Introductory Econometrics

Lecture 23: Binary Choice Models

Instructor: Ma, Jun

Renmin University of China

December 1, 2021

Binary dependent variable

- The explained variable could be binary, e.g. in a population survey dataset, with the subset of women considered, the explained variable can be a binary variable equal to one if the lady was participating work zero if not.
- Let Y_i be the explained variable and let $X_{1i}, X_{2i}, ..., X_{ki}$ be explanatory variables. We have i.i.d. observations $i = 1, 2, ..., n$.
- A linear regression of Y_i on the explanatory variables consistently estimates the best linear approximation to $\mathrm{E}\left[Y_i \mid X_{1i}, ..., X_{ki}\right]$.
- However, apparently, since Y_i is binary we have

$$\mathrm{E}\left[Y_i \mid X_{1i}, ..., X_{ki}\right] = \Pr\left[Y_i = 1 \mid X_{1i}, ..., X_{ki}\right].$$

  Therefore $\mathrm{E}\left[Y_i \mid X_{1i}, ..., X_{ki}\right]$ must be bounded between 0 and 1.
- The predicted value from a linear regression can be bigger than 1 or smaller than 0.

Specifying Logit and Probit models

- Since $\Pr[Y = 1 \mid X_1, ..., X_k]$ must be bounded between 0 and 1, we specify a parametric function form that respects this prior information.
- We consider a class of binary choice models of the form

$$\Pr[Y = 1 \mid X_1, ..., X_k] = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

  where G is a function taking on values strictly between 0 and 1: $0 < G(x) < 1$ for all $x \in \mathbb{R}$.
- The parameters to be estimated are $\beta_0, \beta_1, ..., \beta_k$. The estimated choice probabilities are strictly between 0 and 1.
- G can be taken to be a CDF with $0 < G(x) < 1$ for all $x \in \mathbb{R}$. We can take G to be the standard normal CDF. This is Probit model.
- Alternatively, we can take G to be the logitstic function:

$$G(z) = \frac{\exp(z)}{1 + \exp(z)}.$$

  This is the CDF for a standard logistic random variable. This is called a Logit model.

Latent variable model

- Logit and probit models can be derived from an underlying latent variable model.
- Suppose that we have an unobserved latent variable Y^*, generated by
$$Y^* = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$
where ϵ is independent of X's, e.g. Y^* is the net "return" of working for women.
- We observe $Y = 1[Y^* > 0]$ where $1[\cdot]$ is called the indicator function, which takes on one if the event in the brackets is true, and zero otherwise. Y is a binary random variable.
- We have
$$\begin{aligned} \Pr[Y = 1 \mid X_1, ..., X_k] &= \Pr[Y^* > 0 \mid X_1, ..., X_k] \\ &= \Pr[\epsilon > -(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \mid X_1, ..., X_k] \\ &= 1 - G(-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)) = G(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \end{aligned}$$
if the conditional distribution of ϵ is G.

Identification and normalization

- What if we take G to be the CDF of $\mathrm{N}\left(\mu, \sigma^2\right)$?
- Suppose $k = 1$. We observe

$$\begin{aligned} Y &= 1\left[\beta_0 + \beta_1 X_1 + \epsilon > 0\right] \\ &= 1\left[\frac{\beta_0 + \mu}{\sigma} + \frac{\beta_1}{\sigma} X_1 + \tilde{\epsilon} > 0\right] \end{aligned}$$

  where $\tilde{\epsilon} \sim \mathrm{N}(0,1)$. Let Φ denote the CDF of $\mathrm{N}(0,1)$.
- Denote $\tilde{\beta}_0 = (\beta_0 + \mu)/\sigma$ and $\tilde{\beta}_1 = \beta_1/\sigma$. Now we have

$$\Pr\left[Y = 1 \mid X_1 = x\right] = \Phi\left(\tilde{\beta}_0 + \tilde{\beta}_1 x\right).$$

  One cannot separately estimate β_0, β_1, μ and σ. Only $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are identified and estimable.
- As far as the “partial effect” is concerned, one does not need to separately estimate β_0, β_1, μ and σ. It suffices to estimate $\tilde{\beta}_0$ and $\tilde{\beta}_1$.

Partial effect

- The partial effect of X_j on $\Pr[Y = 1 \mid X_1, ..., X_k]$ is just

$$\frac{\partial \Pr\left[Y = 1 \mid X_1 = x_1, ..., X_j = x_j, ..., X_k = x_k\right]}{\partial x_j} = g\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)\beta_j$$

  where $g = G'$.
- Because G is the CDF of a continuous random variable, g is a probability density function. In Logit and Probit models, G is a strictly increasing CDF and so $g(z) > 0$ for all $z \in \mathbb{R}$.
- The partial effect depends on $(x_1, ..., x_k)$ but always has the same sign as β_j.
- We are often interested in estimating the average partial effect:

$$\mathrm{E}\left[g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)\beta_j\right].$$

Maximum likelihood estimation of Logit and Probit

- To obtain the maximum likelihood estimator, conditional on the explanatory variables, we need the conditional probability mass of Y given $X_1, ..., X_k$.
- We can write this as

$$\begin{aligned} &\Pr\left[Y = y \mid X_1, ..., X_k; \beta_0, \beta_1, ..., \beta_k\right] \\ &= \left[G\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)\right]^y \left[1 - G\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)\right]^{1-y} \end{aligned}$$

  with $y = 0, 1$.
- The log-likelihood function is

$$\begin{aligned} \ell\left(b_0, b_1, ..., b_k\right) = \sum_{i=1}^{n} \{ &Y_i \log\left(G\left(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}\right)\right) \\ &+ \left(1 - Y_i\right) \log\left(1 - G\left(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}\right)\right) \} . \end{aligned}$$

- Because G is strictly between 0 and 1 for Logit and Probit, $\ell(\cdot)$ is well-defined for all values of $b_0, b_1, ..., b_k$.
- The MLE $\hat{\beta}$ maximizes this log-likelihood function.

- If G is the standard Logit CDF, then $\hat{\beta}$ is the Logit estimator. If G is the standard normal CDF, then $\hat{\beta}$ is the Probit estimator.
- Because of the nonlinear nature of the maximization problem

$$\max_{b_0,\ldots,b_k} \ell\left(b_0,\ldots,b_k\right),$$

  we cannot write the maximum likelihood estimator as an explicit function of the data $\{(Y_i, X_{1i},\ldots,X_{ki}) : i = 1,\ldots,n\}$.
- The general theory of maximum likelihood implies that under general conditions, the maximum likelihood estimator is consistent and asymptotically normal: for each $j = 0,\ldots,k$,

$$\sqrt{n}\left(\hat{\beta}_j - \beta_j\right) \rightarrow_d \mathrm{N}\left(0, \mathrm{V}_j\right)$$

  with some asymptotic variance V_j.
- The form of V_j is very complex and not given in the class, but V_j is estimable.

Likelihood ratio test

- To test $H_0 : \beta_j = \beta_j^*$, we construct the usual t-statistic by using an estimate of V_j.
- Instead, we can conduct a likelihood ratio test.
- Suppose we want to test $H_0 : \beta_0 = \beta_0^*; \cdots ; \beta_q = \beta_q^*$ for $q \leq k$. The unconstrained maximum likelihood is

$$\ell_{uc} = \max_{b_0,...,b_k} \ell\left(b_0,...,b_k\right).$$

- The H_0−constrained maximum likelihood is

$$\ell_c = \max_{b_{q+1},...,b_k} \ell\left(\beta_0^*,...,\beta_q^*, b_{q+1},...,b_k\right).$$

- The likelihood ratio statistic is

$$LR = 2\left(\ell_{uc} - \ell_c\right).$$

- Under $H_0 : \beta_0 = \beta_0^*; \cdots ; \beta_q = \beta_q^*$, $LR \to_d \chi^2_{q+1}$.

Bayes theorem

- Continuous (X, Y):

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{\int f_{X|Y}(x \mid y)\, f_Y(y)\, \mathrm{d}y},$$

  where $\int f_{X|Y}(x \mid y)\, f_Y(y)\, \mathrm{d}y = f_X(x)$.

- Discrete (X, Y):

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \cdot \Pr(Y = k)}{\sum_{k=1}^{K} \Pr[X = x \mid Y = k] \cdot \Pr(Y = k)}$$

  where $Y \in \{1, ..., K\}$ and

$$\sum_{k=1}^{K} \Pr[X = x \mid Y = k] \cdot \Pr[Y = k] = \Pr[X = x].$$

Linear discriminant analysis (LDA) for two classes

- Specify:

$$X_1, \ldots, X_k \mid Y = 0 \sim \mathrm{N}(\mu_0, \Sigma)$$
$$X_1, \ldots, X_k \mid Y = 1 \sim \mathrm{N}(\mu_1, \Sigma),$$

  where (μ_0, μ_1) are k-dimensional vectors specifying the means and Σ is the variance-covariance matrix.
- By the Bayes theorem,

$$\Pr[Y = 1 \mid X_1, \ldots, X_k] = \frac{\pi_1 f_1(X_1, \ldots, X_k)}{\pi_0 f_0(X_1, \ldots, X_k) + \pi_1 f_1(X_1, \ldots, X_k)}$$
$$\Pr[Y = 0 \mid X_1, \ldots, X_k] = \frac{\pi_0 f_0(X_1, \ldots, X_k)}{\pi_0 f_0(X_1, \ldots, X_k) + \pi_1 f_1(X_1, \ldots, X_k)},$$

  where $\pi_k = \Pr[Y = k]$ and f_k is the conditional PDF of $(X_1, \ldots, X_k)$ given $Y = k$, $k \in \{0, 1\}$.
- The marginal distribution of Y (π_0, π_1) is left unspecified. (π_0, π_1) are easily estimated by sample averages.
- Estimation of (f_0, f_1) reduces to estimation of (μ_0, μ_1, Σ), which does not require numerical maximization (maximum likelihood).

LDA for $k = 1$

- The normal density has the form

$$f_j(x) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2\right),$$

  where μ_j is the mean and σ_j^2 is the variance, $j = 0, 1$.
- We assume that $\sigma_0^2 = \sigma_1^2 = \sigma^2$. Denote $p_j(x) = \Pr[Y = j \mid X = x]$ and then,

$$\begin{aligned} p_j(x) &= \frac{\pi_j f_j(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \\ &= \frac{\exp(\delta_j(x))}{\exp(\delta_0(x)) + \exp(\delta_1(x))}, \end{aligned}$$

  where the discriminant score $\delta_j(x)$ is defined by

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \log(\pi_j).$$

Estimating the parameters

- Estimator of π_j:

$$\hat{\pi}_j = \frac{n_j}{n},$$

where n_j is the number of observations in the j-th class, $j = 0, 1$.

- Estimator of μ_j:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} 1\,(Y_i = j)\, X_i,$$

average of all the observations from the j-th class.

- Estimator of σ^2:

$$\hat{\sigma}^2 = \sum_{k=0}^{K} \frac{n_j - 1}{n-2} \cdot \hat{\sigma}_j^2$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n} 1\left(Y_i = j\right)\left(X_i - \hat{\mu}_j\right)^2.$$

- $\hat{\sigma}^2$ is a weighted average of the sample variances for each of the classes.
- Then,

$$\hat{\delta}_j(x) = x \cdot \frac{\hat{\mu}_j}{\hat{\sigma}^2} - \frac{\hat{\mu}_j^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_j\right),$$

and we can turn these into estimates for conditional probabilities:

$$\hat{p}_j(x) = \frac{\exp\left(\hat{\delta}_j(x)\right)}{\exp\left(\hat{\delta}_0(x)\right) + \exp\left(\hat{\delta}_1(x)\right)}.$$

Logit/Probit versus LDA

- Logit/Probit:
  - Model the conditional distribution $Y \mid X$.
  - The distribution of X is not modeled.
  - Use MLE to estimate. This requires numerical optimization.
  - Economic justification: random utility model.
- LDA:
  - Model the conditional distribution $X \mid Y$.
  - The distribution of Y is not modeled.
  - Estimation: sample means, variances, and covariances of X.
  - No clear economic model.