

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

Why Not MLE? Drawbacks

Structure incorrect \implies Non-robustness

- ▶ MLE uses entire $D(y | \mathbf{x}) \dots$
- ▶ Inconsistent (in general) if misspecified.
- ▶ Exceptions exist...

Outline

Framework

Example: Probit

Identification and Solution Uniqueness

Asymptotic Properties

Consistency

Asymptotic Normality

Asymptotic Variance Estimation

Example: Probit Avar Estimation

Framework

Truth vs. Model

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

- ▶ Possible values $(\mathbf{y}, \mathbf{x}) \in \mathcal{Y} \times \mathcal{X}$ 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: $\mathcal{F} = \{(\mathbf{y}, \mathbf{x}) \mapsto f(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}) ; \boldsymbol{\theta} \in \Theta\}.$

Assumptions

1. Legitimate densities:

$$f(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}) \geq 0, \text{ for } (\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \in \mathcal{Y} \times \mathcal{X} \times \Theta,$$

$$\int_{\mathcal{Y}} f(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}) v(dy) = 1, \text{ for } (\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta.$$

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

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

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

Call $\boldsymbol{\theta}_o$ “true theta.”

Identification in Maximum Likelihood Context

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

$$f(\mathbf{y} | \mathbf{x}; \theta) \neq f(\mathbf{y} | \mathbf{x}; \theta_o) \text{ for some } (\mathbf{y}, \mathbf{x}) \in \mathcal{Y} \times \mathcal{X}.$$

Conversely: if θ_o not identified, then some $\theta \neq \theta_o$ yields

$$f(\mathbf{y} | \mathbf{x}; \theta) = f(\mathbf{y} | \mathbf{x}; \theta_o) \text{ for all } (\mathbf{y}, \mathbf{x}).$$

- ▶ I.e., θ and θ_o are **observationally equivalent**.
- ▶ Can't tell if data generated from one or the other.

Discussion

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

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

Q: Are (α_o, μ_o) identifiable?

Estimand

Identification implies: θ_o uniquely solves population problem

$$\max_{\theta \in \Theta} E [\ln f (\mathbf{y}_i | \mathbf{x}_i; \theta)] . \quad (\text{PP})$$

(To be shown.)

Equivalently, θ_o (uniquely) solves

$$\min_{\theta \in \Theta} E [-\ln f (\mathbf{y}_i | \mathbf{x}_i; \theta)] .$$

Taking $q(\mathbf{w}, \theta) = -\ln f (\mathbf{y} | \mathbf{x}; \theta) \implies \theta_o$ an M-estimand.

Estimator

Analogy principle suggests sample problem

$$\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N \ell_i(\theta), \quad (\text{SP})$$

with (conditional) likelihood contribution

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

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. $\mathcal{Y} = \{0, 1\}$, has cond'd density

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

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

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

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

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

Example: Probit

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

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

Probit log-likelihood contribution

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

Probit estimator: Any solution to

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

Identification and Solution Uniqueness

Identification and Uniqueness

Claim: θ_o identified \Rightarrow uniquely solves PP,

$$\theta_o = \operatorname{argmax}_{\theta \in \Theta} E [\ln f (\mathbf{y}_i | \mathbf{x}_i; \theta)].$$

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

$$\Rightarrow E [g (Z)] \leq g (E [Z]).$$

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 $\mathbf{x}_i \dots$ [BB]

Identification and Uniqueness

Correct specification + legitimate density \Rightarrow

$$\begin{aligned} \mathrm{E}\left[\frac{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta})}{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}_o)} \middle| \mathbf{x}_i\right] &= \int_{\mathcal{Y}} \frac{f(\mathbf{y} | \mathbf{x}_i; \boldsymbol{\theta})}{f(\mathbf{y} | \mathbf{x}_i; \boldsymbol{\theta}_o)} p_o(\mathbf{y} | \mathbf{x}_i) v(d\mathbf{y}) \\ &= \int_{\mathcal{Y}} f(\mathbf{y} | \mathbf{x}_i; \boldsymbol{\theta}) v(d\mathbf{y}) = 1. \end{aligned}$$

Hence

$$\mathrm{E}\left[\ln\left(\frac{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta})}{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}_o)}\right) \middle| \mathbf{x}_i\right] \leqslant \ln\left(\mathrm{E}\left[\frac{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta})}{f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}_o)} \middle| \mathbf{x}_i\right]\right) = \ln(1) = 0.$$

Rearranging,

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

Identification and Uniqueness

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

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

Taking expectations,

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

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

Identification and Uniqueness

Have shown: θ_o solves PP, i.e.

$$E[\ln f(\mathbf{y}_i | \mathbf{x}_i; \theta_o)] \geq E[\ln f(\mathbf{y}_i | \mathbf{x}_i; \theta)] \text{ for all } \theta \in \Theta.$$

θ_o identified $\implies Z = f(\mathbf{y}_i | \mathbf{x}_i; \theta) / f(\mathbf{y}_i | \mathbf{x}_i; \theta_o)$ non-constant.

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

$$E[\ln f(\mathbf{y}_i | \mathbf{x}_i; \theta_o)] > E[\ln f(\mathbf{y}_i | \mathbf{x}_i; \theta)] \text{ for all } \theta \in \Theta \setminus \{\theta_o\}.$$

Hence, identification implies unique maximizer.

Asymptotic Properties

Asymptotic Properties of MLE

Recall: Every MLE an M-estimator,

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

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. θ_o uniquely minimizes $\theta \mapsto E[q(\mathbf{w}_i, \theta)]$ (“identification”),
2. $\Theta \subseteq \mathbb{R}^P$ compact (i.e. closed + bounded),
3. $q(\mathbf{w}, \cdot)$ continuous (in θ),

(+ technical conditions), then

1. SP has a solution $(\hat{\theta}_N)$, and
2. any selection $\{\hat{\theta}_N\}_{N=1}^\infty$ of minimizers is consistent for θ_o ,

$$\hat{\theta}_N \xrightarrow{P} \theta_o.$$

ML-Estimator Consistency with Compactness

Q: Conditions verified?

1. Unique PP solution? Follows from θ_o identified,

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

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. $Q : \mathbb{R}^P \rightarrow \mathbb{R}$ be uniquely minimized at θ_o ;
2. each (random) $\{\widehat{Q}_N : \mathbb{R}^P \rightarrow \mathbb{R}\}_{N=1}^\infty$ convex; and,
3. $\widehat{Q}_N(\theta) \xrightarrow{p} Q(\theta)$ for each $\theta \in \mathbb{R}^P$.

Then

1. a minimizer $\widehat{\theta}_N$ of \widehat{Q}_N exists with probability $\rightarrow 1$; and
2. for any minimizer selection, $\widehat{\theta}_N \xrightarrow{p} \theta_o$.

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

ML-Estimator Consistency without Compactness

Q: Conditions verified?

1. If θ_o identified (in ML sense), then θ_o uniquely minimizes

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

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

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

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

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

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

Asymptotic Normality

Asymptotic Normality of M-Estimators

Theorem (W. Theorem 12.3)

Provided

- ▶ θ_o unique PP solution + interior to Θ compact,
- ▶ $q(\mathbf{w}, \cdot)$ cont' + twice cont' diff' on $\text{int } \Theta$,
- ▶ $E[\mathbf{s}(\mathbf{w}_i, \theta_o)] = \mathbf{0}$, and $E[\mathbf{H}(\mathbf{w}_i, \theta_o)]$ positive definite,
- ▶ (+ technical),

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

$$\begin{aligned}\sqrt{N}(\hat{\theta}_N - \theta_o) &\xrightarrow{d} N\left(\mathbf{0}, \mathbf{A}_o^{-1} \mathbf{B}_o \mathbf{A}_o^{-1}\right), \\ \mathbf{A}_o &:= E[\mathbf{H}(\mathbf{w}_i, \theta_o)], \\ \mathbf{B}_o &:= E\left[\mathbf{s}(\mathbf{w}_i, \theta_o) \mathbf{s}(\mathbf{w}_i, \theta_o)'\right].\end{aligned}$$

Oh, that Pesky Minus...

Thm. 12.3 designed for *minimization*.

Turn max'n into min'n:

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

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

In what follows,

$$\mathbf{s}_i(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})', \quad (\text{no minus})$$

$$\mathbf{H}_i(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta}), \quad (\text{no minus})$$

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

Information Matrix Equalities, I

Additionally assuming...

- ▶ θ_o interior to Θ ,
- ▶ $\ln f(\mathbf{y}|\mathbf{x}; \cdot)$ twice cont' diff' on $\text{int } \Theta$,
- ▶ (+ technical)

May now apply Thm. 12.3.

But further structure available...

Information Matrix Equalities, II

Under quite mild (smoothness) conditions,

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

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

(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

- ▶ θ_o identified + interior to Θ compact,
- ▶ $\ln f(\mathbf{y}|\mathbf{x}; \cdot)$ cont' + twice cont' diff' on $\text{int } \Theta$,
- ▶ + technical (including CIME justification),

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

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

Asymptotic Variance Estimation

Asymptotic Variance Estimators

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

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

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

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

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

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

Avar Estimation: Discussion

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

Pros/Cons:

1. Always available, 2nd-order diff', p.s.d.
2. Easy to compute, 1st-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

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

Cond'l probit log-likelihood contribution:

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

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

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

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

Probit Avar Estimation

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

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

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

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

Hence

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

Probit Avar Estimation

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

⇒ estimator of probit Avar:

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

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

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