Supplementary Material 2
Semiparametric Binary/Multinomial Choice Models

Hansen (2022, Sections 26.7-26.12)

Yasuyuki Matsumura (Kyoto University)

Last Updated: October 7, 2025

https://yasu0704xx.github.io

Introduction

- A semiparametric single index model is given by

$$Y = g(X^T \beta_0) + u,$$

where

$Y \in \mathbb{R}$: a dependent variable,

$X \in \mathbb{R}^q$: a $q \times 1$ explanatory vector,

$\beta_0 \in \mathbb{R}^q$: a $q \times 1$ vector of unknown parameters,

$u \in \mathbb{R}$: an error term which satisfies $\mathbb{E}(u \mid X) = 0$,

$g(\cdot)$: an unknown distribution function.

- Even though x is a $q \times 1$ vector, $x^T\beta_0$ is a scalar of a single linear combination, which is called a single index.
- By the form of the single index model, we obtain

$$\mathbb{E}(Y \mid X) = g(X^T\beta_0),$$

  which means that the conditional expectation of Y only depends on the vector X through a single index $X^T\beta_0$.
- The model is semiparametric when $\beta \in \mathbb{R}^q$ is estimated with the parametric methods and $g(\cdot)$ with the nonparametric methods.
  - If $g(\cdot)$ is the identity function, then the model turns out to be a linear regression model.
  - If $g(\cdot)$ is the CDF of Normal$(0, 1)$, then the model turns out to be a probit model.
  - If $g(\cdot)$ is the CDF of logistic distribution, then the model turns out to be a logistic regression model.

Textbooks

- Pagan and Ullah (1999, Chapter 7)
- Li and Racine (2007, Chapter 8)
- Horowitz (2009, Chapters 2 and 4)
- 西山・人見（2023，第 3 章）
- 末石（2024，第 4 章）

Contents

Binary Choice Models

- Klein and Spady (1993)
- Lewbel (2000)
- Manski (1975)
- Horowitz (1992)
- Han (1987)
- Mattzkin (1992)

Multinomial Discrete Choice Models

General Semiparametric MLE

Binary Choice Models

Semiparametric Binary Choice Model

- Consider the following binary choice model:

$$Y_i = \begin{cases} 1 \text{ if } Y_i^\star = \alpha + X_i^T\beta + \epsilon_i > 0, \\ 0 \text{ if } Y_i^\star = \alpha + X_i^T\beta + \epsilon_i \leq 0. \end{cases}$$

  To identify β, some standardization on α is required.

- This model can be rewritten as

$$\begin{aligned} \mathbb{E}(Y_i \mid X_i) &= \mathbb{P}(Y_i = 1 \mid X_i) \\ &= \mathbb{P}(\alpha + X_i^T\beta + \epsilon_i > 0) \\ &= \mathbb{P}(\epsilon_i > -X_i^T\beta - \alpha) \equiv g(X_i^T\beta), \end{aligned}$$

  which means that the binary choice model is a special case of the single index models.

- Suppose that $g(\cdot)$ were known. We would estimate β by maximum likelihood methods. The likelihood function would be

$$\begin{aligned} L^{\star}(b) &= \mathbb{P}(\epsilon_i > -X_i^T b - \alpha)^{\sum_{i=1}^n Y_i} \\ &\quad \times \mathbb{P}(\epsilon_i \leq -X_i^T b - \alpha)^{\sum_{i=1}^n (1-Y_i)} \\ &= g(X_i^T b)^{\sum_{i=1}^n Y_i} \times \{1 - g(X_i^T b)\}^{\sum_{i=1}^n (1-Y_i)}, \end{aligned}$$

  and then the log-likelihood function would be

$$L(b) = \sum_{i=1}^{n} [Y_i \log g(X_i^T b) + (1 - Y_i) \log(1 - g(X_i^T b))].$$

Klein and Spady's (1993) Binary Choice Estimator

- In reality, $g(\cdot)$ is unknown.
- Klein and Spady (1993) suggest to replace $g(\cdot)$ with the leave-one-out NW estimator $\widehat{g}_{-i}(X_i^T\beta) = \dfrac{\sum_{j\neq i} K\left(\frac{(X_i - X_j)^T\beta}{h}\right) Y_j}{\sum_{j\neq i} K\left(\frac{(X_i - X_j)^T\beta}{h}\right)}$.
- Making this substitution, and adding a trimming function, this leads to the feasible likelihood criterion:

$$L(\beta) = \sum_{i=1}^{n} [Y_i \log \widehat{g}_{-i}(X_i^T\beta) + (1 - Y_i)\log(1 - \widehat{g}_{-i}(X_i^T\beta))] 1_i(b),$$

where the trimming indicator should not be a function of β, but instead of a preliminary estimator:

$$1_i(b) = 1\left(\widehat{f}_{X^T\tilde{\beta}}(X_i^T\tilde{\beta}) \geq b\right),$$

with a preliminary estimator $\tilde{\beta}$ and density estimator $\widehat{f}_{X^T\tilde{\beta}}(\cdot)$.

- The following asymptotic properties hold:
  - under some regularity conditions, and
  - assuming that the kernel k is of higher-order (must be of fourth-order).
- Define $G(X_i^T\beta) = \mathbb{E}[g(X_i^T\beta_0) \mid X_i^T\beta]$. Then we obtain the asymptotic distribution:

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \text{Normal}(0, \Omega),$$

where the asymptotic variance is given by

$$\Omega = \mathbb{E}\left[\frac{\partial}{\partial\beta}G(X_i^T\beta)\frac{\partial}{\partial\beta}G(X_i^T\beta)^T\frac{1}{g(X_i^T\beta_0)(1-g(X_i^T\beta_0))}\right]^{-1}.$$

- Klein and Spady's (1993) estimator achieves the semiparametric efficiency bound for the single-index binary choice model (not for the general single-index model).

Lewbel's (2000) Estimator

- Consider the following binary choice model:

$$Y_i = 1(v_i + X_i^T \beta + \epsilon_i > 0),$$

  where v_i is a (special) continuous regressor whose coefficient is normalized to be one and X_i is of dimension q.
- Let $f(v|x)$ denote the conditional PDF of v_i given X_i.
- Let $F_\epsilon(\epsilon \mid v, x)$ denote the conditional CDF of ϵ_i conditioned on (v_i, X_i).

- Assumption: Assume that $F_\epsilon(\epsilon \mid v, x) = F_\epsilon(\epsilon \mid x)$.
- In words, here we assume that, conditional on x, ϵ is independent of the special regressor v.
- We also introduce an orthogonality condition: $\mathbb{E}(X_i \epsilon_i) = 0$.
- Identification:

$$\beta = \mathbb{E}[X_i X_i^T]^{-1} \mathbb{E}[X_i \tilde{Y}_i], \text{ where } \tilde{Y}_i = \frac{Y_i - 1(v_i > 0)}{f(v_i \mid X_i)}.$$

- Estimation: Use the sample analogue of identification result, replacing the unkown quantity $f(v_i \mid X_i)$ with its nonparametric kernel estimator [1] .

[1] Random denominator problem to possibly arise

Lewbel's (2000) Estimator: EXTENSION 1

- Assumption: There exists a $p \times 1$ vector Z_i of IVs, which satisfies that $\mathbb{E}(Z_i \epsilon_i) = 0$, $\mathbb{E}(Z_i X_i^T)$ is non-singular, and $F_{\epsilon x}(\epsilon, x \mid v, z) = F_{\epsilon x}(\epsilon, x \mid z)$.
- We do not assume the orthogonality condition here.
- Identification: [2]

$$\beta = \mathbb{E}[Z_i X_i^T]^{-1} \mathbb{E}[Z_i \tilde{Y}_i], \text{ where } \tilde{Y}_i = \frac{Y_i - 1(v_i > 0)}{f(v_i \mid X_i)}.$$

- Estimation: Use the sample analogue of identification result, replacing the unkown quantity $f(v_i \mid X_i)$ with its nonparametric kernel estimator. [3]

[2] This "just-identification" can be easily extended to over-identified cases.

[3] Random denominator problem to possibly arise

Lewbel's (2000) Estimator: EXTENSION 2

- Consider the ordered response model defined as

$$Y_i = \sum_{j=0}^{J-1} j 1(a_j < v_i + X_i^T \beta + \epsilon_i < a_{j+1}),$$

  where $a_0 = -\infty$ and $a_J = +\infty$.

- Y_j is called the response variable, which takes values in the set $\{0, 1, \cdots, J-1\}$.
- $Y_i = j$ if

$$a_j < v_i + X_i^T \beta + \epsilon_i < a_{j+1}.$$

- Let $X_{1i} = 1$ be the intercept and $\beta_1 = 0$ [4] .

[4] Required for identification in latent index models.

- Set $Y_{ji} = 1(Y_i \geq j)$ for $j = 1, \cdots, J-1$.
- Define $\Delta = \mathbb{E}[X_i X_i^T]^{-1}$ and Δ_j as the jth row of Δ.
- Identification:

$$a_j = -\Delta_1 \mathbb{E}\left[X_i \frac{Y_{ji} - 1(v_i > 0)}{f(v_i|X_i)}\right], \text{ for } j = 1, \cdots, J-1; \text{ and}$$

$$b_l = -\Delta_j \mathbb{E}\left[X_i \frac{\frac{\sum_{j=1}^{J-1} Y_{ji}}{(J-1)} - 1(v_i > 0)}{f(v_i|X_i)}\right], \text{ for } l = 2, \cdots, q.$$

- Estimation: Use the sample analogue of identification results, replacing $f(v_i|X_i)$ with its nonparametric kernel estimator.
- Further Extension: These results can be extended to mulitinomial choices models, partially linear latent variable models, and threshold and censored regression models.

Manski's (1975) Maximum Score Estimator

- Consider the binary choice model:

$$Y_i = \begin{cases} 1 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i > 0, \\ 0 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i \leq 0, \end{cases}$$

$$\text{med}(\epsilon_i \mid X_i) = 0 \quad (\iff \text{med}(Y_i^\star \mid X_i) = X_i^T\beta).$$

- Manski's maximum score estimator is defined as

$$\widehat{\beta}_M = \arg\max_\beta \sum_{i=1}^n Y_i 1(X_i^T\beta \geq 0) + (1 - Y_i)1(X_i^T\beta < 0),$$

which is a LAD estimator of a linear median-regression model.

- Under some assumptions, $\widehat{\beta}_M$ has strong consistency.
- The rate of convergence is $n^{-\frac{1}{3}}$ (Kim and Pollard, 1990).
- The limiting distribution is quite complex [5] and therefore not ideal for statistical inferences.
- To approximate the asymptotic distribution, Manski and Thompson (1986) use a bootstrap procedure.

[5] A distribution of a maximum of a multidimensional Brownian motion with quadratic drift.

Horowitz's (1992) Smoothed Maximum Score Estimator

- Consider the binary choice model:

$$Y_i = \begin{cases} 1 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i > 0, \\ 0 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i \leq 0, \end{cases}$$

$$\text{med}(\epsilon_i \mid X_i) = 0 \quad (\iff \text{med}(Y_i^\star \mid X_i) = X_i^T\beta).$$

- Horowitz's modified maximum score estimator is defined as

$$\widehat{\beta}_H = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)G\left(\frac{X_i^T\beta}{h}\right),$$

  where $G(\cdot)$ is a p-times continuously differentiable CDF.
- Recall that Manski's criterion function has the indicator functions, which lead to the lack of continuity.
- Horowitz (1992) modifies Manski's criterion, replacing the indicator functions with a twice continuously differentiable function that retains the essential features.

- Under some assumptions, as $n \to \infty; h = h_n > 0$; and $h \to 0$, $G\left(\frac{X_i^T \beta}{h}\right) \to 1(X_i^T \beta \geq 0)$.
- The convergence rate is $\sqrt{nh}$, and the asymptotic distribution is normal.
- Taking $h = \left(\frac{c}{n}\right)^{\frac{1}{2p+1}}$ for some $0 < c < \infty$, the convergence rate becomes $n^{-\frac{p}{2p+1}}$.
- With sufficiently large p, the convergence rate becomes close to $n^{-\frac{1}{2}}$.

Han's (1987) Maximum Rank (Correlation) Estimator

- Consider the binary choice model:

$$Y_i = \begin{cases} 1 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i > 0, \\ 0 \text{ if } Y_i^\star = X_i^T\beta + \epsilon_i \leq 0, \end{cases}$$

$$\text{med}(\epsilon_i \mid X_i) = 0 \quad (\iff \text{med}(Y_i^\star \mid X_i) = X_i^T\beta).$$

- Han's maximum rank correlation estimator [6] is defined as

$$\widehat{\beta} = \arg\max_{\beta} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} 1(Y_i \geq Y_j) 1(X_i^T\beta \geq X_i^T\beta).$$

[6] Han's estimator is defined using indicator functions as Manski's. While Manski's does not, Han's has $\sqrt{n}$-consistency and asymptotically normally distributed. It can be thought that the double-$\sum$ induce some "smoothness."

- To motivate the maximum rank correlation estimator, observe that

$$\mathbb{P}(Y_i > Y_j \mid X_i, X_j) \geq \mathbb{P}(Y_i \leq Y_j \mid X_i, X_j),$$

  whenever $X_i^T \beta_0 \geq X_j^T \beta_0$, which can be derived from the monotonicity of CDF and the independence of ϵ_i's and X_i's.
- Interpretation: When $X_i^T \beta_0 \geq X_j^T \beta_0$, more likely than not $Y_i \geq Y_j$.
- Let $G_H(\beta)$ denote the criterion function. Han (1987) shows that $\mathbb{E}[G_H(\beta)]$ is maximized at $\beta = \beta_0$.
- Han (1987) also establishes the storng consistency.
- Sherman (1993) shows that the maximum rank correlation estimator is $\sqrt{n}$-consistent and has an asymptotically normal distribution. [7]

[7] U-statistics

More General Cases?

- Matzkin (1992) does not impose any parametric structure on either the systematic function of the observed exogenous variable or on the distribution of the random error term.

Multinomial Discrete Choice Models

Multinomial Discrete Choice Models

- Consider the case where an individual faces $J > 2$ choices.
- Define $Y_{ij} = 1$ if individual i selects alternative $j \in \{1, \cdots, J\}$; and $Y_{ij} = 0$ otherwise.
- Set $F_{ij} = \mathbb{P}(Y_{ij} = 1 \mid X_i) = \mathbb{E}(Y_{ij} \mid X_i)$.
- The multiple choice equation is given in

$$Y_{ij} = F_{ij} + \epsilon_{ij}.$$

- The likelihood function is

$$\sum_{i=1}^{n} \sum_{j=1}^{J} Y_{ij} \ln F_{ij}.$$

Semiparametric Approach

- Set

$$\begin{aligned} F_{ij} &= \mathbb{E}(Y_{ij} \mid X_{i1}^T\beta_1, \cdots X_{iJ}^T\beta_J) \\ &= g(X_{i1}^T\beta_1, \cdots X_{iJ}^T\beta_J), \end{aligned}$$

  where the functional form of $g(\cdot)$ is unknown.

- Estimation methods are developed, for example, by Ichimura and Lee (1991); and Lee (1995).

General Semiparametric MLE

Ai's (1997) Semiparametric Maximum Likelihood Approach

- Ai (1997) considers a general semiparametric maximum likelihood estimation, which covers many semiparametric models such as multi-index models, partially linear models as special cases.