第４講　質的（ダミー）従属変数[1]

従属（被説明）変数が非連続な（離散）値をとる場合を考えてみよう。例えば、

1. 自動車を保有するか？
2. 主婦がパートで働くか？
3. 結婚するか？
4. 就業しているか？

などの結果は２つの選択肢から１つを選ぶことになる。注意すべきは、自動車を保有すると１，保有しなければ０というように数字で表すことはできるが、数字の大きさ自体には意味がないということである。

具体的な例として、結婚するかどうか (Y) を所得あるいは年齢 (X) で説明するケースを考えてみよう。従属変数を

$$\begin{array}{ll} Y=0 & \text{結婚しない} \\ Y=1 & \text{結婚する} \end{array}$$

と定義すれば、選択の結果はダミー変数で表すことができる。このような関係を推定する最も簡単な方法は通常の最小２乗法を使うことである。もう一つの方法は、選択の背後に結婚願望という直接観測できない潜在変数（latent variable）があり、観測値は潜在変数についてのシグナルであると考えるやり方である。この場合ロジット（logit）モデルやプロビット（probit）モデルと呼ばれる推計方法を用いる。

4.1　線形確率モデル（the linear probability model）

線形回帰モデルは説明変数の線形関数によってＹの期待値が決まると仮定し、**線形確率モデル**と呼ばれる。

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, \cdots, n \tag{4.1}$$

X_i は所得（年齢）、u_i は誤差項である。Y_i の条件付き期待値は X_i の線形関数となる。

$$E(Y_i \mid X_i) = \alpha + \beta X_i \tag{4.2}$$

$X = X_i$ の時 $Y_i = 1$ となる確率を P_i とすれば、Y_i の分布は、

Y_i	確率
0	$1 - P_i$
1	P_i

[1] この節の基本的参考文献は Maddala(1983) である。より平易な文献には浅野・中村 (2000、第 10 章) と Maddala(1988) がある。

となり、Y_i の期待値は保有確率 P_i に他ならない。

$$E(Y_i \mid X_i) = 0(1 - P_i) + 1 \cdot P_i = P_i \tag{4.3}$$

従って、条件付き期待値は保有確率で X の線形関数となる。

$$E(Y_i \mid X_i) = \alpha + \beta X_i = P_i \tag{4.4}$$

誤差項 u_i の分布は Y_i の分布から、以下のように与えられる。

Y_i	$u_i\,(= Y_i - P_i)$	確率
0	$-P_i$	$1 - P_i$
1	$1 - P_i$	P_i

u_i は $Y_i = 1$ の時 $1 - P_i$、$Y_i = 0$ の時 $-P_i$ となり、2つの値しかとらない。したがって u_i の分布は正規分布ではない。

分散は、

$$\begin{aligned} Var(u_i \mid X_i) &= E(u_i^2) - E(u_i)^2 \\ &= (-P_i)^2(1 - P_i) + (1 - P_i)^2 P_i \\ &= P_i - P_i^2 \\ &= P_i(1 - P_i) \end{aligned} \tag{4.5}$$

となり、X_i の値とともに変化する。

線形確率モデルの特徴は次の通りである。

1. 保有確率が説明変数の線形関数で表されるため、$(0, 1)$ の範囲をはずれることもある。
2. 誤差項（u_i）は**正規分布には従わない**。
3. 誤差項の分散は $P_i(1-P_i)$ で、X_i とともに変化し**不均一分散（heteroscedasticity）**となる。

誤差項が正規分布に従わず、不均一分散であれば、最小2乗推定量は有効ではない（inefficient）。この場合は次のような加重最小2乗推定を行う。

1. (4.1) 式を OLS 推定する。
2. ウェイト $w_i = \sqrt{\hat{y}_i(1 - \hat{y}_i)}$ を計算する。
3. $y_i / w_i = \alpha + \beta x_i / w_i$ について WLS 推定を行う。

1960-70 年代には線形確率モデルが多用された。Meyer and Pifer (1970) は、銀行倒産について分析したし、Altman (1968) は製造業企業の倒産確率を判別分析（discriminant analysis）を用いて行った。その際、倒産企業と継続企業のサンプルをとり、倒産確率を線形確率モデルや線形判別関数を推計して求めていた。

線形判別関数

n 人のサンプルに対して、k 個説明変数があり、n 人のうち n_1 人がグループ π_1 に属し、n_2 人がグループ π_2 に属する（$n = n_1 + n_2$）ことがわかっているとき、k 個の説明変数による線形判別関数を推計したい。

例えば、ローンの応募者のうち、n_1 人にはローンが与えられ、n_2 人は拒否されたとする。また応募者の社会経済的な属性情報（x_i）はわかっている。

次のような線形関数を定義しよう。

$$Z = \lambda_0 + \sum_{i=1}^{k} \lambda_i x_i$$

２つのグループを判別するための最適な方法は、λ_i が次の比率 η を最小化する場合である。

$$\bar{\eta} = \frac{Z \text{ のグループ間分散（between-group variance）}}{Z \text{ のグループ内分散 (within-group variance)}}$$

このようにして、パラメータ λ_i を決めれば、それを用いて応募者がどちらのグループに入るかを予測できる。

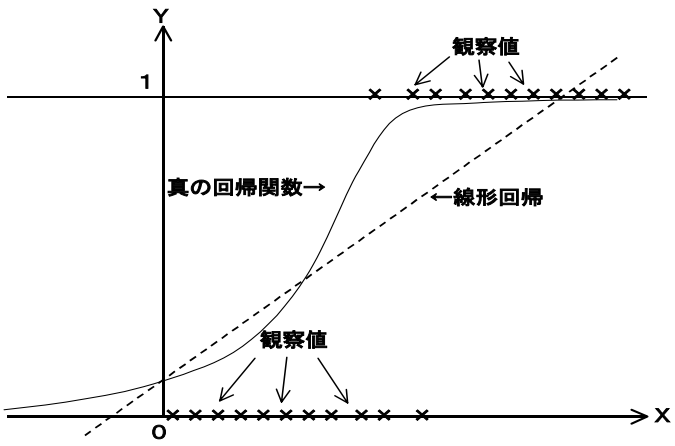
4.2 非線形確率モデル

線形確率モデルの欠陥を避けるには、確率は X の（増加）関数で必ずゼロから１の間にはいるような関数形のモデルを考える必要がある。このような確率を表す関数としては累積密度関数（cumulative density function）が思い浮かぶ。このような累積密度関数の形状を与えて非線形回帰を行えば、X と保有確率の関係を決めているパラメータが推定できる。保有確率を累積密度関数で表す背後には**潜在変数（latent variable）モデル**と呼ばれる選択行動モデルがある。潜在変数モデルでは線形確率モデルとは異なり、所得は結婚の、見えない「願望」に影響し、それを通じて確率が決まると考える。潜在変数（Y^*）は (4.6) のように、X についての線形関数と誤差要因 ε の和と仮定し、

$$Y^* = \alpha + \beta X + \varepsilon \tag{4.6}$$

潜在変数が臨海値を越えると Y の値が１（結婚する）になるとする。

図4.1 ダミー従属変数と線形回帰



結婚（$Y = 1$）の条件は、

$$\begin{aligned} Y^* > 0 &\iff \alpha + \beta X + \varepsilon > 0 \\ &\iff \varepsilon > -\alpha - \beta X \end{aligned} \tag{4.7}$$

となり、結婚確率は誤差項 ε についての確率と書ける。以下では、一般性を失うことなく、ε の分散を 1 とする。i 番目のサンプルの結婚確率は $F(\cdot)$ で ε の累積密度関数を表せば、

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i^* > 0) = \Pr(\varepsilon_i > -\alpha - \beta X_i) \\ &= 1 - F(-\alpha - \beta X_i) \end{aligned} \tag{4.8}$$

また、非婚確率は、

$$\begin{aligned} \Pr(Y_i = 0) &= \Pr(Y_i^* \leq 0) = \Pr(\varepsilon_i \leq -\alpha - \beta X_i) \\ &= F(-\alpha - \beta X_i) \end{aligned} \tag{4.9}$$

と書ける。

実用されている累積密度関数は、（1）ロジスティック分布と（2）標準正規分布の 2 つであり、前者を用いたモデルを**ロジットモデル**、後者を用いたモデルを**プロビットモデル**と呼ぶ。

4.2.1 ロジットモデル（グループデータからの推定）

ロジットモデルでは、結婚確率を次式で表す。

$$P_i = E(Y_i = 1 \mid X_i) = \frac{1}{1 + \exp(-(\alpha + \beta X_i))} \tag{4.10}$$

ここで $Z_i = \alpha + \beta X_i$ とおき、Z_i は X_i に制限がなければマイナス無限大からプラス無限大の値をとり、結婚確率は Z_i がプラス無限大の時 1、マイナス無限大の時 0 となる。

非婚確率は、

$$1 - P_i = \frac{\exp(-Z_i)}{1 + \exp(-Z_i)} \tag{4.11}$$

(4.10) と (4.11) から結婚、非婚の確率の比（オッズ）は、

$$\frac{P_i}{1 - P_i} = \exp(Z_i) \tag{4.12}$$

対数をとり、

$$\ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = \alpha + \beta X_i \tag{4.13}$$

この式の左辺は**ロジット**と呼ばれる。もし結婚確率が観測可能なら、ロジットを X に回帰して (α, β) を推定できる。

個人データからは結婚確率は観測できないが、同じ所得（年齢）の人が多数いる場合、標本割合 $(\hat{P})$ は結婚確率の推定値と考えられる。所得（年齢）が i 番目の値 (X_i) である個人が全部で N_i 人、そのうち既婚者が n_i 人とすれば、グループ（コーホート）の $\hat{P}$ は n_i/N_i となる。

標本の割合と所得（年齢）の関係は誤差項 ε を加えて

$$\ln\left(\frac{\hat{P}_i}{1 - \hat{P}_i}\right) = Z_i = \alpha + \beta X_i + \varepsilon_i \tag{4.14}$$

と表せ、パラメータに関して線型モデルが与えられる。ただし、$\hat{P}_i$ が0か1の 場合は (4.14) 式の左辺は定義できなくなる。

グループデータの場合、標本が大きければグループ i の誤差項 ε_i は近似的に正規分布に従い、分数は $1/(N_i\hat{P}_i(1-P_i))$ となる。

$$\varepsilon_i \sim N\left[0, \frac{1}{N_iP_i(1-P_i)}\right] = N(0, \sigma_i^2) \tag{4.15}$$

それぞれのグループによって誤差項の分散は異なり得るので、不均一となり最小2乗推定量は有効（efficient）でなくなる。このとき、漸近的有効推定量は分散の平方根の逆数で加重した最小2乗法（weighted least square）で与えられる。分散 σ_i^2 は未知だが、標本割合 $\hat{P}_i$ より、

$$\hat{\sigma}_i^2 = \frac{1}{N_i\hat{P}_i(1-\hat{P}_i)} \tag{4.16}$$

で推定値が求まる。

ロジットモデルの推定手続き

1. 各所得（年齢）レベルに応じて結婚確率（$\hat{P}_i$）を求める。
2. ロジット　$L_i = \ln(\hat{P}_i/(1-P_i))$ を計算する。
3. 不均一分散を除去するための加重を行う。

$$\begin{aligned} w_i^{\frac{1}{2}}L_i &= w_i^{\frac{1}{2}}\alpha + w_i^{\frac{1}{2}}X_i\beta + w_i^{\frac{1}{2}}\varepsilon_i \\ \Longrightarrow L_i^{\uparrow} &= \alpha w_i^{\frac{1}{2}} + \beta X_i^* + u_i \end{aligned} \tag{4.17}$$

ここで $w_i = N_i\hat{P}_i(1-\hat{P}_i), L_i^* = w_i^{\frac{1}{2}}L_i, X_i^* = w_i^{\frac{1}{2}}X_i, u_i = w_i^{\frac{1}{2}}\varepsilon_i$ である。

4. L^* を $w^{\frac{1}{2}}$、X^* に最小2乗回帰する。**定数項は使わない**。

4.2.2　プロビットモデル（個別データからの推定）

個人レベルの既婚・未婚と所得（年齢）のデータがある場合、グループデータと違い、加重最小2乗法は使えない。まず所得（年齢）グループごとの結婚確率を求める必要はない。強引にグループの確率を求めてロジット推定を行うことは可能だが、そのような平均化は元データの情報を失うことになり、推定値の精度は下がる。またある所得（年齢）に対応する個人の従属変数の値は0か1であり、対数オッズは定義できない。個人データすべての情報を利用して推定する場合には**最尤法（maximum likelihood）**を用いる（補論3－A参照）。

個人データに対応する**尤度**を求める。観察値が（$Y_1 = 1, \quad Y_2 = 0, \quad Y_3 = 0$）となっていれば、(4.6) の誤差項 ε_i が独立なら**尤度関数**は確率の積、

$$\Pr(Y_1 = 1) \cdot \Pr(Y_2 = 0) \cdot \Pr(Y_3 = 0) \cdot \cdots$$

となる。一般的に尤度関数は (3.8)、(3.9) の表現を用いると、

$$L(\alpha, \beta) = \prod_{Y_i=1}(1 - F(-\alpha - \beta X_i)) \prod_{Y_i=0} F(-\alpha - \beta X_i) \tag{4.18}$$

と表せる。ここで $\prod_{Y_i=1(0)}$ は $Y_i = 1(0)$ となる観察値について積をとる操作を表している。(4.18) は次のように表せる。

$$L(\alpha, \beta) = \pi_i \left\{1 - F(-\alpha - \beta X_i)\right\}^{Y_i} F(-\alpha - \beta X_i)^{(1-Y_i)} \tag{4.19}$$

この対数尤度は、

$$\ln L(\alpha, \beta) = \Sigma_i \left\{Y_i \ln \left\{1 - F(-\alpha - \beta X_i)\right\} + (1 - Y_i) \ln F(-\alpha - \beta X_i)\right\} \tag{4.20}$$

となる。パラメータの推定には $F(\cdot)$ の関数形（ε の分布）を特定化した **最尤法**が使われる。もっともよく使われるのが**標準正規分布**を仮定したもので**プロビット（Probit）モデル**と呼ばれる。

係数の解釈

説明変数 X の値に対応する結婚確率と、X の**結婚確率**への**限界効果**の大きさは次のように表せる。

$$\text{結婚確率} = 1 - F(-\alpha - \beta X_i)$$

限界効果は線形回帰とは異なり、β で表されるのではなく、結婚確率を X で微分して次のように求められる。

$$\frac{\partial P_2(Y_i = 1)}{\partial X} = \frac{\partial (1 - F(-\alpha - \beta X))}{\partial X} = \beta f(-\alpha - \beta X)$$

ここで $f(\cdot)$ は累積密度関数 F を微分したものであり、確率密度関数を $-\alpha - \beta X$ で評価した値である。

Y_i の値が 0 か 1 しかとらない場合に決定係数（R^2）のような当てはまりの良さを示す指標を作るのは難しい。

Effron（1978）は、線形確率モデルの場合、推計値 $\hat{Y}_i$ が求められるので R^2- タイプの指標を考えることができると論じた。すなわち一般的に決定係数は次のように定義される。

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}\right]$$

ここで $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ は定義できる。被説明変数が二項選択の場合には

$$\Sigma (y_i - \bar{y}_i)^2 = \Sigma y_i^2 - n\bar{y}_i^2 = n_1 - n(\frac{n_1}{n})^2 = \frac{n_1 n_2}{n}$$

Effron の R^2- タイプ指標は

$$R^2 = 1 - \frac{n}{n_1 n_2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.21}$$

と定義される。

Amemiya（1981）は次のような定式化を提示した。

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i (1 - \hat{y}_i)} \tag{4.22}$$

これは誤差 2 乗和を分散の逆数で加重したものである。

係数についての検定

β の最尤推定量は**漸近的に正規分布**に従う。標本数が大きければ $\hat{\beta}$ の漸近分布を使い通常の仮説検定が行える。係数の一部についての検定は線形回帰における F 検定と同様に制約なしと制約付きの 2 つの最尤法による推定結果を比べるという方法が使える。もちろん、制約なしの尤度は制約付きの場合より大きくなり、差が大きければ制約は正しくないことを示唆する。検定に使われる統計量は制約なし（Unrestricted）の尤度を L_U、制約付き（Restricted）の尤度を L_R とし、尤度比（L_U/L_R）の対数を 2 倍したものが用いられる。これは尤度比検定と呼ばれ、近似的に自由度＝制約の数のカイ 2 乗分布に従う。

$$2(\ln L_U - \ln L_R) \sim X^2(p) \quad p \text{ は制約数} \tag{4.23}$$

あてはまりの尺度

質的従属変数モデルのあてはまり尺度として最も多く使われているのが**尤度比インデックス**である。一般に複数の説明変数があるケースで潜在変数は、

$$Y^* = \beta_0 + \mathbf{x}\beta + \varepsilon$$

と書けるが、もし説明変数 $\mathbf{x}$ が結婚確率に影響しないのであれば、定数項以外の係数はすべてゼロ（$\beta = \mathbf{0}$）のはずである。$\beta = \mathbf{0}$ という制約をおいたときの β_0 の推定値は、全体の結婚確率を P とすれば、$P = 1 - F(-\beta_0)$ が成立する。対数尤度は、

$$\ln L_0 = n\left[P \ln P + (1-P)\ln(1-P)\right]$$

となる。これが制約付きの対数尤度である。β に制約をつけない場合の対数尤度を $\ln L_U$ という。

ここで次の尤度比インデックス（Likelihood Ratio Index）を定義する。[2]

$$LRI = 1 - \frac{\ln L_U}{\ln L_0} \tag{4.24}$$

もしすべての $\beta = \mathbf{0}$ ならば、$\ln L_U = \ln L_0$ となり $LRI = 0$ となる。すべての既婚者の結婚確率を 1、未婚者の結婚確率を 0 と完璧に予測するなら $L_U = 1(\ln L_U = 0)$ となり $LRI = 1$ となる。

尤度比インデックスに加え、予測の**的中率**（proportion of correct predictions）も 用いられることがある。

X の値に応じて予測された結婚確率を計算し、もし結婚確率が 0.5 以上なら予測は結婚、0.5 未満なら予測は未婚として扱う。すなわち、予測値 $\hat{y}_i^*$ は次のルールに従って 0 か 1 の値をとる。

$$\hat{y}_i^* = \begin{cases} 1 & \hat{y}_i > 0.5 \text{ の場合} \\ 0 & \hat{y}_i < 0.5 \text{ の場合} \end{cases}$$

的中率は次のように定義される。

$$\text{的中率} = \frac{\text{正しい予測値の数}}{\text{全観測値の数}}$$

[2] この定義は McFadden（1974）による。

例えば、次のような表を考えてみよう。

		予測値		
		0	1	合計
観察値	0	36	4	40
	1	8	52	60
	合計	44	56	100

的中率88/100

100 個の観察値のうち 60 が既婚、40 が未婚であるときに、予測値は 56 が既婚、44 が未婚であるとすると、この予測の的中率は 88 ／ 100 となる。この的中率が高いことは必ずしもモデルのあてはまりがよいことを意味しているわけではない。標本の 90 %が既婚の時、説明変数の値に関わりなくすべての標本につき確率 1 で結婚するとする推定値の「的中率」は 90 %であるが、これではモデルの説明力とは関係がないことは明らかであろう。

References

[1] 浅野晳・中村二朗（2000） 『計量経済学』、有斐閣。

[2] Altman, E.I. (1968) "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, September, pp.589-609.

[3] Amemiya, T. (1981) "Qualitative Respponse Models: A Survey," *Journal of Economic Literature*, December, pp.483-536.

[4] Cox, D.R. (1970) *Analysis of Binary Data*, London: Metheun.

[5] Effron, B. (1978) "Regression and ANOVA with Zero-One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, May, pp. 113-21.

[6] Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

[7] Maddala, G.S. (1988) *Introduction to Econometrics*, New Jersey: Prentice-Hall, Inc.

[8] McFadden, D. (1974) "The Measurement of Urban Travel Demand," *Journal of Public Economics*, pp, 303-328.

[9] Meyer, P.A. and Pifer, H.W. (1970) "Prediction of Bank Failures," *Journal of Finance*, pp. 853-868.