ミクロデータのモデル分析
－ LDV(制限従属変数) モデルに注目して－

山下裕歩（京都大学大学院経済学研究科博士後期課程）

1　はじめに

本稿の目的は、近年わが国で公開が進んでいるミクロデータが、どのような方法により経済学上の研究に利用されているのかを概観することである。ミクロデータは個別の家計や企業の行動・選択を表現したものであり、本質的にそのとる値が制限されている場合が多い。これには様々な例が考えられる。

第一の例として、行動結果は本来数値的に表わされるものではないが、その行動結果に対応させて便宜的にある数値が与えられるという意味で制限されている場合が考えられる。例えば、ある個人が労働するか、あるいはしないかの選択が考えられる。労働している場合には従属変数に 1 という数値を与え、労働していない場合には従属変数に 0 という数値が与えられるが、これは便宜上このような数値を与えているに過ぎない。このような数値は質的データと呼ばれる。

第二の例として、ある行動結果を表わす値が連続的な値を取ることができず、非負の整数値しかとり得ないという意味で制限されている場合もある。例えば、ある夫婦が子供を何人持つかを選択する問題を考えてみると、個別の家計にとって非負の整数以外の値はあり得ない。しかしここでは、実際の数値は何人子供を持っているかという人数を表わしており、数値自体が意味を持っている。

第三の例として、ある行動結果は連続的な数値をとり得るが、それはある一定の範囲に限られ、その範囲以外ではある特定の値（閾値）に限定されるという意味で制限されている場合である。例えば、耐久消費財に対する家計の支出はゼロ以下にはなり得ず、データの少なくない割合にゼロという値が現れる。この場合数値は実際の支出額をあらわしている。

あるいはこれらの例とは異なって、データの観測可能性の問題からデータのとり得る値が制限されている場合もある。

第一の例として、ある値以上（以下）に関してはそれ以上（以下）であるということしか分からず、ある特定の値が真の値の代わりに与えられる場合がある。例えば、アンケートで耐久消費財への支出額の回答に対し上限が設定されており、最高支出額が ' 500 万円以上 ' というような場合である。この場合、500 万円以上である家計の真の支出額は観測できない。また、特にリサンプリングデータの場合、個人情報保護の観点からサンプル個人の特定を防ぐために、ある閾値を越える所得水準や支出額などについては正確な数値が分からなくなっている場合も多い。

第二の例として、ある値以上（以下）はそもそも観測されていない場合がある。すなわち、観測されているのがそもそも母集団の一部分であるという場合である。例えば、賃金オファー関数を推計する際、就労者のみを調査対象とする場合が考えられる。

以上のようにミクロデータは様々な形で制限される。重要なことは分析対象であるデータがどのような意味でどのように制限されているのかを考慮した上で利用するモデルを決定しなければならないということである。この「従属変数が制限を受けている」という情報を無視し、単に得られた標本について線形回帰モデルを推定すれば、当てはまりが悪くなるし、係数の解釈が困難になったり、あるいはより重要なことであるが、推定量に望まれる一致性が失われたりするからである。

本稿では、従属変数が何らかの制限を受けるモデル (Limited Dependent Variable Model) にのみ焦点を当てることにする[1]。

次章では、どのようなタイプのデータに対してどのような回帰モデルが用いられるかに重点をおきながら各モデルを概観する。表 1 には、各モデルがまとめてあるので参照されたい。

2 LDV Models

本節では従属変数がそのとり得る値について何らかの制約を受けている場合にどのような回帰モデルが適しているかを考察する。ただし、ここで取り上げるモデルは網羅的にすべての制限従属変数モデルを扱っているわけではない。データの性質とモデルとの対応関係に重点をおきながら各モデルを簡単に説明することにしたい。

2.1 節、2.2 節では従属変数が 0 か 1 のどちらかのみをとる 2 値データである場合を扱う。

2.3 節、2.4 節、2.5 節、2.6 節では従属変数が $0, 1, 2, \cdots, s$ という非負整数値のみをとる離散的データである場合を扱う。これらの節では表面上は同じようなデータを扱っている。しかし、何故このようなデータが得られたかを説明する経済主体の行動原理に対する想定が全く異なっている。これを区別することが重要である。

2.7 節、2.8 節では、従属変数がある範囲では連続的な値をとり得るが、その範囲以外では特定の値しかとり得ない場合を扱う。このような従属変数を扱うモデルは総称してトービット・モデルと呼ばれることが多い[2]。

2.9 節では、従属変数がある範囲では連続的な値をとり得るが、その範囲以外のサンプルは独立変数も含めて観察できない場合を扱う。

最後に、2.10 節では、従属変数によってサンプルが選別されてしまう場合を扱う。これは、内生的標本選別と呼ばれる。

---

[1] discrete dependent variable と limited dependent variable という用語を使い分ける場合もある。この場合、前者は 2.2 節〜2.6 節で言及されるような変数、後者は 2.7 節〜2.10 節で言及されるような変数を表わしている。例えば Greene(1993) ではこのような使い分けをしている。一方、両者をまとめて limited dependent variable と呼ぶ場合もある。例えば Wooldridge(2003) である。本稿では両者をまとめて limited dependent variable と呼ぶことにする。

[2] Wooldridge(2003) は 2.7 節のモデルをトービット・モデル、Greene(1993) では 2.8 節のモデルをトービット・モデルと呼んでいる。Amemiya(1984) では、2.7 節、2.8 節、2.9 節、2.10 節のモデルを全てトービット・モデルと呼んでおり、尤度関数の形状でトービット・モデルを 5 つのタイプに分類している。Amemiya(1984) の分類では、2.7 節、2.8 節、2.9 節のモデルはタイプ 1 のトービット・モデル、2.10 節のモデルはタイプ 2 のトービット・モデルである。

2.1 線型確率モデル

被説明変数 y_i が 0, 1 しかとらない 2 値変数である場合を考える。このとき、線型回帰モデル、

$$y_i = x_i'\beta + u_i \tag{2.1.1}$$

を考える。y_i は 0 か 1 しかとらないので、β_k を x_k が 1 単位増加したときの y_i の変化量と考えることはできない。しかし β_k は強外生性 $E(u_i|\boldsymbol{X}) = 0$ $(i = 1, 2, \cdots, n)$ が満たされることを仮定すると次のように解釈できる。(1) 式の両辺の条件付期待値をとると、

$$E[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i'\beta \tag{2.1.2}$$

を得る。今、y_i は 2 値変数であるから、常に $P[y_i = 1|\boldsymbol{x}_i] = E[y_i|\boldsymbol{x}_i]$ が成り立つ。すなわち、

$$P[y_i = 1|\boldsymbol{x}_i] = \boldsymbol{x}_i'\beta \tag{2.1.3}$$

が成り立つのである。つまり、β_k は x_k が 1 単位増加したときの成功確率の変化を表わすのである。故に、これは線型確率モデルと呼ばれる。

線型確率モデルは、推計が最小 2 乗法により簡単に行えるし、また上に述べたように係数の解釈も明確ある。しかし、一方で次のような欠点を持っている。

まず、第 1 に推定結果で確率が 0 以下になったり 1 以上になったりすることがあげられる。これは推定結果は確率を与えているという解釈上明らかにおかしい。

第 2 に、β_k が x_k に対して一定であることも欠点である。例えば女性が労働するかどうかの説明変数として子供の数を考える場合、一人目の子供と二人目の子供では、労働する確率に与える影響は一人目の子供の方が大きいと考えられるが、線型確率モデルでは、一人目の与える効果も二人目の与える効果も同じである。

また第 3 に、y_i は二項変数であるから分散は、

$$\mathrm{Var}[y_i|\boldsymbol{x}_i] = P(y_i = 1|\boldsymbol{x}_i)[1 - P(y_i = 1|\boldsymbol{x}_i)] = (\boldsymbol{x}_i'\beta)(1 - \boldsymbol{x}_i'\beta) \tag{2.1.4}$$

となる。すなわち、成功確率 $P(y_i = 1|\boldsymbol{x}_i)$ が説明変数 $\boldsymbol{x}_i$ に依存している限り、不均一分散が存在することになる。ただし、不均一分散の問題については、大標本理論に基づき robust standard error で解決可能である。また、実際には OLS standard error と robust standard error はほとんど差がないことが多い[3]。

2.2 プロビット・モデルとロジット・モデル

2.1 節で述べたられたように線型確率モデルにはいくつかの欠点がある。これらの欠点を解消するのがプロビット・モデルとロジット・モデルである。

まず、(2.1.3) 式を次のように修正する。

$$P[y_i = 1|\boldsymbol{x}_i] = G(\boldsymbol{x}_i'\beta) \tag{2.2.1}$$

[3]例えば、Wooldridge(2003),*Introductory Econometrics*,(7.29)(8.37) 式を参照。

ここで、$G(z)$ はどんな実数 z に対しても $0 < G(z) < 1$ を満たす関数である。

プロビット・モデルはこの $G(z)$ を標準正規分布の累積分布関数、

$$G(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\nu^2/2) d\nu \equiv \Phi(z) \tag{2.2.2}$$

で特定化する。

一方、ロジット・モデルはこの $G(z)$ を標準ロジスティック関数、

$$G(z) = \frac{\exp(z)}{1 + \exp(z)} \equiv \Lambda(z) \tag{2.2.3}$$

で特定化する。

このように、$G(z)$ を特定化すれば、(2.2.1) において $0 < P[y_i = 1|x_i] < 1$ となることが保証され、線型確率モデルの欠点が回避される。

ところで、ロジット・モデルとプロビット・モデルは古典的な線型回帰モデルの諸仮定を満たす latent variable model から導き出すことができる。今、観察され得ない変数 (latent variable) を y_i^* とし、y_i と y_i^* は、

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \quad y_i = 1[y_i^* > 0] \tag{2.2.4}$$

で決定されていると考える。ここで $1[\cdot]$ はインディケーター関数 (indicator function) である[4]。また、e は標準正規分布、あるいは標準ロジスティック分布に従うものと仮定する。どちらの場合にしろ e の分布は 0 の左右で対称であり $1 - G(-z) = G(z)$ を満たしている。

さて、$y_i = 1$ となるのは (2.2.4) 式で $y_i^* > 0$ のときであるから、

$$\begin{aligned} P(y_i = 1|\boldsymbol{x}_i) &= P(y_i^* > 0|\boldsymbol{x}_i) = P(e_i > -\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{x}_i) \\ &= 1 - G[-\boldsymbol{x}_i'\boldsymbol{\beta}] = G[\boldsymbol{x}_i'\boldsymbol{\beta}] \end{aligned} \tag{2.2.5}$$

となる。これは (2.2.1) 式と全く同じであり、(2.2.1) 式の背後にはこのような経済主体の行動が想定されているのでる。

線型確率モデルは、OLS により、また場合によっては GLS、FGLS や WLS で推計することができた。しかし、プロビット・モデル、ロジット・モデルではこれらの手法は利用できないので、最尤法が用いられる。

今、大きさ n の無作為標本が得られたとする。既に述べたように、

$$P[y_i = 1|\boldsymbol{x}_i] = G(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.2.6}$$

$$P[y_i = 0|\boldsymbol{x}_i] = 1 - G(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.2.7}$$

であるから、ある標本 i について $\boldsymbol{x}_i$ を条件としたその確率関数 $f(y_i|\boldsymbol{x}_i)$ は、

$$f(y_i|\boldsymbol{x}_i) = [G(\boldsymbol{x}_i'\boldsymbol{\beta})]^{y_i}[1 - G(\boldsymbol{x}_i'\boldsymbol{\beta})]^{1-y_i} \tag{2.2.8}$$

と書ける。従って、標本 i の対数尤度 $l_i(\boldsymbol{\beta})$ は (2.2.8) 式の両辺の自然対数をとって、

$$l_i(\boldsymbol{\beta}) = y_i \log[G(\boldsymbol{x}_i'\boldsymbol{\beta})] + (1 - y_i) \log[1 - G(\boldsymbol{x}_i'\boldsymbol{\beta})] \tag{2.2.9}$$

[4] $y = 1$ if $y^* > 0$, $y = 0$ if $y^* \leq 0$ である。

である。(2.2.9) 式を全ての標本について足すと対数尤度関数が得られる。

$$\mathscr{L} = \sum_{i=1}^{n} l_i(\beta) \tag{2.2.10}$$

この対数尤度を最大化する $\hat{\beta}$ が求める推定量である。プロビット・モデルのときはプロビット推定量、ロジット・モデルのときはロジット推定量と呼ばれる。2 値選択行動モデルで、我々が知りたいのは x_k が $y = 1$ となる条件付確率 $P(y = 1|x)$ に与える影響である。線型確率モデルでは β_k の値がそのまま解釈できたが、ロジット・モデル、プロビット・モデルでは注意が必要である。ロジット・モデル、プロビット・モデルでは β は観察されない変数 y^* への影響を表わしていることになる。$P(y = 1|x)$ に与える影響は、

$$\frac{\partial P(y = 1|x)}{\partial x_k} = \frac{dG(z)}{dz}\frac{dz}{dx_k} = g(x'\beta)\beta_k \tag{2.2.11}$$

で与えられる。ここで $z = x'\beta$ であり、$g(\cdot)$ は分布関数を z で微分したものだから確率密度関数である。この式から分かるように x_k が $P(y = 1|x)$ に与える影響の正負は β_k の正負と同じである。また、この式の値は x に依存するので一定ではない。

2 値データの場合、ロジットとプロビットのどちらを用いるかについて、決定的な優位関係はない[5]。経済学ではプロビット・モデルが自然科学ではロジット・モデルが使われることが多い。

プロビット推定量、ロジット推定量は、一致性、漸近的正規性、漸近的効率性を持つことが知られている。漸近的正規性から、計算された推定値と標準誤差を使って通常の t 検定を行うことができる。

2.3 順序プロビット・モデルと順序ロジット・モデル

2.2 節では 2 値データの取り扱いについて述べたが、3 値以上のデータを得る場合もある。経済主体の反応が 3 値以上の場合、その 3 つ以上の選択肢がどのような関係にあるかに注意しなければならない。本節では選択肢間に順序関係のある場合、2.4 節では選択肢間に順序関係のない場合、2.5 節では選択肢間に順序関係のあるものとないものが混合している場合を扱う。

本節で見る順序反応モデルは 3 つ以上の選択があり、それらがある一つの隠れた要因 (latent variable) によって順序付けられており、選択結果に $0, 1, 2, \cdots$ という数値が与えられている場合に用いられる。このようなデータを通常の線型回帰モデルで分析するのは不適切である。何故なら、線型回帰モデルでは 0 と 1 の差を 1 と 2 の差と同等に扱うことになるが、選択肢に与えられた $0, 1, 2, \cdots$ という数値は便宜上のものであり、これはあくまで経済主体の選択肢に対するランキングを表しているに過ぎないからである。この回答結果はランキングであるというデータの特性を反映するために用いられるのが順序反応プロビット・モデルや順序反応ロジット・モデルである。

本節では、選択肢が 3 つの 3 値モデルに限定するが、任意の選択肢数への拡張は容易である。次のように、ある政策の是非に対するアンケートの結果として、

[5] Amemiya(1981) は、両者の相違と使い分けについて議論している。また、2.4 節で見る多項反応モデルの場合は両者には大きな相違がある。

1. $y_i = 0,\ y_i^* \leq 0$：反対
2. $y_i = 1,\ 0 < y_i^* \leq \alpha$：どちらともいえない
3. $y_i = 2,\ \alpha < y_i^*$：賛成

となるような場合を考えよう。α は β と共に推計される未知パラメータである。ここで y_i^* は latent variable（観察されない変数）であり、

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{2.3.1}$$

によって決定されている変数である。原理的にはアンケートの回答者はこの y^* を報告することもできたのであるが、アンケートが選択式であるため自分の y^* に最も近い回答を行ったものと考えられる。e_i に正規分布を仮定するのが順序反応プロビット・モデル、e_i にロジスティック分布を仮定するのが順序反応ロジット・モデルである。

さて、$y_i = 0$ となるのは、(2.3.1) 式で $y_i^* \leq 0$ のときであるから、

$$\begin{aligned} P(y_i = 0|\boldsymbol{x}_i) &= P(y_i^* \leq 0|\boldsymbol{x}_i) = P(e_i \leq -\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{x}_i) \\ &= G(-\boldsymbol{x}_i'\boldsymbol{\beta}) \end{aligned} \tag{2.3.2}$$

が成り立つ[6]。同様に、

$$P(y_i = 1|\boldsymbol{x}_i) = G(\alpha - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(-\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.3.3}$$

$$P(y_i = 2|\boldsymbol{x}_i) = 1 - G(\alpha - \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.3.4}$$

が成り立つ。従って、サンプル i の対数尤度関数は、

$$l_i(\alpha, \boldsymbol{\beta}) = \begin{cases} \log[G(-\boldsymbol{x}_i'\boldsymbol{\beta})] & \text{if } y_i = 0 \\ \log[G(\alpha - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(-\boldsymbol{x}_i'\boldsymbol{\beta})] & \text{if } y_i = 1 \\ \log[1 - G(\alpha - \boldsymbol{x}_i'\boldsymbol{\beta})] & \text{if } y_i = 2 \end{cases} \tag{2.3.5}$$

である。従って標本全体の対数尤度関数は、$\mathscr{L} = \Sigma_i l_i(\alpha, \boldsymbol{\beta})$ であり、これを最大化する $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ が求める推定量である。

最後に係数の解釈について述べる。前節と同様に、x_k が条件付確率 $P(y_i = 0|\boldsymbol{x}_i)$、$P(y = 1|\boldsymbol{x})$、$P(y = 2|\boldsymbol{x})$ に与える影響は $\boldsymbol{x}$ に依存しており一定ではない。より重要なのは次のことである。x_k が $P(y = 2|\boldsymbol{x})$ に与える影響の正負は β_k の符号と一致し、逆に $P(y_i = 0|\boldsymbol{x}_i)$ に与える影響の正負は β_k の符号と逆になる。しかし、中間の選択肢が選ばれる確率 $P(y = 1|\boldsymbol{x})$ に与える影響の正負は確定しない。

2.4 多項反応ロジット・モデル

2.3 節では、経済主体の反応が 3 値以上で、かつ選択肢間に順序関係がある場合の分析手法を扱った。本節では、選択肢間に順序関係のない場合の分析手法を概観する。

選択肢間に順序関係がないと考えられる例として次のような場合が考えられる。東京-大阪間の移動手段について 3 つの選択肢（新幹線 $y_i = 0$、飛行機 $y_i = 1$、高速バス $y_i = 2$）

[6] ここで、$G(\cdot)$ は標準正規分布あるいは標準ロジスティック分布の累積分布関数である。

がある状況を想定しよう[7]。この場合、ある観察され得ない因子 (latent variable) によって選択結果が全てのサンプルについて同方向に順序付けられているとは考えられない。そこで、i 番目のサンプルが新幹線 $y_i = 0$、飛行機 $y_i = 1$、高速バス $y_i = 2$ を選択したときの効用 $U_{ij}, j = 0, 1, 2$ が次のような確率効用モデルに従うことを仮定する[8]。

$$U_{i0} = \mu_{i0} + e_{i0} \tag{2.4.1}$$

$$U_{i1} = \mu_{i1} + e_{i1} \tag{2.4.2}$$

$$U_{i2} = \mu_{i2} + e_{i2} \tag{2.4.3}$$

ここで、μ は各選択肢が選択されたときの効用が説明変数の関数として体系的に表現できる部分であり、e は確率項である。さて、実際にサンプル i が選択したのが新幹線 $y_i = 0$ だったとしよう。サンプル i が新幹線を選択したのは他の選択肢より効用が高かったためであると考えるのが妥当である。つまり、

$$U_{i0} \geq U_{i1} \Leftrightarrow (\mu_{i0} - \mu_{i1}) + (e_{i0} - e_{i1}) \geq 0 \tag{2.4.4}$$

$$U_{i0} \geq U_{i2} \Leftrightarrow (\mu_{i0} - \mu_{i2}) + (e_{i0} - e_{i2}) \geq 0 \tag{2.4.5}$$

が成り立っているはずである。(2.4.4)(2.4.5) が同時に成り立つ確率が $y_i = 0$ となる尤度である。この確率を計算するには確率変数 $e^*_{i1} \equiv (e_{i0} - e_{i1}), e^*_{i2} \equiv (e_{i0} - e_{i2})$ の分布を知らなければならない。

多項反応プロビット・モデルは (e^*_{i1}, e^*_{i2}) が 2 変量正規分布 $N(\mathbf{0}, \Omega)$ に従うことを仮定する。この場合 $P(y_i = 0)$ の計算に際し 2 重積分を計算しなければならない。この 2 重積分は解析的には計算できず数値計算をしなければならないが、多くの場合計算不可能である[9]。そこで通常は、多項反応の場合にはロジット・モデルが用いられる。

多項反応ロジット・モデルでは、e_{i0}, e_{i1}, e_{i2} が互いに独立な確率変数でその累積分布関数 $F(z)$ が次式で表わされる対数ワイブル分布（タイプ 1 極値分布）であると仮定される。

$$F(z) = \exp\{-\exp(-z)\} \tag{2.4.6}$$

この時[10]、

$$\begin{aligned}
P(y_i = 0) &= P(\mu_{i0} - \mu_{i1} + e_{i0} \geq e_{i1} \text{ and } \mu_{i0} - \mu_{i2} + e_{i0} \geq e_{i2}) \\
&= \int_{-\infty}^{\infty} f(e_{i0}) \left[\int_{-\infty}^{e_{i0}+\mu_{i0}-\mu_{i1}} f(e_{i1}) de_{i1} \cdot \int_{-\infty}^{e_{i0}+\mu_{i0}-\mu_{i2}} f(e_{i2}) de_{i2} \right] de_{i0} \\
&= \int_{-\infty}^{\infty} e^{-e_{i0}} \exp(-e^{-e_{i0}}) \cdot \exp(-e^{-e_{i0}-\mu_{i0}+\mu_{i1}}) \cdot \exp(-e^{-e_{i0}-\mu_{i0}+\mu_{i2}}) de_{i0} \\
&= \int_{-\infty}^{\infty} \exp\left[-e_{i0} - e^{-e_{i0}} \left(1 + \frac{e^{\mu_{i1}}}{e^{\mu_{i0}}} + \frac{e^{\mu_{i2}}}{e^{\mu_{i0}}} \right) \right] de_{i0} \\
&= \int_{-\infty}^{\infty} \exp\left[-e_{i0} - e^{-e_{i0}} \left(\frac{\sum_{k=0}^{2} e^{\mu_{ik}}}{e^{\mu_{i0}}} \right) \right] de_{i0}
\end{aligned}$$

[7]一般に任意の選択肢数に拡張することは容易である。

[8]応用例によっては、確率効用関数ではなくこれを確率利潤関数として定式化することもできる。

[9]より一般的に選択肢が m 個の場合 $m-1$ 重積分を計算しなければならない。これはより困難である。

[10]以下の数式展開で $f(z)$ は z の確率密度関数であり、$f(z) = dF(z)/dz = \exp(-z) \cdot \exp\{-\exp(-z)\}$ である。

ここで、$\lambda_0 = \log\left(\sum_{k=0}^{2} e^{\mu_{ik}}/e^{\mu_{i0}}\right)$ とおくと、

$$\begin{aligned} P(y_i = 0) &= \int_{-\infty}^{\infty} \exp(-e_{i0} - e^{-(e_{i0}-\lambda_0)})de_{i0} \\ &= \exp(-\lambda_0)\int_{-\infty}^{\infty} \exp\{-(e_{i0}-\lambda_0) - e^{-(e_{i0}-\lambda_0)}\}de_{i0} \\ &= \exp(-\lambda_0)\int_{-\infty}^{\infty} f(e_{i0}-\lambda_0)de_{i0} = \exp(-\lambda_0) \\ &= \frac{\exp(\mu_{i0})}{\sum_{k=0}^{2}\exp(\mu_{ik})} = \frac{e^{\mu_{i0}}}{e^{\mu_{i0}} + e^{\mu_{i1}} + e^{\mu_{i2}}} \end{aligned} \tag{2.4.7}$$

となる。同様に考えれば各選択肢が選択される確率は次式で表わされる。

$$P(y_i = j) = \frac{\exp(\mu_{ij})}{\sum_{k=0}^{2}\exp(\mu_{ik})}, \quad j = 0, 1, 2 \tag{2.4.8}$$

次に考えなければならないのは、効用のうち説明変数の関数として体系的に表現できる部分 μ がどのように決まるかである。この定式化によってロジット・モデルは分類される。

2.4.1 条件付ロジット・モデル (Conditonal Logit Model)

今、移動手段の選択に関して効用を決定する説明変数が選択肢の属性 (attributes) であるとしよう。例えば、その移動手段の運賃や所要時間である[11]。つまり、説明変数ベクトルを w_{ij} とすると、

$$\mu_{ij} = w_{ij}'\beta, \quad j = 0, 1, 2 \tag{2.4.9}$$

と書ける。ここで注意すべきことはパラメータベクトル β は全ての選択肢で共通であるということである。これと (2.4.8) 式から、

$$P(y_i = j) = \frac{\exp(w_{ij}'\beta)}{\sum_{k=0}^{2}\exp(w_{ik}'\beta)}, \quad j = 0, 1, 2 \tag{2.4.10}$$

となる。この式の自然対数をとり、サンプルについて足し合わせると対数尤度関数が得られ、それを β について最大化すると最尤推定量 $\hat{\beta}$ が得られる。

条件付ロジット・モデルの主要な目的は、推定の時点では考慮されていなかった新しい選択肢が選ばれる確率を予測することである。例えば、新しい選択肢としてリニアモーターカー $y_i = 3$ が選択できるとなったとき、これが選択される可能性は、既に推定されているパラメータ $\hat{\beta}$ と属性ベクトル w_{ij} から計算される。

2.4.2 多項ロジット・モデル (Multinomial Logit Model)

今度は説明変数がサンプルである経済主体の特徴 (characteristics) であるとしよう。例えば、移動手段の選択の説明変数が、経済主体の所得や年齢に依存すると考える場合であ

[11] これらは同じ新幹線を選択したとしても各自宅からの移動開始とすればサンプルによって異なり得る。

る。つまり、説明変数ベクトルを x_i とすると、

$$\mu_{ij} = x_i'\alpha_j, \quad j = 0, 1, 2 \tag{2.4.11}$$

と書ける。ここで注意すべきことはパラメータベクトル α_j は選択肢によって異なるということである。これと (2.4.8) 式から、

$$P(y_i = j) = \frac{\exp(x_i'\alpha_j)}{\sum_{k=0}^{2}\exp(x_i'\alpha_k)}, \quad j = 0, 1, 2 \tag{2.4.12}$$

となる。通常、多項ロジット・モデルでは、α_0 は 0 に基準化される。選択行動において問題となるのは効用の絶対水準ではなく大小関係だからである。この基準化を行った場合、(2.4.12) 式は、

$$P(y_i = j) = \frac{\exp(x_i'\alpha_j)}{1 + \sum_{k=1}^{2}\exp(x_i'\alpha_k)}, \quad j = 1, 2 \tag{2.4.13}$$

$$P(y_i = 0) = \frac{1}{1 + \sum_{k=1}^{2}\exp(x_i'\alpha_k)} \tag{2.4.14}$$

と書きかえられる[12]。この式の自然対数をとり、サンプルについて足し合わせると対数尤度関数が得られ、それを α_j について最大化すると最尤推定量 $\hat{\alpha}_j$ が得られる。

多項ロジット・モデルの主要な目的は、特徴ベクトル x_i を持つサンプルに含まれていなかった新しい経済主体が各選択肢を選択する確率を計算することである。

2.4.3 混合ロジット・モデル (Mixed Logit Model)

以上の条件付ロジット・モデルと多項ロジット・モデルは、説明変数が選択肢の属性か、サンプルの特徴かのどちらかのみである場合である。しかし、より一般的に両方を含む場合、すなわち、

$$\mu_{ij} = w_{ij}'\beta + x_i'\alpha_j, \quad j = 0, 1, 2 \tag{2.4.15}$$

の場合も考えられる。これを混合ロジット・モデルという。

2.4.4 ロジット・モデルの特徴

まず、推定されたパラメータの解釈は、2 値データの時と同じように注意が必要である。ここでは、多項ロジット・モデルの場合を考えることにすると、推定された $\hat{\alpha}_j$ は、説明変数 x_i が効用 μ_{ij} に与える限界効果を表わしている。我々が知りたいのは、選択肢 j が選択される確率に与える影響である。これは、(2.4.13) から次式で表される。

$$\frac{\partial P(y=j)}{\partial x} = P(y=j)\cdot\left[\alpha_j - \sum_{k=0}^{2} P(y=k)\cdot\alpha_k\right] = P(y=j)[\alpha_j - \bar{\alpha}] \tag{2.4.16}$$

[12] ここで、$j = 0, 1$ とすれば 2.2 節で見た二項反応の場合のロジット・モデルが得られる。これは、2.2 節の latent variable は、選択肢 $y = 0$ を選んだときの効用を 0 と基準化した上での二つの効用関数の差であったと解釈されることを意味する。2 項選択では、効用関数の差は一組なので一つの latent variable で定式化可能なのである。

条件付ロジット・モデルの場合も (2.4.10) から同様に計算できる。

次に、ロジット・モデルでは、二つの選択肢が選択される確率の比（オッズ比）がそれ以外の選択肢の存在に依存しないという性質がある。またオッズ比の自然対数が説明変数の線型関数となる。これは (2.4.8) 式から次式が導かれることから分かる。

$$\frac{P_{ij}}{P_{ik}} = \frac{\exp(\mu_{ij})}{\exp(\mu_{ik})}, \quad j \neq k,\ j,k = 0,1,2 \tag{2.4.17}$$

$$\log\left(\frac{P_{ij}}{P_{ik}}\right) = \mu_{ij} - \mu_{ik}, \quad j \neq k,\ j,k = 0,1,2 \tag{2.4.18}$$

多項反応ロジット・モデルがこのような特徴を持つのは、e_{i0}, e_{i1}, e_{i2} が互いに独立であると仮定したからである。McFadden はこのような性質を「無関係な選択肢からの独立性 (independence from irrelevant alternatives)」と呼んでいる。選択肢間に類似性が無ければ、この性質が成り立つと考えられる。

2.5 ネステッド・ロジット・モデル (Nested Logit Model)

2.4 節の多項反応ロジット・モデルは「無関係な選択肢からの独立性」が成り立たなければ妥当なモデルとはならない。しかし、この基準は選択肢に類似性のあるものが含まれている場合には成り立たない。この基準が満たされていないときに用いられるのがネステッド・ロジット・モデルである。

例えば、2.4 節の移動手段の選択の例において、3 つの選択肢が（新幹線 $y_i = 0$、A 航空会社 $y_i = 1$、B 航空会社 $y_i = 2$）というような場合である。この場合、$y_i = 1$ と $y_i = 2$ には類似性があり、効用関数の誤差項 e_{i1} と e_{i2} が独立であるとは考えがたい。つまり、この場合効用関数の誤差項 e_{i1}, e_{i2} が独立であると仮定するロジット・モデルは適していない。そこで、このような場合には次のように考える。経済主体はまず第 1 段階に新幹線を使うか航空機を使うかを選択する。もし第 1 段階で航空機が選択されたならば、次に第 2 段階として A 航空会社か B 航空会社かを選択する。

このような段階的選択は、以下のように (2.4.6) 式を変更することによって表現できる。e_{i0}, e_{i1}, e_{i2} は一般化された極値分布、

$$F(e_{i0}, e_{i1}, e_{i2}) = \exp[-G\{\exp(-e_{i0}), \exp(-e_{i1}), \exp(-e_{i2})\}] \tag{2.5.1}$$

に従うことを仮定する。さらに、この $G(\cdot)$ を次式で特定化する。

$$G(z_0, z_1, z_2) = z_0 + \left(z_1^{1/(1-\sigma)} + z_2^{1/(1-\sigma)}\right)^{1-\sigma} \tag{2.5.2}$$

これは、e_{i0} と (e_{i1}, e_{i2}) は独立であるが、e_{i1} と e_{i2} は相関があり、その相関係数がほぼ σ であることを示している[13]。もし、$\sigma = 0$ とすれば、e_{i0}, e_{i1}, e_{i2} は独立で、それぞれの分布関数は (2.4.6) になり、ロジット・モデルに帰着する。

[13] σ は厳密には相関係数ではないが、両者は極めて近い値をとることが知られている。

(2.5.1)(2.5.2) の下では、各選択肢が選択される確率は次式で表わされる。

$$P(y_i=0)=\frac{\exp(\mu_{i0})}{\exp(\mu_{i0})+[\exp\{\mu_{i1}/(1-\sigma)\}+\exp\{\mu_{i2}/(1-\sigma)\}]^{1-\sigma}} \tag{2.5.3}$$

$$P(y_i=1,2)=\frac{\exp\{\mu_{i1}/(1-\sigma)\}+\mu_{i1}/(1-\sigma)\}]^{1-\sigma}}{\exp(\mu_{i0})+[\exp\{\mu_{i1}/(1-\sigma)\}+\exp\{\mu_{i2}/(1-\sigma)\}]^{1-\sigma}} \tag{2.5.4}$$

$$P(y_i=j|y_i=1,2)=\frac{\exp\{\mu_{ij}/(1-\sigma)\}}{\exp\{\mu_{i1}/(1-\sigma)\}+\exp\{\mu_{i2}/(1-\sigma)\}},\ j=1,2 \tag{2.5.5}$$

これは (2.4.8) 式に対応する式である。μ_{ij} をどう定式化するかは、多項反応ロジット・モデルの場合分けと同様である。これをもとにして尤度関数が計算される。「無関係な選択肢からの独立性」が成り立つかどうかは、帰無仮説 $H_0:\sigma=0$、対立仮説 $H_1:\sigma\neq 0$ として t 検定あるいは尤度比検定で検定される。

2.6 ポアソン回帰モデル (Poisson Regression Model)

本節では、従属変数が非負整数値のみをとるカウントデータである場合を分析するモデルを概観する。カウントデータとは、例えば1年間に何回事故が発生したかなどの数を数えたデータである。もちろんこのようなデータに対し線型回帰モデルを当てはめることも可能である[14]。しかし、サンプルの多くが0や1といった小さな値をとり、またとり得る値が離散的であるという特徴を有効に用いれば当てはまりが改善されるかもしれない。そこで、用いられるのがポアソン回帰モデルである。

さて、従属変数が非負であるという性質を捉えるために、従属変数 y_i の期待値は次式で決定されるものと仮定する。

$$E[y_i|\boldsymbol{x}_i]=\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.6.1}$$

この式は両辺の自然対数をとると、

$$\log[E(y_i|\boldsymbol{x}_i)]=\boldsymbol{x}_i'\boldsymbol{\beta} \tag{2.6.2}$$

となり対数線型性を持っている[15]。この式は係数解釈の際に役立つ。x_k の1単位の変化は、$E(y|\boldsymbol{x})$ の $100\beta_k$ %の変化を表わすのである。また、説明変数が被説明変数に与える影響は $\partial E(y|\boldsymbol{x})/\partial x_k=\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\beta_k$ と計算される。つまり、プロビット・モデルやロジット・モデルと同様、$\partial E(y|\boldsymbol{x})/\partial x_k$ は $\boldsymbol{x}$ に依存するので一定ではない。

次に、ポアソン回帰モデルでは y_i はポアソン分布に従うことが仮定される。y_i の条件付期待値は既に (2.6.1) 式で与えられているので、$y_i=h,(h=0,1,2,\cdots)$ となる確率は、

$$P(y_i=h|\boldsymbol{x}_i)=e^{-E[y_i|\boldsymbol{x}_i]}\frac{(E[y_i|\boldsymbol{x}_i])^h}{h!}=\exp[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]\frac{[\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]^h}{h!} \tag{2.6.3}$$

で表わされる。これより、n 個の無作為標本が得られた場合の対数尤度関数は次式のようになる。

$$\mathscr{L}(\boldsymbol{\beta})=\sum_{i=1}^{n}\{-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})+y_i\boldsymbol{x}_i'\boldsymbol{\beta}-\log(y_i!)\} \tag{2.6.4}$$

[14]線型モデルを OLS で推定することはベンチマークとして有益である。

[15]カウントデータは 0 を含むので、(2.6.2) 式を直接 OLS で推計することはできない。

この対数尤度関数を最大化する $\hat{\beta}$ がポアソン最尤推定量である。

本節で見たポアソン回帰モデルはカウントデータの分析において最も自然なモデルである。しかし、(2.6.3) 式による定式化では $V(y|x) = E(y|x)$ となるが、これは制約としてかなり厳しいものである。実際にカウントデータは不均一分散性を示すことが多い。そこで、多くの拡張されたポアソン回帰モデルが提案されている[16]。

2.7 端点解反応モデル

前節までででは、従属変数が離散的である場合を扱うモデルを見た。本節以降では従属変数 y がある領域では概ね連続的な値を取るが、少なくない割合である特定の一定値となる場合を扱う。さらに、本節ではこのような制限従属変数が、経済主体の行動の結果として観察される場合を考察する。

ところでこのようなデータに対しても、通常の OLS を行うことは可能であり、線型確率モデルとプロビット・モデル、ロジット・モデルの関係と同様に、特にデータの中間域では当てはまりが良いことが多い。しかし、このような取り扱いは閾値を境にして行動が本質的に変化しているというデータが持っている重要な情報を無視することになる。また、予測値が負になる、理論的には不均一分散の可能性が高い[17]、説明変数の限界効果が一定となる、といった欠点がある。

データの持つ情報を有効に使い、その上で非負なる予測値を返し、また説明変数が非説明変数に与える限界効果が変化しうるモデルが求められる。それが端点解反応モデルである。

端点解反応モデルは従属変数がある条件を満たしたときにのみ連続的な変数として観察でき、そうでないときはある特定の値（閾値）を取るようなデータを分析する際に用いられる。例えば、

$$y_i^* = x_i'\beta + u_i, \quad u_i|x_i \sim N(0, \sigma^2) \tag{2.7.1}$$

$$y_i = \begin{cases} y_i^* & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}. \tag{2.7.2}$$

というような場合である[18]。ここで、y_i の確率密度は y_i^* が正の時 y_i^* の確率密度と同じである。さらに、

$$\begin{aligned} P(y_i = 0|x_i) &= P(y_i^* < 0|x_i) \\ &= P(u_i < -x_i'\beta|x_i) = P[u_i/\sigma < -x_i'\beta/\sigma|x_i] \\ &= \Phi(-x_i'\beta/\sigma) = 1 - \Phi(x_i'\beta/\sigma) \end{aligned} \tag{2.7.3}$$

である。ここで $\Phi(z)$ は標準正規分布の累積分布関数である。以上より、(y_i, x_i) がランダムサンプルであれば、x_i を与えた時の y_i の確率密度関数は、

$$\begin{aligned} (2\pi\sigma^2)^{-1/2}\exp[-(y_i - x_i'\beta)^2/(2\sigma^2)] &= (1/\sigma)\phi[(y_i - x_i'\beta)/\sigma], \ y_i > 0 \\ P(y_i = 0|x_i) &= 1 - \Phi(x_i'\beta/\sigma) \end{aligned}$$

[16]詳しくは、Green(1993)、937-940 ページ、Wooldridge(2003)、575-576 ページ参照。

[17]不均一分散の問題は robust standard error で解決可能である。

[18]この例ではある特定の値（閾値）が 0 である。

で与えられる。ここで ϕ は標準正規分布の密度関数である。これより、各観測値の対数尤度関数は、

$$\begin{aligned} l_i(\boldsymbol{\beta},\sigma) &= 1(y_i=0)\log[1-\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)] \\ &\quad +1(y_i>0)\log[(1/\sigma)\phi\{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})/\sigma\}] \end{aligned}$$

となる[19]。大きさ n の標本の対数尤度関数は $\Sigma_{i=1}^n l_i(\boldsymbol{\beta},\sigma)$ であり、$(\hat{\boldsymbol{\beta}},\hat{\sigma})$ が最尤法で推定される。

次に推定したパラメータの持つ意味を考えよう。推定結果と (2.7.3) 式から $P(y=0|\boldsymbol{x}), P(y>0|\boldsymbol{x})$ を推定できる。次に我々が知りたいのは二つの期待値 $E(y|\boldsymbol{x}), E(y|y>0,\boldsymbol{x})$ である。これらについて次の関係が成り立つ。

$$E(y|\boldsymbol{x}) = P(y>0|\boldsymbol{x})E(y|y>0,\boldsymbol{x}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)E(y|y>0,\boldsymbol{x}) \tag{2.7.4}$$

つまり、$E(y|y>0,\boldsymbol{x})$ が分かれば $E(y|\boldsymbol{x})$ も分かるので $E(y|y>0,\boldsymbol{x})$ を求めることが次の課題である。(2.7.1)(2.7.2) から、

$$\begin{aligned} E(y|y>0,\boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{\beta}+E(u|u>-\boldsymbol{x}'\boldsymbol{\beta}) \\ &= \boldsymbol{x}'\boldsymbol{\beta}+\sigma E[u/\sigma|u/\sigma>-\boldsymbol{x}'\boldsymbol{\beta}/\sigma] \\ &= \boldsymbol{x}'\boldsymbol{\beta}+\frac{\sigma\phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)}{\Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)} \\ &= \boldsymbol{x}'\boldsymbol{\beta}+\sigma\lambda(\boldsymbol{x}'\boldsymbol{\beta}/\sigma) \end{aligned} \tag{2.7.5}$$

となる。3 番目の等号は、z を標準正規分布に従う確率変数とすると、任意の定数 c に対して、$E(z|z>c)=\phi(c)/[1-\Phi(c)]$ が成り立つことと正規分布が対称であることから導かれる。最後の等号では、$\lambda(c)=\phi(c)/\Phi(c)$ と定義しており、この λ は逆ミルズ比あるいはハザード比と呼ばれる。この (2.7.5) は、$y_i>0$ となったサンプルについてのみ OLS を行った場合に β が一致推定量にならない理由を示している。逆ミルズ比が omitted variable となっているのである。

(2.7.4)(2.7.5) 式から $E(y|\boldsymbol{x})$ が求められる。

$$\begin{aligned} E(y|x) &= \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)[\boldsymbol{x}'\boldsymbol{\beta}+\sigma\lambda(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)] \\ &= \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma)\boldsymbol{x}'\boldsymbol{\beta}+\sigma\phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma) \end{aligned} \tag{2.7.6}$$

証明は省略するが、この式は右辺は任意の $\boldsymbol{x}$ と $\boldsymbol{\beta}$ に対して正になることが示される。

次に、説明変数が従属変数に与える限界効果は線形モデルと比較するとかなり複雑なものになる。x_k が $E(y|y>0,\boldsymbol{x})$ や $E(y|\boldsymbol{x})$ に与える限界効果の符号は β_k と同じであるがその大きさは $\boldsymbol{x}$ や他のパラメータに依存する。実際に $E(y|y>0,\boldsymbol{x})$ と $E(y|\boldsymbol{x})$ に与える限界

[19] $1(\cdot)$ はインジケーター関数である。

効果を計算すると、

$$\begin{aligned}\frac{\partial E(y|y>0,x)}{\partial x_k} &= \beta_k + \beta_k \cdot \frac{d\lambda[x'\beta/\sigma]}{dc} \\ &= \beta_k\{1 - \lambda[x'\beta/\sigma](x'\beta/\sigma + \lambda[x'\beta/\sigma])\} \qquad (2.7.7)\end{aligned}$$

$$\begin{aligned}\frac{\partial E(y|,x)}{\partial x_k} &= \frac{\partial P(y>0|x)}{\partial x_k} \cdot E(y|y>0,x) + P(y>0|x)\frac{\partial E(y|y>0,x)}{\partial x_k} \\ &= \beta_k \Phi(x'\beta/\sigma) \qquad (2.7.8)\end{aligned}$$

のようになる。ここで、証明は省略するが (2.7.7)(2.7.8) で β_k に掛け合わされている項は 0 より大きく 1 より小さいことが示される。

以上より、端点解反応モデルは通常の OLS を行った場合の欠点である、予測値が負になる、説明変数の限界効果が一定であるといった欠点を回避しているし、また閾値の存在を十分に考慮している。つまり、正の領域では概ね連続的な値を取るが少なくない割合でゼロとなるようなデータを従属変数とするときには非常に適したモデルであることがわかる。

しかし、想定される経済モデルとの適合性を考えると端点解反応モデルが適さないこともある。ここでは、端点解反応モデルが適切ではない例を一つ述べることにする。端点解反応モデルでは $y > 0$ となる確率 $P(y > 0|x)$ と、$y > 0$ という条件下での y の期待値 $E(y|y > 0, x)$ の両者に x_k が与える限界効果が共に β_k の符号と同じであり共通している。

しかし、経済学的にはこれらが逆方向に働く可能性が考えられる。例えば、生命保険の補償範囲が被説明変数、年齢が説明変数であることを考えよう。若い人は一般的に生命保険に無関心である傾向があると考えられる。そうだとすれば、年齢が上がるにつれて生命保険に加入する可能性は高くなる。しかし、一旦生命保険に加入したとするとその補償範囲は年齢が上昇すると共に小さいものになると考えられる。というのも、死が近づくにつれて生命保険の重要性が低下するからである。つまり、年齢は生命保険に入るかどうかの確率と、既に生命保険に加入しているという条件下での補償範囲には逆方向の影響を与えるのである。もし、この想定が正しいならば端点解反応モデルは本来の経済的行動を正しく描写し得ないのである。この例が示しているのは、どの LDV モデルを使うかを選択するとき、データと回帰モデルの表面的な適合性だけでなく、回帰モデルと想定する経済モデルの適合性にも注意を払わなければならないということである。

端点解反応モデルが適したモデルかどうかを評価する一つの方法はプロビット・モデルの結果と比較してみることである[20]。(2.2.7) 式と (2.7.3) 式から分かるように、プロビット・モデルによる推定値 $\hat{\beta}_k$ と端点解反応モデルの推定値 $\hat{\beta}_k/\hat{\sigma}$ が直接比較できる。もし、両者に大差がなければ端点解反応モデルはふさわしいモデルであるといえる。しかし、上に述べたように、x_k の影響が $P(y > 0|x)$ と $E(y|y > 0, x)$ に対して反対方向に働く場合、端点解反応モデルの推定値 $\hat{\beta}_k/\hat{\sigma}$ は逆方向に働く力を平均化した値となり、$P(y > 0|x)$ に与える影響だけを推計しているプロビット・モデルによる推定値と大きく乖離してくるのである。端点解反応モデルがふさわしくない場合、$P(y > 0|x)$ と $E(y|y > 0, x)$ に与える影響を別々に考察するハードル・モデル (hurdle model) が代替的モデルとなる[21]。

[20] $y_i > 0$ となっているサンプルは $y_i = 1$ と置き換えてプロビット・モデルを推定する。

[21] hurdle model については Greene(1993)、943 ページを参照。

2.8 途中打ち切り回帰モデル (Censored Regression Model)

前節で概観したモデルが用いられた理由は、経済主体の行動の結果を反映して従属変数の分布が制限を受けるという重要な特性を考慮に入れるためであり、データの観測可能性について何ら問題がないことが前提であった。

これに対し、本節では主に調査方法等から発生する censored data を扱うモデルを概観する。censored data とは、ある値（閾値）以上（あるいは以下）の値は正確には観察されず、その値以上（あるいは以下）であるということだけが分かるようなデータである。

ここでは、右側からの censoring の場合、すなわち閾値 c_i より大きいデータはその閾値より大きいということしか分からない場合を考えることにする。もし、uncensored な観測値 $y_i < c_i$ だけを用いて OLS を行えば、端点解反応モデルの場合と同様の理由で、β の最小二乗推定量は一致推定量とならない。そこで、次のようなモデルを考える[22]。

$$y_i = x_i'\beta + u_i, \quad u_i|x_i, c_i \sim N(0, \sigma^2) \tag{2.8.1}$$

$$w_i = \min(y_i, c_i) \tag{2.8.2}$$

ここで、c_i に添字 i がついているのは閾値がサンプルごとに異なる場合をも考慮するためであり、これは実際のデータでもしばしば起こり得ることである。

(2.8.1)(2.8.2) を前提として β の最尤推定量を求めることができる。censoring されていないサンプル $y_i = w_i$ の確率密度は $(1/\sigma)\phi[(w_i - x_i'\beta)/\sigma]$ である。一方、censoring されている確率は、

$$P(w_i = c_i|x_i) = P(y_i \geq c_i|x_i) = P(u_i \geq c_i - x_i'\beta) = 1 - \Phi[(c_i - x_i'\beta)/\sigma] \tag{2.8.3}$$

である。故に無作為標本 (x_i, w_i) の対数尤度関数は、

$$\begin{aligned} l_i(\beta, \sigma) = \; & 1(w_i = c_i) \log[1 - \Phi(c_i - x_i'\beta/\sigma)] \\ & + 1(w_i < c_i) \log[(1/\sigma)\phi\{(w_i - x_i'\beta)/\sigma\}] \end{aligned}$$

である。$\Sigma_{i=1}^n l_i(\beta, \sigma)$ を最大化する $(\hat{\beta}, \hat{\sigma})$ が最尤推定量である。

ところで、この対数尤度関数は 2.7 節の端点解反応モデルと本質的に同一である[23]。しかし、推定された係数の解釈は 2.7 節と異なる。2.7 節で我々が知りたかったのは、x_k が latent variable y^* ではなく y の期待値に与える影響であった。一方、途中打ち切り回帰モデルの場合は、w ではなく y に対して x_k が持つ影響が関心事になることが通常である。故に、途中打ち切り回帰モデルの場合、β_k の解釈は線型回帰モデルと同様に考えられるのである。

2.7 節の端点解反応モデルと本節の途中打ち切り回帰モデルは従属変数の制限のされ方は本質的に同じであり、そのためモデルの形式的な構造も同じである。しかし、そのような従属変数が観察された理由が決定的に異なっている。経済主体の行動結果として従属変

[22] ここでは u_i が正規分布であることを仮定しているので censored normal regression model と呼ばれる。

[23] Amemiya(1984) では、2.7 節、2.8 節、2.9 節のどのモデルもタイプ I のトービット・モデルと分類している。これは尤度関数の形状が共通しているからである。本稿では、従属変数の制限のされ方が背後にある経済学的理論によるものなのか、あるいはデータ観測の不完全さによるものかを区別することに重要性を与えているので 2.7 節、2.8 節、2.9 節のモデルを区別した。

数が制限されている場合、観察されたデータは経済主体の行動様式について有意義な情報を含んでおり、それを積極的に利用するのが端点解反応モデルである。一方、データの観測上の問題から従属変数が制限されている場合、制限されていること自体は本来望ましくない。実際、本節では (2.8.1) 式で $u|x,c$ に正規性と均一分散を仮定したが、これが妥当でない場合最尤推定量は一致性を持たなくなる。もし、データが censoring されていなければ、(2.8.1) を OLS で推計すれば、$u|x$ に正規性と均一分散を仮定しなくても推定量は一致性を持つのである。データが観測上の問題点から censoring されている場合、本来なら censoring されていないデータが望まれるが、それが不可能な場合、次善の策として少なくとも censoring がしてあるという情報を有効に利用するために途中打ち切り回帰モデルが用いられるのである。

ただし、censoring されたデータしか知り得ない場合もある。この場合は途中打ち切り回帰モデルは最善のモデルである。その例として期間分析 (duration analysis) が挙げられる。期間分析では、ある事象が再発するまでの期間が被説明変数となる。この場合、データの一部分にまだ再発していないものが含まれてくる。これは再発しないことを意味しているのではなくその時点ではまだ再発していないだけである。つまり、「前回の発生からその時点までに経過した時間」が「再発するまでの期間」の censoring されたデータとなる。

2.9 切断回帰モデル (Truncated Regression Model)

本節では、truncated data を扱うモデルを概観する。truncated data とはある値（閾値）以上（あるいは以下）の値は全く観察されないようなデータである。言い換えれば、母集団の一部分に関しては、y_i だけでなく x_i も観察されないようなデータである。x_i も観測されないので、truncated data は censored data よりも情報が少ないデータである。このような truncated data は、例えば女性の労働時間を決定する式を推計するため、従属変数として労働時間を、独立変数として個人の属性（経験年数、学歴等）をデータとして取る際に、働いている人だけを調査対象とした場合などに得られる[24]。

まず、母集団が次のような通常の線型回帰モデルに従っていると考える[25]。

$$y_i = x_i'\beta + u_i, \quad u_i|x_i \sim N(0,\sigma^2) \tag{2.9.1}$$

もし、得られた標本 (y_i, x_i) がランダムサンプルであるならば、OLS が最適な推定方法である。しかし、ここでは右側からの truncation を考えることにすると、標本が観察されるのは $y_i \leq c_i$ の時のみである[26]。故に、$y_i \leq c_i$ を満たすあるサンプルが観測される確率密度 $f(y_i|x_i,c_i)$ は、

$$f(y_i|x_i,c_i) = \frac{\sigma^{-1}\phi[(y_i - x_i'\beta)/\sigma]}{\Phi[(c_i - x_i'\beta)/\sigma]} \tag{2.9.2}$$

[24] もし仮に労働していない女性についても属性（経験年数、学歴等）を観察できるのであれば、2.8 節の途中打ち切り回帰モデル (Censored Regression Model) が適切な回帰モデルである。

[25] ここでは u_i が正規分布であることを仮定しているので truncated normal regression model と呼ばれる。

[26] ここで c_i は truncation の閾値であり、これはサンプルごとに異なり得るので添字 i が付いている。通常 c_i は独立変数 x_i やモデル外のサンプルの属性に依存する。

である。(2.9.2) 式の自然対数をとり、それを各 i について合計したものを最大化することによって最尤推定量 $(\hat{\beta}, \hat{\sigma})$ が得られる。途中打ち切り回帰モデル同様、(2.9.1) 式で $u|x$ に正規性と均一分散を仮定したが、この仮定が無ければ最尤推定量 $(\hat{\beta}, \hat{\sigma})$ に一致性が無くなる。

2.10 標本選別モデル (Sample Selection Model)

得られた標本が必ずしも無作為標本でない場合がある。無作為標本ではない場合、標本選別があるといわれるが、標本選別を考えるにあたって、外生的標本選別と内生的標本選別を区別することがまず重要である。

2.10.1 内生的標本選別と外生的標本選別

内生的標本選別と外生的標本選別の区別を考えるために、まず母集団からの無作為標本は次のような線型モデルに従っていることを仮定しよう。

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad E(u_i|\boldsymbol{x}_i) = 0 \text{ for all } i = 1, 2, \cdots, n \tag{2.10.1}$$

もし、全ての標本について $(y_i, \boldsymbol{x}_i)$ が観察可能ならば、単に OLS を使えば不偏性、一致性を持つ推定量が得られる。ここでは何らかの理由で、いくつかの標本については y_i か $\boldsymbol{x}_i$ が観察できない場合を考える。そのためにまず、標本選別を表すインディケーター関数 $s_i = 1$[if we observe all of $(y_i, \boldsymbol{x}_i)$] $s_i = 0$[if we can not observe some of $(y_i, \boldsymbol{x}_i)$] を定義する。問題は、$s_i = 1$ になるサンプルだけを用いて OLS を行ったときに推定量がどのような性質を持つかということである。標本について完全な情報が得られたものだけを用いて OLS を行うということは次式を推定することに等しい。

$$s_i y_i = s_i \boldsymbol{x}_i'\boldsymbol{\beta} + s_i u_i \tag{2.10.2}$$

この式を OLS で推定した結果が一致性を持つためには、

$$E(s_i u_i) = 0 \tag{2.10.3}$$

$$E[(s_i x_{ik})(s_i u_i)] = E(s_i^2 x_{ik} u_i) = E(s_i x_{ik} u_i) = 0 \tag{2.10.4}$$

が成り立たなければならない。すなわち、説明変数 $s_i x_{ik}$ と誤差項 u_i に相関がなく、誤差項の期待値が 0 であれば OLS 推定量は一致性を持つ。また、不偏性が成立するためにはより強い条件、

$$E(s_i u_i | s_i \boldsymbol{x}_i) = 0 \tag{2.10.5}$$

が必要である。(2.10.5) は (2.10.3)(2.10.4) が成り立つための十分条件である。

まず、外生的標本選別について考える。標本選別が独立変数 $\boldsymbol{x}_i$ によって行われることを外生的標本選別という[27]。言い換えれば、s_i が独立変数 $\boldsymbol{x}_i$ のみの関数であるならば外生的標本選別である。s_i が独立変数 $\boldsymbol{x}_i$ のみの関数であるなら $s_i x_{ik}$ も独立変数 $\boldsymbol{x}_i$ のみの関数

[27]独立変数はモデルの外生変数だからである。

となる。故にこのとき、母集団からの無作為標本が $E(u_i|x_i) = 0$ を満たすという仮定と、条件付期待値の線型性 $E[f(x)y|x] = f(x)E(y|x)$ から、

$$E(s_i u_i|s_i x_i) = s_i E(u_i|s_i x_i) = s_i E(u_i|x_i) = 0 \tag{2.10.6}$$

を得る。つまり (2.10.5) が成立する。従って、外生的標本選別の場合、(2.10.2) を OLS で推定すれば不偏一致推定量が得られることになり問題は無い。

次に、内生的標本選別を考える。標本選別がモデルの内生変数によって行われることを内生的標本選別という。内生的標本選別がある場合、観察されたサンプルについてのみ OLS を行うとバイアスがもたらされる。内生的標本選別については次節で述べる。

2.10.2 標本選別モデル

2.9 節の truncated data は標本選別がモデルの従属変数（内生変数）y_i によって行われるので内生的標本選別の一種であるが、ここではより一般化された場合を扱う。2.9 節で扱った truncated data は、従属変数がある閾値を超える（あるいは下回る）場合には、従属変数と独立変数が両者とも観察されないというものであった。本節ではより一般的に、従属変数が観察できるかどうかが、経済主体の行動によって決定されるものと想定し、その行動がモデルとして明示化される。つまり、標本選別をもたらす経済主体の行動はモデルの内生変数となっている。

例えば、既婚女性への賃金オファー関数を推計することを考える。もし、無作為抽出されたサンプルが労働しているならば、彼女の受け取っている賃金が賃金オファーであると考えられる。彼女の受け取っている賃金を従属変数、各個人の特徴（年齢、経験年数、学歴等) を独立変数とした回帰モデルが想定される。一方、労働していないサンプルについては賃金を観察することができない。かといってこのサンプルに対して賃金オファーが無いわけではない。そこで、働いていないのは、この個人による合理的な選択の結果であると考え、その行動をも含めて回帰モデルを想定するのである。

まず、母集団からの無作為標本は (2.10.1) に従っているものと仮定する。その上で、次のようなモデルを想定する。

$$s_i = 1[y^* = z_i'\gamma + u_{1i} \geq 0] \tag{2.10.7}$$

$$s_i y_i = s_i x_i'\beta + s_i u_{2i} \tag{2.10.8}$$

ここで、独立変数 z は無作為抽出された全てのサンプルについて完全に観察できるものと仮定する。独立変数 x については $s = 1$ になるサンプルのみについて観察されれば良い。また、u_1, u_2 は平均 0、分散 1、σ_2^2、共分散 σ_{12} の 2 変量正規分布に従うものと仮定する[28]。賃金オファー関数の例で言えば、latent variable y^* は、賃金オファーと留保賃金の差であると考えられる。留保賃金より賃金オファーが高ければ、労働市場に参入するのである。その結果 $s = 1$ となり、実際に受け取っている賃金が賃金オファー y として観察される。

[28]すなわち、u_1 と u_2 の相関係数は $\rho = \sigma_{12}/\sigma_2$ である。60 年代には、u_{1i}, u_{2i} が独立であると仮定している研究も多い。しかしこれは推定上の利点からであり、多くの場合非現実的な仮定である。

以下ではこのモデルをヘックマンの2段階推定法に沿って見ていく[29]。$s_i = 1$ を条件として (2.10.8) 式の期待値をとれば、

$$\begin{aligned} E(y_i|s_i = 1) &= x_i'\beta + E(u_{2i}|s = 1) \\ &= x_i'\beta + \sigma_{12} \cdot \lambda(z_i'\gamma) \end{aligned} \tag{2.10.9}$$

を得る。ここで、$\lambda(z_i'\gamma) = \phi(z_i'\gamma)/\Phi(z_i'\gamma)$ であり逆ミルズ比と呼ばれる。この式から、

$$y_i = x_i'\beta + \sigma_{12} \cdot \lambda(z_i'\gamma) + \epsilon_i,\ E(\epsilon_i|s_i = 1) = 0 \tag{2.10.10}$$

を得る。この式は、$s_i = 1$ になったサンプルについて y_i を x_i のみに回帰すると、逆ミルズ比 $\lambda(z_i'\gamma)$ が omitted variable となり推定値が一致性推定量で無くなることを示している。逆に、$s_i = 1$ であるサンプルについて y_i を x_i と $\lambda(z_i'\gamma)$ に回帰すれば一致推定量が得られる。しかし、$\lambda(z_i'\gamma)$ は未知パラメータ γ を含むので $\lambda(z_i'\gamma)$ の真の値を知ることはできない。そこでまず γ の一致推定量を求める。モデルの仮定から、

$$P(s_i = 1|z_i) = \Phi(z_i'\gamma) \tag{2.10.11}$$

となり、s はプロビット・モデルに従うことが分かる[30]。従って、プロビット推定量 $\hat{\gamma}$ を求めることができ、これを元に $\lambda(z_i'\gamma)$ を推定できる。

ヘックマンの2段階推定法をまとめると次のようになる。

1. 無作為標本全てを用いて (2.10.11) 式を推定し、プロビット推定量 $\hat{\gamma}$ を求める。これを用いて、逆ミルズ比 $\lambda(z_i'\hat{\gamma})$ を計算する。
2. 第1段階で求め逆ミルズ比を使って、(2.10.10) を $s_i = 1$ であるサンプルについてOLSで推定し、$\hat{\beta}$ を求める。$\hat{\beta}$ は一致推定量で漸近的正規性を持つことが知られている。

3 おわりに

ミクロデータの計量分析を行う場合、重要なのは次の関係に注意することである。第1に、分析対象であるデータと回帰モデルの適合性である。第2に、用いる回帰モデルと分析者が想定する経済モデルの適合性である。これらを考慮して、用いる回帰モデルを決定する必要がある。また、係数の解釈が線型回帰モデルと異なる場合が多く、このことにも注意する必要がある。

本稿ではミクロデータの分析というタイトルを掲げながらも、ミクロデータ分析の中で最も重要なものの一つであるパネルデータ分析には一切触れなかったことを断っておきたい。家計データなどでは、サンプルとなった家計を追跡調査しパネル化されているものも多く、単年度のクロスセクションデータでは得られない重要な情報を含んでいる。パネルデータ分析については、参考文献 [11] 等を参照されたい。

[29] このモデルを前節までと同様に最尤法で推定する方法もある。しかし、最尤法の場合、尤度関数が大域的凸性を満たさないことや他の理由から、ヘックマンの2段階推定法が用いられることが多い。両者の相違や使い分けについては、牧・宮内・浪花・縄田・「応用計量経済学II」第4章を参照。

[30] (2.2.1) 式参照。

表1

LDV Models	従属変数の特徴	推定法
線型確率モデル	2値データ (0,1) 例）y=0 労働しない y=1 労働する	OLS
プロビット・モデル ロジット・モデル	2値データ (0,1) 例）y=0 労働しない y=1 労働する	最尤法
順序プロビット・モデル 順序ロジット・モデル	多値データ (0,1,2,…,n) 順序関係有り 例）試食の結果として、 y=0 まずい y=1 ふつう y=2 おいしい	最尤法
多項反応ロジット・モデル ・条件付ロジット・モデル （説明変数が選択肢の属性） ・多項ロジット・モデル （説明変数がサンプルの特徴） ・混合ロジット・モデル	多値データ (0,1,2,…,n) 順序関係無し、「無関係な選択肢からの独立性」あり 例）移動手段の選択として、 y=0 新幹線 y=1 飛行機 y=2 高速バス	最尤法
ネステッド・ロジット・モデル	多値データ (0,1,2,…,n) 「無関係な選択肢からの独立性」なし 例）移動手段の選択として、 y=0 新幹線ひかり号 y=1 新幹線のぞみ号 y=2 飛行機	最尤法
ポアソン回帰モデル	カウントデータ (0,1,2,…,n) 例）年間交通事故発生回数	最尤法
端点解反応モデル (Type1 Tobit)	従属変数の値に関して、無視し得ない割合に0が現れ、それ以外は連続的な値をとるデータ 例）耐久消費財に対する支出額	最尤法
途中打ち切り回帰モデル (Type1 Tobit)	従属変数がある閾値を境に正確な値は観察不可能であるが、同じサンプルに対応する独立変数は観察されるデータ （censored data） 例）ある事象が再発生するまでの期間	最尤法
切断回帰モデル (Type1 Tobit)	従属変数がある閾値を境に観察不可能であり、同じサンプルに対応する独立変数も観察不可能なデータ （truncated data） 例）就労者のみを対象とした労働時間	最尤法
標本選別モデル (Type2 Tobit)	従属変数が一部のサンプルについて観察不可能なデータ 例）賃金オファー	ヘックマンの2段階推定法

参考文献

[1] 小林正人、「順序プロビット・モデルのテストと社債格付データへの応用」、『金融研究』、第 20 号 (別冊 1)、2001

[2] 牧厚志、宮内環、浪花貞夫、縄田和満、「応用計量経済学 II」、多賀出版、1997

[3] 松田芳郎、伴金美、美添泰人編、「講座ミクロ統計分析第 2 巻－ミクロ統計の集計解析と技法」、日本評論社、2000

[4] 縄田和満、「トービット・モデルによる金融資産分析への応用について」、『フィナンシャル・レビュー』、第 23 号、1992

[5] 和合肇、伴金美、「TSP による経済データの分析」、東京大学出版会、1995

[6] Amemiya,Takeshi., ' Qualitative Response Models:A Survey',*Journal of Economic Literature*,No.19,1981,481-536

[7] Amemiya,Takeshi., ' Tobit Models:A Survey',*Journal of Econometrics*,No.24,1984,3-61

[8] Hayashi,Fumio., ' *Econometrics* ' ,Princeton University Press,2000

[9] Maddala,G.S.,' *Limited-Dependent and Qualitative Variables in Econometrics* ',Cambridge University Press,1983

[10] Wooldridge,Jeffrey M., ' *Introductory Econometrics* ' 2nd edition,Thomson Learning,2003

[11] Wooldridge,Jeffrey M.,' *Econometric Analysis of Cross Section and Panel Data*',MIT Press,2002

[12] Mroz,T.A, ' The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions ' ,*Econometrica*,No.55,1987,765-799

[13] Greene,William.H, ' *Econometric Analysis* ' ,Prentice-Hall,1993