Maximum Likelihood Methods

Jesper Riis–Vestergaard Sørensen

University of Copenhagen, Department of Economics

Maximum Likelihood Estimation: Aim

- Previously: Modelled [feature(s)] of distribution of $y|\mathbf{x}$.
  - E.g. $E(y|\mathbf{x})$ and $\operatorname{var}(y|\mathbf{x})$.
- Maximum likelihood estimation (MLE) more ambitious.
- Model for *entire distribution* $D(y|\mathbf{x})$

Why MLE? Advantages

Efficiency

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

May estimate any feature

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

Why Not MLE? Drawbacks

Nonrobustness

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

Outline

Framework

Truth vs. Model

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

- Possible values $(\mathbf{y}, \mathbf{x}) \in \mathscr{Y} \times \mathscr{X}$ for $(\mathbf{y}_i, \mathbf{x}_i)$.
- Discrete and/or continuous elements allowed.
- Only discrete: Integrals $\rightarrow$ Sums.

Parametric model: $f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^P$.

Model Assumptions

Parametric model: $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^P$.

Assume

1. Legitimate densities:

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \geqslant 0, \text{ all } (\mathbf{y},\mathbf{x},\boldsymbol{\theta}),$$
$$\int_{\mathscr{Y}} f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\, v(\mathrm{d}y) = 1, \text{ all } (\mathbf{x},\boldsymbol{\theta}).$$

2. Correct specification: For *some* $\boldsymbol{\theta}_o \in \Theta$,

$$p_o(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_o), \text{ all } (\mathbf{y},\mathbf{x}).$$

Identification in Maximum Likelihood Context

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

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

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

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

data generating process would show the same densities
you would not be able to tell true \theta and candidate \theta's apart

- I.e., $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_o$ are observationally equivalent.

Example

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: Is (α_o, μ_o) identified?

```
no, they are conditionally dependent on each other
Would not be able seperate them apart


A potential reason for non-convergence
```

Estimand

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

$$\underbrace{\max_{\boldsymbol{\theta}\in\Theta} E\left[\ln f\left(\mathbf{y}_i|\,\mathbf{x}_i;\boldsymbol{\theta}\right)\right]}_{\texttt{maximizing the expected log density}}. \qquad \text{(PP)}$$

Equivalently, $\boldsymbol{\theta}_o$ solves

$$\underbrace{\min_{\boldsymbol{\theta}\in\Theta} E\left[-\ln f\left(\mathbf{y}_i|\,\mathbf{x}_i;\boldsymbol{\theta}\right)\right]}_{\texttt{minimzing the negative log density}}.$$

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

Estimation

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

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

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

- Every MLE an M-estimator!

Example: Probit

Example: Probit

Binary outcome y_i, i.e. $\mathscr{Y} = \{0, 1\}$,

$$p_o\left(y|\mathbf{x}\right) = p_o\left(1|\mathbf{x}\right)^y \left[1 - p_o\left(1|\mathbf{x}\right)\right]^{1-y}, \quad y \in \{0, 1\}.$$

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

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

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

$$\mathrm{P}\left(y_i = 1|\mathbf{x}_i = \mathbf{x}\right) = \Phi\left(\mathbf{x}\boldsymbol{\theta}_o\right) \text{ for all } \mathbf{x} \in \mathscr{X}.$$

Example: Probit

Model implies density of $y_i | \mathbf{x}_i$,

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

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 solves

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

Identification and Solution Uniqueness

Identification and Uniqueness

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

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

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

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

Inequality *strict* provided g *strictly* concave + Z *nonconstant*.

Identification and Uniqueness

- Fix $\boldsymbol{\theta} \in \Theta$.
- $g := \ln(\cdot)$
- $Z := f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})/f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_o)$
- Cond'n on $\mathbf{x}_i \Rightarrow$ [FILL IN]

Identification and Uniqueness

- Correct specification + legitimate density $\Rightarrow$

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

- Hence

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

- Rearranging,

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

Identification and Uniqueness

- Have shown:

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

- Taking expectations,

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

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

Identification and Uniqueness

- Have shown: $\boldsymbol{\theta}_o$ solves PP,

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

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

- $\ln(\cdot)$ *strictly* concave, so Jensen $\Rightarrow$

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

- Hence, identification implies unique maximizer.

Asymptotic Properties

Asymptotic Properties of MLE

- Recall: Every MLE an M-estimator,

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

- May appeal to general results:
  - Consistency (W. Thm. 12.2)
  - Asymptotic normality (W. Thm. 12.3).
- Will verify relevant conditions.

Consistency

Recap: Consistency of M-Estimators

Theorem

If

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

(+ technical conditions), then

1. *Minimizer* $\widehat{\boldsymbol{\theta}}$ *of* $N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \cdot)$ *exists,*
2. $\widehat{\boldsymbol{\theta}}$ *consistent for* $\boldsymbol{\theta}_o$*.*

Consistency of ML-Estimators

Q: Conditions verified?

1. Θ compact? Assumed...
2. $q(\mathbf{w}, \cdot) = -\ln f(\mathbf{y}|\,\mathbf{x}; \cdot)$ continuous?
   - Assume $f(\mathbf{y}|\,\mathbf{x}; \cdot)$ cont's.
   - **Probit:** Φ *is* cont's.
3. Unique PP solution?
   - Follows from $\boldsymbol{\theta}_o$ identified.

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

Compact + LL cont' + ML identification $\Rightarrow$ MLE consistency.

Asymptotic Normality

Recap: Asymptotic Normality of M-Estimators

Theorem

Provided

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

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

Oh, that Pesky Minus...

- Thm. 12.3 designed for *min*imization.
- Turn max'n into min'n:

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

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

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

Information Matrix Equalities, I

- Additionally assuming...
  - $\boldsymbol{\theta}_o$ interior to Θ,
  - $\ln f(\mathbf{y}|\mathbf{x};\cdot)$ twice cont' diff' on int Θ,
  - (+ technical)
- May now apply Thm. 12.3.
- But further structure available...

Information Matrix Equalities, II

this is due to interchanging expectations (E) with derivatives
since expectations translate to integrals(?)

- Under quite mild conditions,

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

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

- (Un)Conditional Information Matrix Equality.
- Implies $\mathbf{A}_o = \mathbf{B}_o$.
- Asymptotic variance simplifies.

Asymptotic Normality of ML-Estimators

Theorem

Provided

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

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

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

Asymptotic Variance Estimation

Asymptotic Variance Estimators

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

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

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

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

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

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

Avar Estimation: Discussion

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

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

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 sample.

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 LL:

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

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

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

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

```
outer product product are like squaring
```

Probit Avar Estimation

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

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

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

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

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 + plug in $\widehat{\boldsymbol{\theta}}$ + invert.

$\Rightarrow$ variance matrix estimator for probit:

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

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

- Positive definite when invertible.
- Same result possible from 2^{nd}-order diff'n. (Check!)