ECON 480-3
LECTURE 12: BINARY CHOICE

Ivan A. Canay
Northwestern University

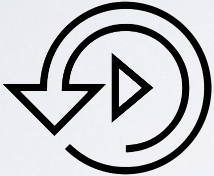
Past & Future

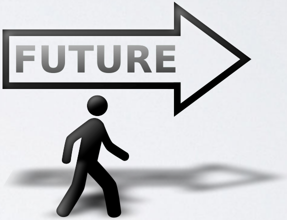
LAST CLASS

- Regression Tress
- Classification Tress
- Random Forests

TODAY

- Related to Classification Tress
- Latent Index and Identification
- Identification via Median Independence
- Parametric Models: Logit & Probit





Setup

- **Today** we consider the problem of estimating

$$P\{Y = 1|X\}$$

  where Y is **binary**, i.e., takes values in $\{0, 1\}$

- **Two problems**
  - Predicting Y given X (e.g., propensity score)
  - Viewing $P\{Y = 1|X\}$ as a model to identify partial effects.
- We consider **parametric** and **semi-parametric models**.
- Both based on the so-called **Linear Index** where (Y, X, U) is such that
  - Y takes values in $\{0, 1\}$
  - U take values in $\mathbf{R}$
  - X takes values in $\mathbf{R}^{k+1}$ with $X_0 = 1$.
  - $P\{Y = 1|X\} = P\{Y = 1|X'\beta\}$ for some $\beta = (\beta_0, \beta_1, \dots, \beta_k)' \in \mathbf{R}^{k+1}$

Linear Index

- Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = I\{X'\beta - U \geqslant 0\} . \tag{1}$$

- This is known as a Threshold crossing model or Single index model or Linear index model

- Y often indicates a utility-maximizing decision maker's observable choice between two alternatives.

- **Latent index**: $X'\beta - U$ can be interpreted as the difference in the utility between the two choices.

- We first discuss conditions for **identification** of this model.

Definition of Identification

- Let P denote the distribution of the observed data.
- Denote by $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ a **model** for P.
- θ could have infinite dimensional components.
- Model is correctly specified: $P \in \mathbf{P}$.
- Interest might be in θ or a function $\lambda(\theta)$.

Identification

Let $\Theta_0(P)$ be the collection of θ such that $P = P_\theta$, i.e.

$$\Theta_0(P) = \{\theta \in \Theta : P_\theta = P\}.$$

We say that θ is identified if $\Theta_0(P)$ is a singleton for all $P \in \mathbf{P}$.

Note: $\lambda(\theta)$ may be identified even if θ is not.

Identification: Parametric Binary Model

- In the binary choice model the parameter is $\theta = (\beta, P_X, P_{U|X})$.
- Θ is the set of all possible values of θ.
- Identification **almost** follows from the following assumption:

Assumption (Parametric)

P1 $P_{U|X} = N(0, \sigma^2)$.

P2 *There exists no $A \subseteq \mathbf{R}^{k+1}$ such that A has probability one under P_X and A is a proper linear subspace of $\mathbf{R}^{k+1}$*

- Given assumption P1 we may replace $P_{U|X}$ with σ: $\theta = (\beta, P_X, \sigma)$.
- Proof approach: suppose that there are two values of θ,

$$\theta = (\beta, P_X, \sigma) \text{ and } \theta^* = (\beta^*, P_X^*, \sigma^*) ,$$

such that $\theta \neq \theta^*$ and $P = P_\theta = P_{\theta^*}$. Then reach a contradiction.

Identification: Proof

- The marginal dist. of X is identified from the joint dist. of $(Y, X) \Rightarrow$ it **must be** that $P_X = P_X^*$.
- **P1 implies**:
$$P_\theta\{Y = 1|X\} = \Phi\left(\frac{X'\beta}{\sigma}\right) \quad \text{and} \quad P_{\theta^*}\{Y = 1|X\} = \Phi\left(\frac{X'\beta^*}{\sigma^*}\right) .$$
- Since $P_\theta = P_{\theta^*}$ by **assumption**, it must be
$$\frac{\beta}{\sigma} = \frac{\beta^*}{\sigma^*} . \tag{2}$$
- We **cannot conclude** that $\beta = \beta^*$ and $\sigma = \sigma^*$.
- **Indeed**: our analysis shows that any θ and θ^* for which (2) holds and $P_X = P_X^*$ satisfies $P_\theta = P_{\theta^*}$.
- We cannot identify $\theta = (\beta, P_X, \sigma)$ **BUT** we can identify $\lambda(\theta) = (P_X, \beta/\sigma)$.

Identification: Comments

- "Normalization": researchers typically assume further that $|\beta| = 1$, $\beta_0 = 1$, or $\sigma = 1$.
- The model with $\sigma = 1$ is called **Probit** and it identifies $\theta = (\beta, P_X, 1)$.
- To see this, note that from P1 and $\sigma = 1$

$$P_\theta\{Y = 1|X\} = \Phi\left(X'\beta\right) = \Phi\left(X'\beta^*\right) = P_{\theta^*}\{Y = 1|X\}$$

  holds a.s. for $\beta \neq \beta^*$ iff

$$P_X\{X'\beta = X'\beta^*\} = 1 \ , \tag{3}$$

  which violates P2 with $A = \{x \in \mathbf{R}^{k+1} : x'(\beta - \beta^*) = 0\}$.

- Other parametric assumptions possible: **Logit**.
- **Question**: is θ identified without parametric assumptions on $P_{U|X}$?

QUESTIONS?

Mean independence

- **First idea**: mimic the linear model.
- **Linear model**: all we needed from $P_{U|X}$ was $E[U|X] = 0$.
- Replacing P1 with $E[U|X] = 0$ does **not** work
  – Manski (1988) shows nothing is learned about $(\beta, P_{U|X})$.
- Note even useful to identify $\lambda(\theta) = \beta$ in this case.
- **In general**: mean independence assumptions are rather useless in non-linear models.

Median independence

- **Median independence**: $\lambda(\theta) = \beta$ is identified under reasonable conditions if $\mathrm{Med}(U|X) = 0$.

Assumption (Semi-Parametric)

S1 $\mathrm{Med}(U|X) = 0$ *with probability 1 under* P_X

S2 *There exists no* $A \subseteq \mathbf{R}^{k+1}$ *such that* A *has probability one under* P_X *and* A *is a proper linear subspace of* $\mathbf{R}^{k+1}$

S3 $|\beta| = 1$.

S4 P_X *is such that at least one component of* X *has support equal to* $\mathbf{R}$ *conditional on the other components with probability 1 under* P_X*. Moreover, the corresponding component of* β *is non-zero.*

- S1 is weaker than P1
- S2 is the same as P2
- S3 is a normalization similar to $\sigma = 1$ in the Probit case.
- S4 is new: stronger assumption on P_X and also on β.

Identification: Median Independence

The following lemma will help us prove the result.

Lemma

Let $\theta = (\beta, P_X, P_{U|X})$ satisfying $S1$ be given. Consider any β^*. If

$$P_\theta\left\{X'\beta^* < 0 \leqslant X'\beta \cup X'\beta < 0 \leqslant X'\beta^*\right\} > 0 \tag{4}$$

then there exists no $\theta^* = (\beta^*, P^*_X, P^*_{U|X})$ satisfying $S1$ and also having $P_\theta = P_{\theta^*}$.

PROOF: Suppose by contradiction that (4) holds yet there exists such θ^*.

Because $P_\theta = P_{\theta^*}$ then $P_X = P^*_X$. Now note that $Y = I\{X'\beta - U \geqslant 0\}$ so

$$\begin{aligned} P_\theta\{Y = 1|X\} \geqslant \frac{1}{2} &\iff P_\theta\{X'\beta \geqslant U\} \geqslant \frac{1}{2} \\ &\iff X'\beta \geqslant 0 \quad \text{by Assumption } S1 . \end{aligned}$$

Likewise

$$\begin{aligned} P_{\theta^*}\{Y = 1|X\} \geqslant \frac{1}{2} &\iff P_{\theta^*}\{X'\beta^* \geqslant U\} \geqslant \frac{1}{2} \\ &\iff X'\beta^* \geqslant 0 \quad \text{by Assumption } S1 . \end{aligned}$$

$$P_\theta\left\{X'\beta^* < 0 \leqslant X'\beta \cup X'\beta < 0 \leqslant X'\beta^*\right\} > 0$$

Our condition implies that with positive probability, either

$$X'\beta^* < 0 \leqslant X'\beta$$

or

$$X'\beta < 0 \leqslant X'\beta^* ,$$

which implies that either

$$P_{\theta^*}\{Y = 1|X\} < \frac{1}{2} \leqslant P_\theta\{Y = 1|X\}$$

or

$$P_\theta\{Y = 1|X\} < \frac{1}{2} \leqslant P_{\theta^*}\{Y = 1|X\} .$$

This contradicts the fact that $P_\theta = P_{\theta^*}$ and completes the proof.

THEOREM

Under assumptions $S1 - S4$, $\lambda(\theta) = \beta$ is identified.

PROOF: Assume wlog that the component of X specified in $S4$ is the ***k*th component** and that $\beta_k > 0$.

Let θ satisfying $S1$-$S4$ be given. Consider any $\beta^* \neq \beta$.

Wish to show there is no $\theta^* = (P^*_X, \beta^*, P^*_{U|X})$ satisfying $S1$-$S4$ s.t $P_\theta = P_{\theta^*}$.

From the previous Lemma it **suffices** to show that:

$$P_\theta \left\{ X'\beta^* < 0 \leqslant X'\beta \cup X'\beta < 0 \leqslant X'\beta^* \right\} > 0 \, .$$

We now divide the proof in **three cases** according to sign(β^*_k)

Identification: Median Independence

Case 1 Suppose $\beta_k^* < 0$. Then,

$$P_\theta\{X'\beta^* < 0 \leqslant X'\beta\} = P_\theta\left\{X_k > -\frac{X'_{-k}\beta^*_{-k}}{\beta_k^*},\ X_k \geqslant -\frac{X'_{-k}\beta_{-k}}{\beta_k}\right\}.$$

This probability is positive by $S4$

Case 2 Suppose $\beta_k^* = 0$. Then,

$$P_\theta\{X'\beta^* < 0 \leqslant X'\beta\} = P_\theta\left\{X'_{-k}\beta^*_{-k} < 0,\ X_k \geqslant -\frac{X'_{-k}\beta_{-k}}{\beta_k}\right\} \quad (5)$$

$$P_\theta\{X'\beta < 0 \leqslant X'\beta^*\} = P_\theta\left\{X'_{-k}\beta^*_{-k} \geqslant 0,\ X_k < \frac{X'_{-k}\beta_{-k}}{\beta_k}\right\} \quad (6)$$

If $P_\theta\{X'_{-k}\beta^*_{-k} < 0\} > 0$ then (♣) is positive by $S4$

If $P_\theta\{X'_{-k}\beta^*_{-k} \geqslant 0\} > 0$ then (6) is positive by $S4$.

Identification: Median Independence

Case 3 Suppose $\beta_k^* > 0$. Then,

$$P_\theta\{X'\beta^* < 0 \leqslant X'\beta\} = P_\theta\left\{-\frac{X'_{-k}\beta_{-k}}{\beta_k} \leqslant X_k < -\frac{X'_{-k}\beta^*_{-k}}{\beta^*_k}\right\} \quad (7)$$

$$P_\theta\{X'\beta < 0 \leqslant X'\beta^*\} = P_\theta\left\{-\frac{X'_{-k}\beta^*_{-k}}{\beta^*_k} \leqslant X_k < -\frac{X'_{-k}\beta_{-k}}{\beta_k}\right\} \quad (8)$$

- **Problem** if

$$P_\theta\left\{\frac{X'_{-k}\beta_{-k}}{\beta_k} = \frac{X'_{-k}\beta^*_{-k}}{\beta^*_k}\right\} = 1 \quad (\clubsuit)$$

- **Assumption** $S3$: implies that β^* is not a scalar multiple of β, Therefore,

$$\frac{\beta^*_{-k}}{\beta^*_k} \neq \frac{\beta_{-k}}{\beta_k}$$

- It follows from $S2$ and $S3$ that $\clubsuit$ cannot happen.
- Adding $S4$ then implies that at at least one of (7) and (8) must be positive. This concludes the proof.

QUESTIONS?

Estimation: parametric cases

- Previous Theorem identifies β only: not enough for marginal effects (later)
- Go back to parametric case where
$$P\{Y=1|X\}=F(X'\beta)$$
with $F(\cdot)$ being
  1. **PROBIT**: $F(x)=\Phi(x)$
  2. **LOGIT**: $F(x)=\frac{\exp(x)}{1+\exp(x)}$
- **Data**: a random sample of size n from the distribution of (Y,X), i.e., $(Y_1,X_1),\ldots,(Y_n,X_n)$
- The model is parametric, so we can do **Maximum Likelihood Estimation**.
- First write the probability mass function (pmf) of Y_i
$$f_\beta(Y_i|X_i)=F(X_i'\beta)^{Y_i}(1-F(X_i'\beta))^{1-Y_i}$$
- Now we can write the **log-likelihood**.

MLE

- **Log-likelihood function**:

$$\ell_n(b) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(f_b(Y_i|X_i)\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \ln\left(F(X_i'b)\right) + (1-Y_i)\ln\left(1-F(X_i'b)\right)\right\}$$

- Can be shown β is the unique maximizer of $Q(b) = E[\ell_n(b)]$.
- Let $\hat{\beta}_n$ be the MLE.
- By usual MLE results,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

where $\mathbb{V} = \mathbb{I}_\beta^{-1}$ and

$$\mathbb{I}_\beta = -E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left(f_\beta(Y_i|X_i)\right)\right].$$

Asymptotic Variance

- By the information equality

$$\mathbb{I}_\beta = -E\left[\frac{\partial^2}{\partial\beta\partial\beta'}\ln\left(f_\beta(Y_i|X_i)\right)\right] = E\left[\frac{\partial}{\partial\beta}\ln\left(f_\beta(Y_i|X_i)\right)\frac{\partial}{\partial\beta'}\ln\left(f_\beta(Y_i|X_i)\right)\right].$$

- Since

$$\frac{\partial}{\partial\beta}\ln\left(f_\beta(Y_i|X_i)\right) = \left[\frac{Y_i - F(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\right]F'(X_i'\beta)X_i$$

We get that

$$\begin{aligned}\mathbb{I}_\beta &= E\left[\left[\frac{Y_i - F(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\right]^2 F'(X_i'\beta)^2 X_i X_i'\right]\\ &= E\left[\frac{F'(X_i'\beta)^2}{F(X_i'\beta)(1-F(X_i'\beta))}X_i X_i'\right].\end{aligned}$$

The second equality comes from the law of iterated expectations and law of total variance (480-2).

Interpreting β

- For the moment, consider X_j continuously distributed.
- In linear regression with $E[U|X]$ we had

$$\frac{\partial E[Y|X]}{\partial X_j} = \beta_j \, .$$

- In Binary models we rather have

$$\frac{\partial E[Y|X]}{\partial X_j} = \frac{\partial P\{Y=1|X\}}{\partial X_j} = \frac{\partial F(X'\beta)}{\partial X_j}\beta_j \, .$$

- PROBIT: $F' = \phi$ so that

$$\frac{\partial P\{Y=1|X\}}{\partial X_j} = \phi(X'\beta)\beta_j \, .$$

- LOGIT: $F' = F(1-F)$ so that

$$\frac{\partial P\{Y=1|X\}}{\partial X_j} = F(X'\beta)(1-F(X'\beta))\beta_j \, .$$

Interpreting β - cont.

- We can still extract information by simply inspecting β
- **Fact 1:** ratio of β has meaning in terms of partial effects

$$\frac{\frac{\partial P\{Y=1|X\}}{\partial X_j}}{\frac{\partial P\{Y=1|X\}}{\partial X_k}} = \frac{\beta_j}{\beta_k} \; .$$

- **Fact 2:** Since $F' > 0$, $\mathrm{sign}(\beta_j)$ identifies the sign of the marginal effect.
- **Fact 3:** easy to get upper bound on marginal effects from β

PROBIT

$$\frac{\partial P\{Y=1|X\}}{\partial X_j} \leqslant 0.4\beta_j \quad \text{since } \phi(x) \leqslant \phi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4 \; .$$

LOGIT

$$\frac{\partial P\{Y=1|X\}}{\partial X_j} \leqslant \frac{1}{4}\beta_j \quad \text{since } F(1-F) \leqslant \frac{1}{4} \; .$$

Estimation of marginal effects

- Marginal effects for X_j depends on the entire vector X.
- We can compute the average/mean marginal effect,

$$E\left[\frac{\partial P\{Y=1|X\}}{\partial X_j}\right] = E[F'(X'\beta)]\beta_j$$

- And estimate this by

$$\frac{1}{n}\sum_{i=1}^{n} F'(X_i'\hat{\beta}_n)\hat{\beta}_{n,j} \; .$$

- Distinction between that and marginal effects “at the average”, i.e.

$$F'(E[X]'\beta)\beta_j \; ,$$

which can be estimated by

$$F'(\bar{X}_n'\hat{\beta}_n)\hat{\beta}_{n,j} \; .$$

- Stata offers both options with the option margins.

Estimation of marginal effects II

- Partition $X = (X_1, D)$ where $X \in \mathbf{R}^k$ and $D \in \{0,1\}$. Partition $\beta = (\beta_1, \beta_2)$ accordingly. In this case using
$$E\left[\frac{\partial P\{Y=1|X\}}{\partial D}\right] = E[F'(X'\beta)]\beta_2$$
does not make a lot of sense.
- The marginal effect in this case is
$$P\{Y=1|X_1, D=1\} - P\{Y=1|X_1, D=0\} = F(X_1'\beta_1 + \beta_2) - F(X_1'\beta_1) \ .$$
- Averaging X_1 out,
$$E\left[F(X_1'\beta_1+\beta_2) - F(X_1'\beta_1)\right] \ .$$
- And we can estimate this by
$$\frac{1}{n}\sum_{i=1}^{n} F(X_1'\hat{\beta}_{n,1} + \hat{\beta}_{n,2}) - F(X_{1,i}'\hat{\beta}_{n,1}) \ .$$
- Alternative: marginal effect on the treated by conditioning on $D = 1$.

Estimation of marginal effects III

- **Note:** It often makes sense to report marginal effects in a table
- This requires standard errors for those marginal affects.
- In the continuous case

$$\frac{\partial P\{Y = 1|X\}}{\partial X_j} = F'(X'\beta)\beta_j = h(\beta)$$

  for a **known function** $h(\beta)$. Similarly for the discrete case.
- Can compute standard errors via the **Delta Method**.
- Stata has options for this: see margins.

Logit and the odds ratio

- In statistic and Biostatistic the Logit model has particular appeal.
- Let $p_i = P\{Y_i = 1|X_i\}$. Since

$$p_i = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \quad \Rightarrow \quad \frac{p_i}{1-p_i} = \exp(X_i'\beta)$$

  and so taking logs

$$\ln\left(\frac{p_i}{1-p_i}\right) = X_i'\beta \ .$$

- $p/(1-p)$ is the odds ratio or relative risk. Say $Y = 1$ if you live and $Y = 0$ if you die in a clinical trial. An odds ratio of 2 means that the odds of survival are twice those of death.
- β_j is the marginal effect of X_j on the log odds ratio.
- Interpretation: $\beta_j = 0.1$ means the relative probability of survival increases by 10% (roughly)
- Such easy rounding works for small values of β_j.

QUESTIONS?

Linear probability model

- Some people still advocate the use of the linear probability model where

$$Y = X'\beta + U \tag{9}$$

and $E[U|X] = 0$.

- Reason: β directly delivers “marginal effects”, easy to accommodate instrumental variables, panels with fixed effects, etc.
- If Y is binary, 2SLS still admits LATE interpretation, etc.
- These extensions are hard in Probit/Logit: e.g., bivariate Probit and other more recent methods.
- However: hard to interpret the linear model causally as $E[Y|X]$ cannot be linear in most cases

  The true $E[Y|X]$ may arise from a causal model, but the regression is only providing a linear approximation to the true $E[Y|X]$.

- Still, may use the linear model as a descriptive tool to approximate $E[Y|X]$ - will still be the best linear approximation and predictor.
- **But $E[Y|X = x]$ is fundamentally non linear**.

Linear probability model

- **Consequence**: LPM often delivers predicted probabilities outside $[0, 1]$.
- Angrist and Pischke (p.103): "...[linear regression] may generate fitted values outside the LDV boundaries. This fact bothers some researchers and has generated a lot of bad press for the linear probability model."
- Well said...however, later on they add...
- Angrist and Pischke (p.197): "Yet we saw that the added complexity and extra work required to interpret the results from latent index models may not be worth the trouble".
- This statement is controversial, at the very least. You should read MHE with care...

Comments

- Logit, Probit, and LPM yield quite different estimates $\hat{\beta}_n$.
- Expected: if we use the upper bounds for marginal effects, we get

$$\begin{aligned} \hat{\beta}_{\text{logit}} &\approx 4\hat{\beta}_{\text{ols}} \\ \hat{\beta}_{\text{probit}} &\approx 2.5\hat{\beta}_{\text{ols}} \\ \hat{\beta}_{\text{logit}} &\approx 1.6\hat{\beta}_{\text{probit}} \,. \end{aligned}$$

- However, average marginal effects from Logit, Probit, and even LPM are often **"close"**.
- Partly due because there is **averaging** going on.

Extensions

- Similar ideas to those discuss here apply to other settings.
- **Ordered choice**: individual decides how many units to buy from the same item.
- **Unordered choice**: individual decides to buy 1 of many different alternatives.
- **Conditional Logit** and **multinomial Logit** arise.
- Most popular example: “BLP” in IO.
- These topics are covered in second year IO classes.

THE END