Models of Limited Dependent Variables

Zhentao Shi

The Chinese University of Hong Kong

Fundamental Task

- Use X to predict y
- Beyond continuous random variables
  - Binary
  - Multi-responses
  - Integer
  - Mixed type: censoring, truncation
  - Self-selection

- Applied microeconomics
- Biostatistics

Panoramic View

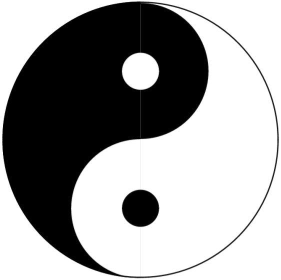
- MLE is a unifying framework
- Regressors X_i enter the model as a **single index** $X_i'\beta$
- Distributional assumptions are chosen for convenience
- Economists interpret the single index as a utility

Section 1

Binary Choices

Binary Outcome

- Outcome $y_i \in \{0,1\}$
- Binary regression
    - College entrance
    - Loan decision
    - Spam filter



Linear Probability Models

- Keep using linear regression $y_i = X_i'\beta + \varepsilon_i$
- Conditional mean

$$\Pr\left[y_i = 1 \mid X_i\right] = E\left[y_i \mid X_i\right] = X_i'\beta$$

- Error term $\varepsilon_i \in \{-X_i'\beta, 1 - X_i'\beta\}$ is binary. Conditional heteroskedastic.
- Predicted range: $E\left[y_i \mid X_i\right] = X_i'\beta$ can go out of $[0, 1]$.
  - $X_i'\beta$ is a single index.

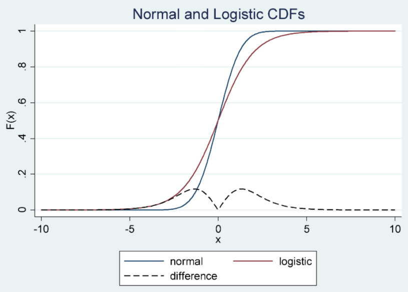
- Ironically, the linear model is popular (for causal inference)!

Generalized Linear Model

- To ensure predicted probability inside $[0,1]$, pick some $G(\cdot) : \mathbb{R} \to [0,1]$ to model

$$E\left(y_i = 1 \mid X_i\right) = G(X_i'\beta)$$

- Popular choices
  - **Probit**: $G(x) \sim$ Normal CDF
  - **Logit**: $G\left(x\right) \sim$ Logistic CDF
- Facts about Logistic CDF
  - $\Lambda = \Lambda\left(x\right) = \frac{1}{1+\exp(-x)}$
  - $\frac{\mathrm{d}\Lambda}{\mathrm{d}x} = \Lambda\left(1-\Lambda\right)$



Latent Utility Model

- Latent utility $y_i^* = X_i'\beta + \varepsilon_i$
- Observed outcome $y_i = \mathbb{I}\left\{y_i^* \geq 0\right\}$
- If $\varepsilon_i \mid X_i \sim$ Logistic, then

$$\begin{aligned}\Pr\left(y_i = 1 \mid X_i\right) &= \Pr\left(X_i'\beta + \varepsilon_i \geq 0 \mid X_i\right) \\ &= \Pr\left(-\varepsilon_i \leq X_i'\beta \mid X_i\right) \\ &= \Lambda(X_i'\beta)\end{aligned}$$

  where the last line holds if ε_i is symmetric around 0.
- The scale of β is not identifiable. Normalization is needed.

Log-Likelihood

Conditional likelihood of $y_i|X_i$ is

$$\left\{\Lambda\left(X_i'\beta\right)\right\}^{y_i}\left\{1-\Lambda\left(X_i'\beta\right)\right\}^{1-y_i}$$

A sample of N observations is

$$L_N(\beta)=\prod_{i=1}^{N}\left\{\Lambda\left(X_i'\beta\right)\right\}^{y_i}\left\{1-\Lambda\left(X_i'\beta\right)\right\}^{1-y_i}$$

Log-likelihood

$$\ell_N\left(\beta\right)=\sum_{i=1}^{N}\left\{y_i\log\left(\Lambda\left(X_i'\beta\right)\right)+\left(1-y_i\right)\log\left(1-\Lambda\left(X_i'\beta\right)\right)\right\}$$

Properties

- The score ($\Lambda(X_i'\beta)$ is simplified as Λ_i)

$$
\begin{aligned}
S_N(\beta) &= \sum_{i=1}^{N} \left\{ \frac{y_i}{\Lambda_i} \cdot \Lambda_i \left(1-\Lambda_i\right) X_i - \frac{(1-y_i)}{1-\Lambda_i} \cdot \Lambda_i \left(1-\Lambda_i\right) X_i \right\} \\
&= \sum_{i=1}^{N} \left\{ y_i \left(1-\Lambda_i\right) - \left(1-y_i\right) \Lambda_i \right\} X_i \\
&= \sum_{i=1}^{N} \left(y_i - \Lambda_i\right) X_i
\end{aligned}
$$

- Negative-definite second derivative

$$
\frac{\partial \ell_N\left(\beta\right)}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} \Lambda_i \left(1-\Lambda_i\right) X_i X_i'.
$$

- Globally concavity implies uniqueness of maximizer.

Goodness of Fit for Binary Classification

$$\text{McFadden } R^2 = 1 - \log L_1 / \log L_0$$

- $\log L_1$: maximum of likelihood
- $\log L_0$: the null model (no X, intercept only)

$$\log L_0 = N_1 \log \hat{p}_1 + (N - N_1) \log (1 - \hat{p}_1)$$

  where $\hat{p}_1 = N_1 / N$
- $\log L_0 < \log L_1 < 0$ implies $\frac{\log L_1}{\log L_0} \in [0, 1]$

Maximum Likelihood

- All nice properties of ML hold
  - Consistency
  - Asymptotic normality
- Misspecification?
- Choices of loss functions

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} Loss(\theta; data_i)$$

Loss Functions

Hastie, Tibshirani, Friedman (2008): The Elements of Statistical Learning

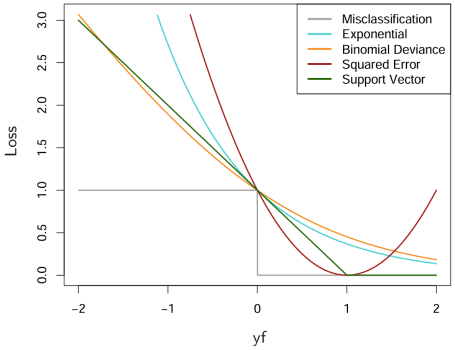


FIGURE 10.4. *Loss functions for two-class classification. The response is $y = \pm 1$; the prediction is f, with class prediction $\text{sign}(f)$. The losses are misclassification: $I(\text{sign}(f) \neq y)$; exponential: $\exp(-yf)$; binomial deviance: $\log(1+\exp(-2yf))$; squared error: $(y-f)^2$; and support vector: $(1-yf)_+$ (see Section 12.3). Each function has been scaled so that it passes through the point $(0,1)$.*

Prediction and Evaluation

Natural prediction:

$$\hat{y}_i = 1 \text{ if } \Pr\left(y_i \mid X_i\right) \geq 0.5$$

Outcomes: n_{11}: correct positive; n_{01}: false positive

	$\hat{y}_i = 0$	1	Total
$y_i = 0$	n_{00}	n_{01}	N_0
1	n_{10}	n_{11}	N_1
Total			N

- Hendrick-Merton: $\frac{n_{00}}{N_0} + \frac{n_{11}}{N_1}$
- Kuiper Score: $\frac{n_{11}}{N_1} - \frac{n_{01}}{N_0}$

Data Example

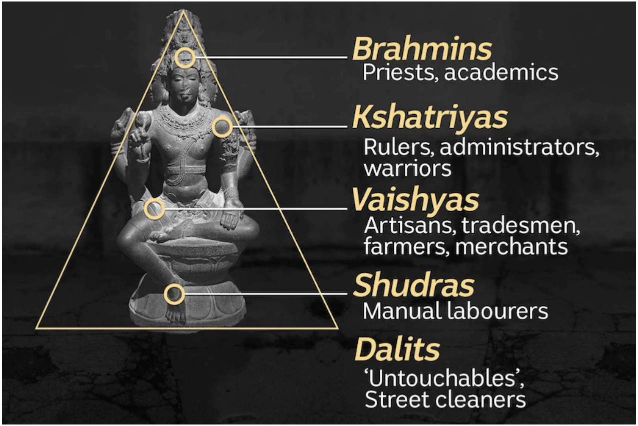
https://www.kaggle.com/code/jipann/logistic-regression

Section 2

Multiple Choices

Ordered Response

- More than two categories
- Categories are naturally ordered

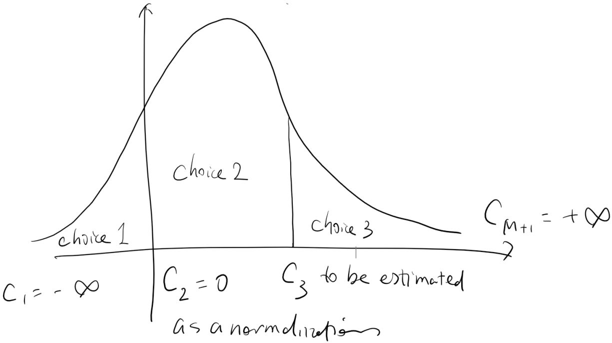


Utility

- Latent utility: $y_i^* = X_i'\beta + \varepsilon_i$ while the observed outcome

$$y_i = j, \quad \text{if } c_j < y_i^* \leq c_{j+1}$$

  - Normalization is needed for identification
  - M categories ($M \geq 3$)
  - $c_1 = -\infty$, $c_{M+1} = +\infty$, $c_2 = 0$



Probability

- Parameter: $\theta = (\beta, c_3, \ldots, c_M)$

$$
\begin{aligned}
P_{ij} &= \Pr\left(c_j < y_i^* \leq c_{j+1} \mid X_i\right) \\
&= \Pr\left(c_j < X_i'\beta + \varepsilon_i \leq c_{j+1} \mid X_i\right) \\
&= \Pr\left(c_j - X_i'\beta < \varepsilon_i \leq c_{j+1} - X_i'\beta \mid X_i\right) \\
&= G\left(c_{j+1} - X_i^{'}\beta\right) - G\left(c_j - X_i^{'}\beta\right)
\end{aligned}
$$

Likelihood

- Unobservable error $\varepsilon_i \mid X_i$ is assumed to be either logistic or normal
- Likelihood of individual observation

$$\Pr\left(y_i = j\right) = \sum_{j=1}^{M} P_{ij} \mathbb{I}\left\{y_i = j\right\}$$

- Likelihood of the N-observation sample

$$L\left(\theta\right) = \prod_{i=1}^{N} \Pr\left(y_i = j\right)$$

Choice of Transportation

Machine Learning: Handwriting

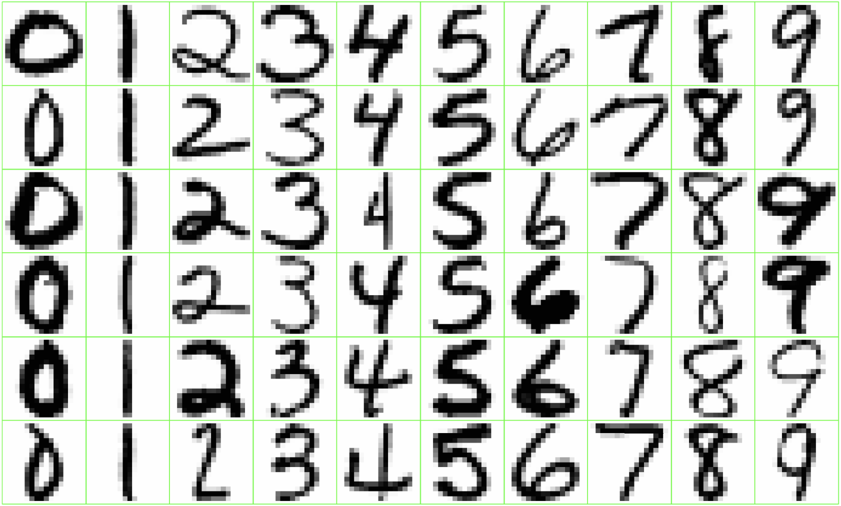


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

Multinomial Choice

- Level of individual-choice utility

$$\mu_{ij} = W'_{ij}\beta + Z'_i\beta_j$$

- Choice-specific regressors W_{ij}
  - e.g. Distance to stations (β is value of time)
- Choice-invariant regressors Z_i
  - e.g. Motion sickness (β_j is the effect; bus is bad)

Latent Utility

- For each $j = 1, 2, \ldots, M$, the utility $y_{ij}^* = \mu_{ij} + \varepsilon_{ij}$, where μ_{ij} is the choice level index, and for ε_{ij} is the error term.
- The observed choice

$$y_{ij} = \mathbb{I}\left\{ y_{ij}^* \geq \max_{k=1,\ldots,M} y_{ik}^* \right\}$$

- $\Pr\left(y_i = j \mid \mu_{i1}, \ldots, \mu_{iM}\right) = \Pr\left(y_{ij}^* \geq y_{i1}^*, \ldots, y_{ij}^* \geq y_{iM}^*\right)$

Distributional Assumption

- The probability depends on the joint distribution of $\left(\varepsilon_{ij}\right)_{j=1}^{M}$
- If $\varepsilon_{ij} \sim$ Type I extreme value distribution and ε_{ij} i.i.d. across choices, then

$$\Pr\left(y_i = j \mid \mu_{i1}, \ldots, \mu_{iM}\right) = \frac{\exp\left(\mu_{ij}\right)}{\sum_{k=1}^{M} \exp\left(\mu_{ik}\right)}$$

Normalization

- Normalize $\mu_{i1} = 0$ for all i. Usually for the "other" group.
- Equivalent to $\beta_{j=1} = 0$ (including the intercept)
- Parameters: $(\beta;$ and $\beta_2, \beta_3, \ldots, \beta_M)$

$$L(\theta) = \prod_{i=1}^{N} \left\{ \sum_{j=1}^{M} \mathbb{I}(y_i = j) \left(\frac{\exp(\mu_{ij})}{1 + \sum_{k=2}^{M} \exp(\mu_{ik})} \right) \right\}$$

where $\mu_{i1} = 0$ by standardization.

Independence of Irrelevant Alternative

- The concise form leverages that ε_{ij} is i.i.d. across choices $j = 1, 2, \ldots, M$
- Dilemma of "red bus" versus "blue bus"
- Must pay attention to the specification of choice set
- There are methods to fix it; still depending on specification.

- Daniel McFadden (Nobel Prize 2000)

Section 3

Integer Outcomes

Counting Model

- Outcomes take non-negative integers
  - Number of children
  - Number of hospital visits
  - Number of patents
- Poisson model: $y \sim \text{Poisson}(\lambda)$:

$$\Pr(y = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{for } k = 0, 1, 2, \ldots$$

Poisson Regression

- Poisson regression: Suppose

$$\lambda_i = \exp(X_i'\beta).$$

Model the single index λ_i by $\exp(X_i'\beta)$ gives log-likelihood

$$\log \Pr(y_i|X_i) = -\exp(X_i'\beta) + y_i \cdot X_i'\beta - \log k!$$

- Log-likelihood function of an N-observation sample:

$$\ell_N(\beta) = -\sum_{i=1}^{N} \exp(X_i'\beta) + \sum_{i=1}^{N} y_i X_i'\beta,$$

where $\log k!$ can be omitted.

Poisson MLE

- Score:

$$s_N(\beta) = \frac{\partial \ell_N(\beta)}{\partial \beta} = -\sum_{i=1}^{N} \exp(X_i'\beta) X_i + \sum_{i=1}^{N} y_i X_i.$$

- Second derivative is negative definite:

$$\frac{\partial^2 \ell_N(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \exp(X_i'\beta) X_i X_i'$$

- $\ell_N(\beta)$ is strictly concave in β.

Pseudo Poisson MLE

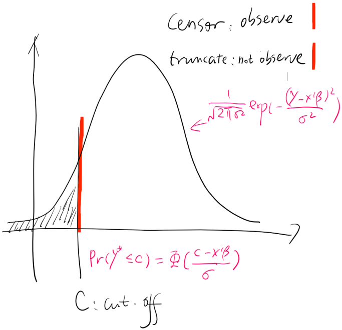
- Conditional mean model $E[y|X] = \exp(X'\beta)$
- If y is continuously distributed, the Poisson model must be misspecified
- e.g., Bilateral international trade between pairs of countries.

Data example:
`https://www.kaggle.com/code/jipann/poisson-regression`

Section 4

Incomplete Data

Censored or Truncated Data



- Latent utility: $y_i^* = X_i'\beta + \varepsilon_i$
- Observed outcome: $\begin{cases} y_i = y_i^*, & \text{if } y_i^* > c \\ y_i = c, & \text{if } y_i^* \le c \end{cases}$
- Real example `https://www.kaggle.com/datasets/lightonkalumba/us-womens-labor-force-participation`

Probabilities

- Assume $\varepsilon_i \mid X_i \sim N\left(0, \sigma^2\right)$
- The probability mass

$$\begin{aligned}\Pr\left(y_i = c\right) = \Pr\left(y_i^* \leq c\right) &= \Pr\left(X_i'\beta + \varepsilon_i \leq c\right) \\ &= \Pr\left(\frac{\varepsilon_i}{\sigma} \leq \frac{c - X_i'\beta}{\sigma}\right) = \Phi\left(\frac{c - X_i'\beta}{\sigma}\right)\end{aligned}$$

  where $\Phi(\cdot)$ is the CDF of $N(0,1)$.
- The density of the continuous region remains the same.

Likelihood

The likelihood consists of two components:

$$f\left(y_i \mid X_i\right) = \Phi\left(\frac{c - X_i'\beta}{\sigma}\right) \times \mathbb{I}\left(y_i = c\right) + \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - X_i'\beta\right)^2\right) \times \mathbb{I}\left(y_i > c\right)$$

- Mixed type of discrete and continuous random variable
- Mixed of probability mass function and density

$$\int_{-\infty}^{\infty} f\left(y \mid X\right) \mathrm{d}y = 1$$

Truncation

- If data of those with $y_i = c$ are completely unobservable
- We have data (Y_i, X_i) for those $y_i > c$ only.
- The likelihood with a condition on the outcome:

$$f(y_i \mid y_i \geq c, X_i) = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - X_i'\beta\right)^2\right)}{1 - \Phi\left(\frac{c - X_i'\beta}{\sigma}\right)}$$

- Due to truncation, OLS cannot consistently estimate β
- Must use MLE

Tobit II Models

- Wage (continuous): $y_i^* = X_{1i}'\beta_1 + \varepsilon_{1i}$
- Choice (binary) : $h_i^* = X_{2i}'\beta_2 + \varepsilon_{2i}$ and the observed outcome

$$h_i = \begin{cases} 1, & \text{if } h_i^* \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- Observe $y_i = y_i^*$ if $h_i = 1$
- The two equations have different regressors, coefficients and and errors terms.
- More flexible than the Tobit I model

Joint Normal

- Assume $\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix} \right)$ where $\sigma_{22} = 1$ is a normalization
- Parameter $\theta = (\beta_1, \beta_2, \sigma_{11}^2, \sigma_{12})$ can be estimated by MLE
- The conditional likelihood involves the bivariate normal distribution and its integrals
- No existing routine in `py::statmodels`

Conditional Expectation

- The conditional mean

$$
\begin{aligned}
E\left(y_i^* \mid h_i = 1\right) &= X_{1i}'\beta_1 + E\left(\varepsilon_{1i} \mid h_i = 1\right) \\
&= X_{1i}'\beta_1 + \sigma_{12}\lambda\left(X_{2i}'\beta_2\right)
\end{aligned}
$$

  due to the joint normality, where $\lambda\left(x\right) = \phi\left(x\right)/\Phi\left(x\right)$ is called the **inverse Mill's ratio**. $\phi(\cdot)$ is the pdf of $N(0,1)$
- The regression model can be estimated by `heckit`.
- In theory, `heckit` is inefficient...
- Real example: `https://www.kaggle.com/code/jipann/censored-regression`
- James Heckman (Nobel Prize 2000)

Summary

- Binary choice
- Multiple choices: ordered or unordered
- Poission regression
- Censored data and truncated data
- Selection model

- All are applications of MLE