

第11章 離散選択モデル

質的従属変数 (Qualitative Dependent Variable) の一種：離散選択モデル (Discrete Choice Model)

通常的回帰モデル：

$$Y_i = \alpha + \beta X_i + u_i \quad u_i \sim N(0, \sigma^2) \quad i = 1, 2, \dots, n$$

u_i が $-\infty$ から ∞ の範囲の値を取る連続型確率変数なので、 Y_i も連続型確率変数である。

Y_i が離散型確率変数であればどうなるか？

(* 復習) 離散型確率変数と連続型確率変数：

離散型確率変数 X ：不連続な値を取る

- サイコロの出た目 (X の取る値は $1, 2, 3, 4, 5, 6$)
- コインを投げて表・裏 (X の取る値：表の場合は 1 ，裏の場合は 0)

...

連続型確率変数 X ：ある区間内 ($-\infty$ から ∞ の区間も含む) のどの実数値も取り得る

11.1 二値選択モデル (Binary Choice Model)

アンケート調査を行う。

回答は YES か NO の 2 つから 1 つを選択することとする。

$$Y_i = \begin{cases} 1, & i \text{ 番目の人が YES と答えたとき} \\ 0, & i \text{ 番目の人が NO と答えたとき} \end{cases}$$

通常の次の回帰モデルを考える。

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \dots, n$$

u_1, u_2, \dots, u_n は互いに独立で, 平均ゼロ・分散 σ^2 とする。

$E(u_i) = 0$ なので , Y_i の期待値は ,

$$E(Y_i) = \alpha + \beta X_i$$

となる。

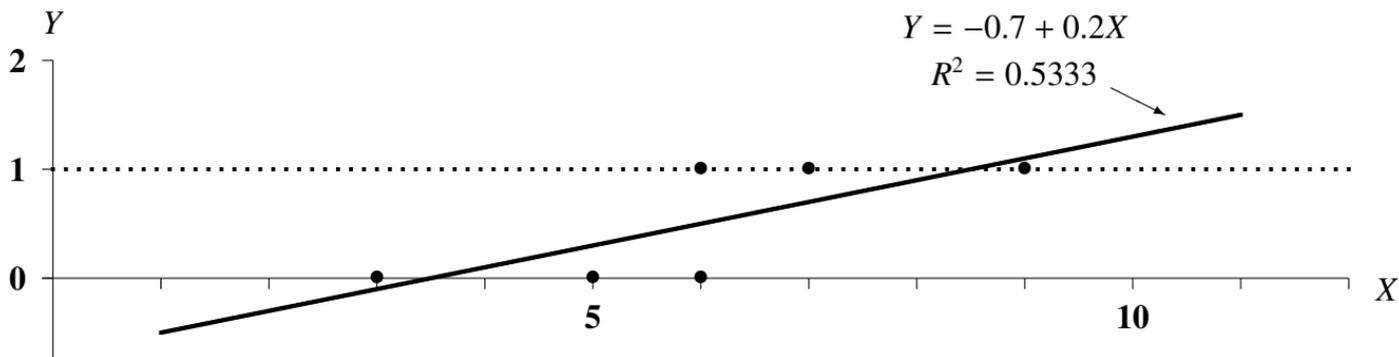
線形関数なので , $\alpha + \beta X_i$ は $-\infty$ から ∞ の値を取ることになる。

この定式化は適切か？

数值例：

i	1	2	3	4	5	6
Y_i	0	0	0	1	1	1
X_i	3	5	6	6	7	9

最小二乗法で α, β を推定する。



Gretl による最小二乗法の結果：

? ols y const x

モデル 1: 最小二乗法 (OLS), 観測: 1-6
従属変数: y

	係数	標準誤差	t 値	p 値
const	-0.700000	0.586657	-1.193	0.2987
x	0.200000	0.0935414	2.138	0.0993 *
Mean dependent var	0.500000	S.D. dependent var	0.547723	
Sum squared resid	0.700000	S.E. of regression	0.418330	
R-squared	0.533333	Adjusted R-squared	0.416667	
F(1, 4)	4.571429	P-value(F)	0.099301	
Log-likelihood	-2.068328	Akaike criterion	8.136656	
Schwarz criterion	7.720175	Hannan-Quinn	6.469448	

一方, $E(Y_i)$ を計算する。

(* 復習) 離散型確率変数の期待値 :

確率変数 X は x_1, x_2, \dots を取り, X が x_i を取る確率を p_i とする。

すなわち, $P(X = x_i) = p_i$

ある関数 $g(\cdot)$ について, $g(X)$ の期待値は次のように計算される。

$$E(g(X)) = \sum_i g(x_i)p_i$$

$g(\cdot)$ は $g(X) = X$ や $g(X) = (X - \mu)^2$ など

(* 復習) ベルヌイ分布 :

確率変数 X は , 確率 p で 1 を取り , 確率 $1 - p$ で 0 を取る

このとき , X の確率関数 $f(x)$ は ,

$$f(x) = p^x(1 - p)^{1-x} \quad x = 0, 1$$

と表される。

平均 $\mathbf{E}(X)$ と分散 $V(X)$ は ,

$$\mu = \mathbf{E}(X) = \sum_{x=0}^1 xf(x) = 0 \times (1 - p) + 1 \times p = p$$

$$\sigma^2 = \mathbf{V}(X) = \mathbf{E}((X - \mu)^2) = \sum_{x=0}^1 (x - \mu)^2 f(x) = (0 - p)^2(1 - p) + (1 - p)^2 p = p(1 - p)$$

となる。

Y_i の取る値は 0 か 1 のどちらかなので , $P(Y_i = 0) + P(Y_i = 1) = 1$ となる。

$P(Y_i = 1) = p$ とすると , $P(Y_i = 0) = 1 - p$ となる。

したがって , $E(Y_i) = 0 \times P(Y_i = 0) + 1 \times P(Y_i = 1) = P(Y_i = 1) = p$ となり , 確率 p に等しいので , $0 < E(Y_i) < 1$ となる。

確率を表すということは , $E(Y_i)$ が , ゼロより小さくなる , 1 より大きくなることはあり得ない。

$E(Y_i)$ を $\alpha + \beta X_i$ とするのは不適切である。

誤差項 u_i について： Y_i の期待値は，

$$E(Y_i) = P(Y_i = 1) = p$$

なので，

$$Y_i = P(Y_i = 1) + u_i$$

と誤差項を加えて書き換えることができる。

$u_i = Y_i - P(Y_i = 1)$ なので， u_i は確率 $P(Y_i = 1)$ で $1 - P(Y_i = 1)$ の値を取るか，確率 $P(Y_i = 0) = 1 - P(Y_i = 1)$ で $-P(Y_i = 1)$ の値を取ることになる。

すなわち，誤差項 u_i も離散型確率変数となる。

推定方法について： $P(Y_i = 1)$ は分布関数 $F(\cdot)$ に関連付けられて、説明変数 X_i の関数として、 $F(\alpha + \beta X_i)$ と表す。

$$P(Y_i = 1) = F(\alpha + \beta X_i)$$

多変数の場合は、

$$P(Y_i = 1) = F(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})$$

となる。

$F(\cdot)$ には、標準正規分布やロジスティック分布が使われる。

$F(\cdot)$ に標準正規分布が使われた場合はプロビット・モデル (**probit model**) と呼ばれ、ロジスティック分布が使われた場合はロジット・モデル (**logit model**) と呼ばれる。

標準正規分布：

$$\text{密度関数： } f(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right)$$

$$\text{分布関数： } F(x) = \int_{-\infty}^x (2\pi)^{-1/2} \exp\left(-\frac{1}{2}z^2\right) dz$$

ロジステック分布：

$$\text{密度関数： } f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}$$

$$\text{分布関数： } F(x) = \frac{1}{1 + \exp(-x)}$$

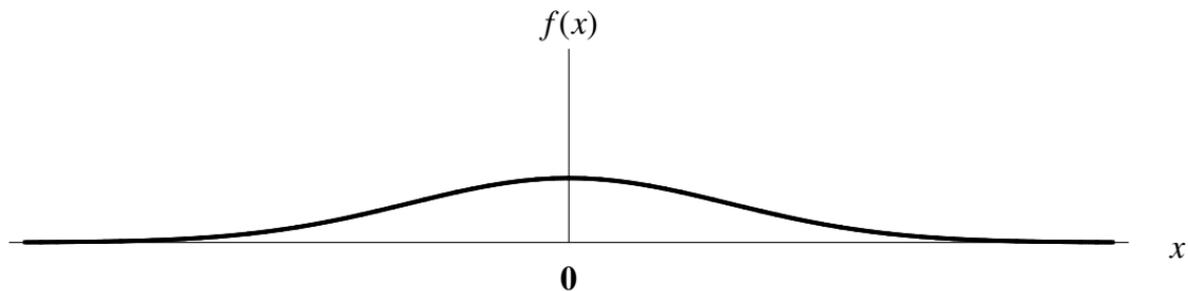
$F(\cdot)$ に他の分布関数を用いてもよい。

(* 復習) 密度関数と分布関数 :

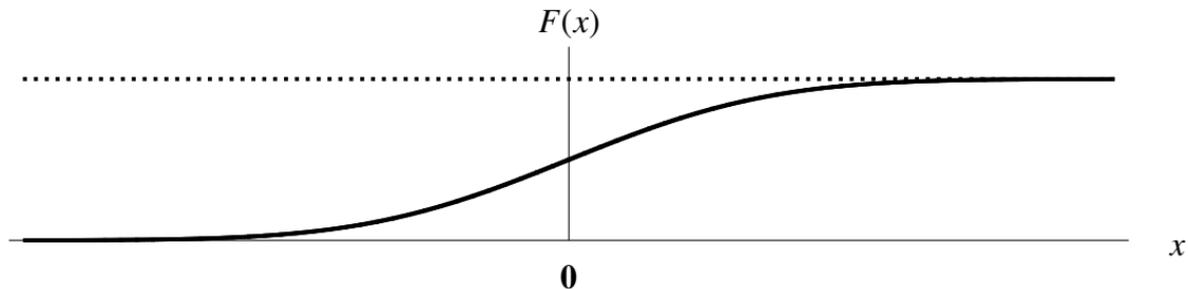
$$\frac{dF(x)}{dx} = f(x)$$

標準正規分布のケース :

密度関数:



分布関数:



最尤法: Y_i の確率関数は,

$$f(Y_i) = F(\alpha + \beta X_i)^{Y_i} (1 - F(\alpha + \beta X_i))^{1 - Y_i} \equiv F_i^{Y_i} (1 - F_i)^{1 - Y_i} \quad Y_i = 0, 1$$

となる。 $F_i = F(\alpha + \beta X_i)$ としている。

Y_1, Y_2, \dots, Y_n の結合確率関数 (同時確率関数) は,

$$f(Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^n f(Y_i) = \prod_{i=1}^n F_i^{Y_i} (1 - F_i)^{1 - Y_i} \equiv l(\alpha, \beta)$$

となる。

尤度関数 $l(\alpha, \beta)$ を α, β について最大にする $\hat{\alpha}, \hat{\beta}$ を求めることになる。 \Rightarrow 最尤法

? probit y const x

モデル 2: プロビット・モデル, 観測: 1-6

従属変数: y

標準誤差はヘッシアン (Hessian) に基づく

	係数	標準誤差	z	限界効果

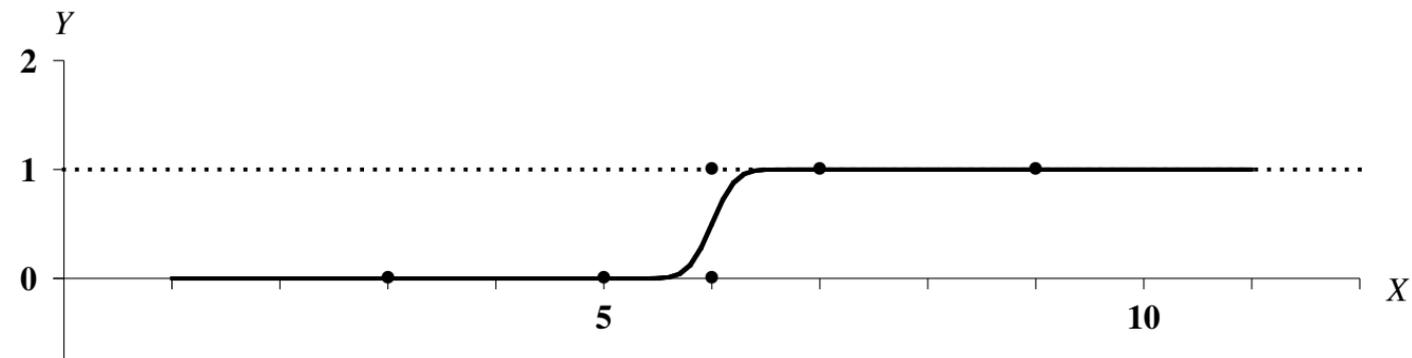
const	-34.8819	13019.0	-0.002679	
x	5.81364	2169.83	0.002679	2.31931
Mean dependent var	0.500000	S.D. dependent var	0.547723	
McFadden R-squared	0.666667	Adjusted R-squared	0.185768	
Log-likelihood	-1.386294	Akaike criterion	6.772589	
Schwarz criterion	6.356108	Hannan-Quinn	5.105381	

「正しく予測された」ケース数 = 5 (83.3%)

$f(\beta'x)$ (説明変数の平均における) = 0.399

尤度比検定: カイ二乗 (1) = 5.54518 [0.0185]

	予測値	
	0	1
実績値 0	3	0
1	1	2



太線は，説明変数 X を与えたもとでの Y の予測値，すなわち，今までの記号では $\hat{Y}_i = F(\hat{\alpha} + \hat{\beta}X_i)$ を表す。

限界効果 2.31931 の意味： 限界係数（太線の接線の傾き）：

$$\frac{d\hat{Y}_i}{dX_i} = \frac{dF(\hat{\alpha} + \hat{\beta}X_i)}{dX_i} = f(\hat{\alpha} + \hat{\beta}X_i)\hat{\beta}$$

となり， X_i の値に依存する。

限界効果： 2.31931 = $\frac{dF(\hat{\alpha} + \hat{\beta})\bar{X}}{d\bar{X}} = f(\hat{\alpha} + \hat{\beta}\bar{X})\hat{\beta}$ と説明変数の平均値で評価する。

$f(\cdot)$ ， $F(\cdot)$ は標準正規分布，または，ロジスティック分布，または，他の分布

上の推定では標準正規分布，次の推定ではロジスティック分布

? logit y const x

モデル 3: ロジット・モデル, 観測: 1-6

従属変数: y

標準誤差はヘッシアン (Hessian) に基づく

	係数	標準誤差	z	限界効果

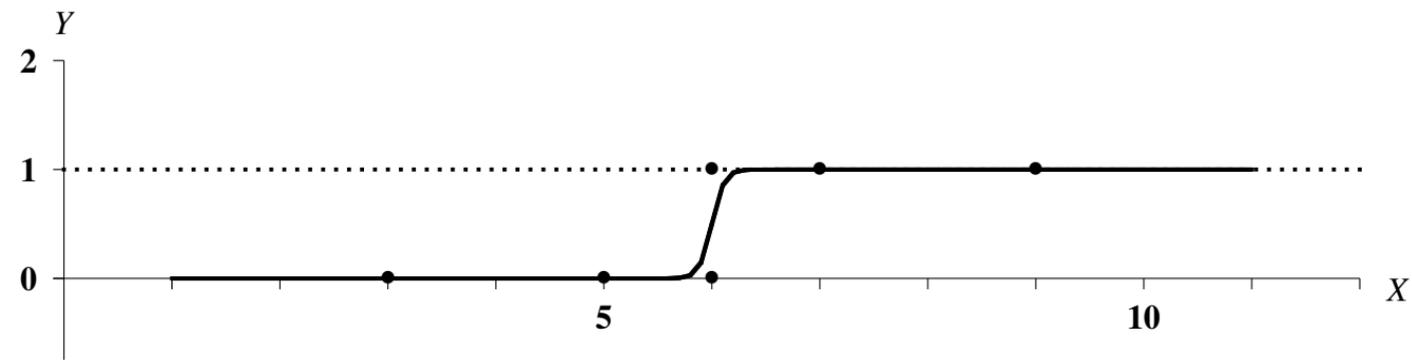
const	-106.179	29537.1	-0.003595	
x	17.6964	4922.86	0.003595	4.42411
Mean dependent var	0.500000	S.D. dependent var	0.547723	
McFadden R-squared	0.666667	Adjusted R-squared	0.185768	
Log-likelihood	-1.386294	Akaike criterion	6.772589	
Schwarz criterion	6.356108	Hannan-Quinn	5.105381	

「正しく予測された」ケース数 = 5 (83.3%)

$f(\beta'x)$ (説明変数の平均における) = 0.250

尤度比検定: カイ二乗 (1) = 5.54518 [0.0185]

	予測値	
	0	1
実績値 0	3	0
1	1	2



● 定期試験

1. 日時： **2021年2月2日 AM8:50~10:20 (1限)**
2. 場所： **自宅・大学など** どこからでも構いませんが、ネットワーク環境が良好な場所 (<--- **持ち帰り試験**)
3. **AM8:30~9:20** の間に、試験問題を [ここ](#) か [ここ](#) (または、[ここ](#) か [ここ](#) <--- ミラー・サーバー) からダウンロードしてください (4箇所のファイルはどれも同じ内容)。
(*) AM8:30から試験問題配布開始 ---> ネットワーク事情を考慮して20分前から
AM9:20までダウンロード可能 ---> 遅刻は試験開始30分まで
4. 教科書・参考書・講義ノートなど何を見ても構いません。特に、統計表・分布表などは、試験時間中に見れるように各自用意しておいてください。
5. 答案作成後、スキャナーでスキャンする、iphoneなどで写真を撮る、などでpdfファイル、または、画像ファイルを作ってください。
添付ファイルは複数になっても構いませんが、ページ番号を書くなどして、ファイルの順番がわかるようにして下さい。
答案は **AM10:30** (10分間の猶予) までに **谷崎宛 (tanizaki [at] econ.osaka-u.ac.jp)** に **メール添付** で送ってください。 **「練習問題 (課題)」の提出先とは異なります！！**
答案が添付されているかどうかを、十分に確認してからメールしてください。
6. *****重要事項*****
 - 答案には **名前** と **学籍番号** を必ず書いてください。氏名・学籍番号がないとプリンタに印刷すると誰の答案か分からなくなります (氏名・学籍番号がないと採点しません)。
 - 答案は必ず **メール添付** で送ってください。 クラウドなどは使わないようにしてください。
メールの **Subject** (件名) には必ず「**計量**」の単語を含めてください。Subject (件名) で仕分けをしています。「計量」という単語が Subject になければ、「ごみ箱」に入る可能性が大きくなります。
 - ネットワークの通信量を減らすために、答案のファイルサイズは**1MB** 以下にしてください。この件は厳格にするつもりはありませんが、できるだけファイルサイズを小さくしてください。
ただし、答案が読めるような画質にして下さい。

画像のファイルサイズを減らすためには、例えば、[IrfanView \(https://www.irfanview.com/\)](https://www.irfanview.com/) というソフトを使えば、ファイルサイズを小さくできます。前もって、ファイルサイズの調整を練習しておいてください。
ファイルサイズの比較： [BMP ファイル](#) (file size: 6,221KB), [JPEG ファイル](#) (file size: 838KB), [JPEG file](#) (file size: 90KB <--- 3つのファイルの中では最小)
このように、JPEG ファイルの大きさを調整することができます。
 - 2度、3度と答案を送ってこないようにして下さい。 **複数回、答案を送ってきた場合は最初の答案を採点します。** 通常の試験では、一旦提出すれば、修正は認められません。
7. 試験時間中に質問 (トラブルを含む) があれば **tanizaki [at] econ.osaka-u.ac.jp** にメールしてください。
この場合も、**Subject** (件名) には必ず「**計量**」の単語を含めてください。
8. 当日、どのようにしても試験問題をダウンロードできない場合は、**AM8:50 から AM9:20 まで**の間にメールしてください。
試験開始後、メール添付で試験問題を送ることにします。
9. これでも再試験が必要な場合は個別に連絡してください。
また、当日試験が受けられない場合、欠席事由によっては追試験が認められる場合があります。
追試験の件は教務係に問い合わせてください (再試験は追試験とは異なります)。

● 成績評価

1. 「4年生だから」という理由で、救済することはありません。特別にレポートを課して救済するという事もあります。