第５講　切断従属変数

従属変数が質と量の両方の性質を持つようなデータを考えてみよう。例えば、

1. 自動車関係経費への支出
2. アウトドアレジャーへの支出
3. 株式の保有高

これらの変数を個人レベルで観察すると、一部の個人がプラスで残りはゼロの値をとる。すなわち、このタイプのデータは (1) 保有しているかどうか（０か１）の選択（質的変数）と (2) 保有している場合にはどれぐらいか（量的変数）が複合されたものであることがわかる。

ロジットやプロビットモデルでは潜在変数の値は直接観察できないが、この場合、潜在変数（支出）が臨海値（０）を越えるとその値が観察されるようになるという性質を持っている。

このようなデータは厳密には次の２種類に分けることができる。すなわち、従属変数がある値以上、以下、あるいは上下で切断されていても、説明変数はすべての標本について観察される場合を**検閲標本（censored sample）**と呼び、説明変数についても切断されなかった標本のみが観察される場合を**切断標本（truncated sample）**と呼ぶ。 以下ではこの両者の区別は行わない。

5.1　切断正規分布（truncated normal distribution）

切断された変数は、例えば、平均が μ の正規分布から選ばれた無作為標本（y）について、ある水準以下のデータは検閲（censor）されて消されているようなケースに相当する。

$$\begin{array}{ll} y > a & \text{なら}\quad y \text{の値がそのまま公開される} \\ y \leq a & \text{なら}\quad y = 0 \text{として扱われる} \end{array}$$

y の密度関数を $f(\cdot)$、累積密度関数を $F(\cdot)$ で表すと、検閲をパスした標本の密度関数は

$$\frac{f(y)}{1 - F(a)}, \quad y > a \tag{5.1}$$

となる。

図 5.1 のように検閲をパスした標本は密度関数 $f(\cdot)$ を a で切断した右側の密度と同じ形になる（赤線の部分）。条件付きの密度関数はその下の面積が１になるように調整し、$f(\cdot)$ を y が a を超える確率 $(1 - F(a))$ で割ることにより得られる。

図5．1

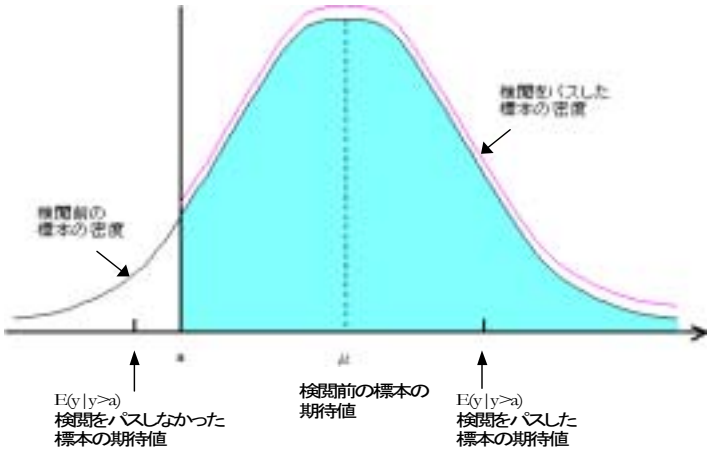
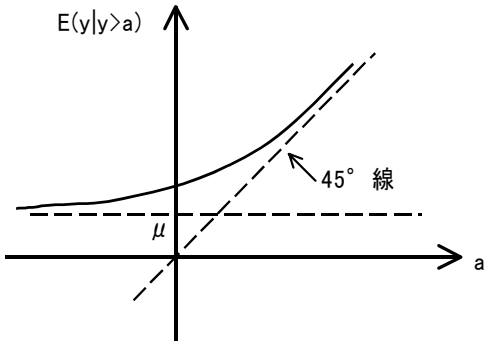


図５．２　切断された正規分布の期待値



切断された分布の条件付き期待値は (5.1) の密度関数より

$$E(y \mid y > a) = \int_a^\infty \frac{yf(y)}{1 - F(a)} dy \tag{5.2}$$

により求められる。これは無条件の期待値 μ より大きくなる。

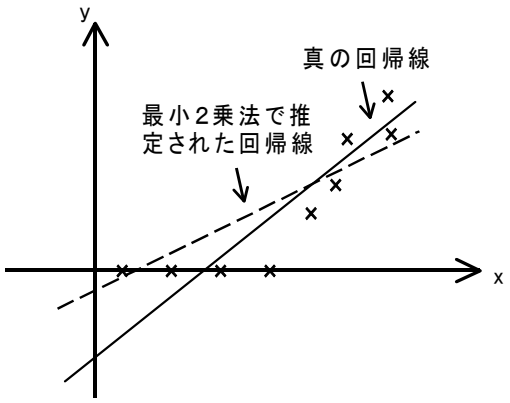
一般的に、切断された分布の期待値と元の分布の期待値については、次のような関係が成り立つ。

$E(y \mid y > a) > E(y)$　：　下から切断された分布の期待値
$E(y \mid y > a) < E(y)$　：　上から切断された分布の期待値

さらに y が正規分布に従う場合、条件付き期待値 $E(y \mid y > a)$ と切断点 a の関係は 図 5.2 のようになる。

条件付き期待値は a が小さくなると無条件の期待値 μ に漸近し、大きくなるに従って a に近づく。

図５．３　トービット･モデルの考え方



5.2　トービット・モデル

観察される値 (y) と潜在変数 (y^*) の間に次の関係が存在し、誤差項 ε_i は x と無相関で独立に平均 0 の正規分布に従うとする。

$$\begin{aligned} y_i^* &= \alpha + \beta x_i + \varepsilon_i \quad i = 1, \cdots, n \\ \varepsilon &\approx N(0, \sigma^2) \end{aligned} \tag{5.3}$$

$$\begin{aligned} y_i^* > 0 \text{ なら } y_i &= y_i^* \\ y_i^* \leq 0 \text{ なら } y_i &= 0 \end{aligned} \tag{5.4}$$

ここでは、潜在変数については古典的正規回帰モデルの仮定が成立しているが、潜在変数が負ならば切断され、その値は直接観察できないと考える。このようなモデルを提唱者トービンにちなんで**トービット・モデル**と呼ばれている。(図 5.3)

ここでは真の回帰線のパラメータ (α, β) と分散 σ^2 の値を求めることが問題となる。支出がゼロとなる確率は x の値に依存して変化し、ε についての条件として書き表せる。

$$\begin{aligned} &\Pr(y = 0 \mid x) = \Pr(y^* \leq 0 \mid x) = \Pr(\alpha + \beta x + \varepsilon \leq 0 \mid x) \\ &\quad = \Pr(\varepsilon \leq -\alpha - \beta x \mid x) = F(-\alpha - \beta x) \end{aligned}$$

ここで $F(\cdot)$ は ε の累積密度関数である。確率は標準正規分布の累積密度関数 $\Phi(\cdot)$ を使って書き直すと、

$$\Pr(y = 0 \mid x) = \Phi\left(\frac{-\alpha - \beta x}{\sigma}\right) \tag{5.5}$$

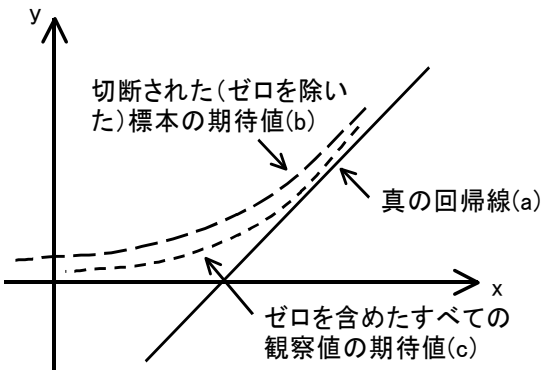
プラスの支出の確率は、

$$\Pr(y > 0 \mid x) = 1 - \Phi\left(\frac{-\alpha - \beta x}{\sigma}\right) \tag{5.6}$$

となる。支出がプラスで切断された標本の条件付き期待値は次のように表せる。

$$\begin{aligned} E(y \mid y^* > 0, x) &= E(\alpha + \beta x + \varepsilon \mid \varepsilon > -\alpha - \beta x) \\ &= \alpha + \beta x + E(\varepsilon \mid \varepsilon > -\alpha - \beta x) \end{aligned} \tag{5.7}$$

図5．４　切断された標本の期待値



ε の条件付き期待値 $E(\varepsilon \mid \varepsilon > -\alpha - \beta x)$ は分布が下から切断されているため、プラスとなる。切断された標本の期待値は常に潜在変数の期待値より大きくなる。

図 5.4 は真の回帰線 (a) と切断された標本の条件付き期待値 (b) およびゼロを含むすべての観察値の期待値 (c) の関係を示したものである。(c) は、

$$\begin{aligned} E(y \mid x) &= E(y \mid y = 0, x) = \Pr(y = 0 \mid x) + E(y \mid y > 0, x) \Pr(y > 0 \mid x) \\ &= E(y \mid y > 0, x) \Pr(y > 0 \mid x) \\ &= (\alpha + \beta x + E(\varepsilon \mid \varepsilon > -\alpha - \beta x))(1 - \Phi((-\alpha - \beta x)/\sigma)) \end{aligned} \tag{5.8}$$

となり、(5.7) と比べて切断された標本の期待値より小さくなる。つまり、$(b) > (c)$ となる。

これらの関係を推計する場合には最小２乗法を用いるとゼロに偏った推計となることが、図 5.3 と図 5.4 から明らかである。

5.3　最尤法

従属変数がプラスの場合、尤度関数は密度の積、ゼロの場合、尤度関数は確率の積として表せる。プラスの場合 $y = \alpha + \beta x + \varepsilon$ なので、$\varepsilon = y - \alpha - \beta x$ と書ける。$\varepsilon \approx N(0, \sigma^2)$ であり ε/σ は標準正規分布に従う。標準正規分布の密度関数 $\phi(\cdot)$ により y の密度は次のように表せる。

$$f(y) = \frac{1}{\sigma}\phi\left(\frac{y - \alpha - \beta x}{\sigma}\right) \tag{5.9}$$

(5.5) のようにゼロが起こる確率は $\Pr(y = 0 \mid x) = \Phi\left(\frac{-\alpha-\beta x}{\sigma}\right)$ である。尤度関数は、

$$L(\alpha, \beta, \sigma) = \prod_{y_i > 0} \frac{1}{\sigma}\phi\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right) \prod_{y_i = 0} \Phi\left(\frac{-\alpha - \beta x_i}{\sigma}\right) \tag{5.10}$$

となる。対数尤度を α, β, σ について最大化することによって、最尤推定量が求まる（**トービット推定**という）[1]。

[1] トービットの最尤推定量の詳しい導出については Maddala(1983) chapter6 を参照されたい。

5.4 トービット・モデルの限界

1. **誤差項に関する想定・非正規性と不均一分散性**

   古典的回帰モデルにおいて最小２乗推定量は誤差項が正規分布に従わなくても**最小分散不偏**であり、また誤差項が不均一分散でも**不遍性**と**一致性**は保たれていた。 つまり最小２乗法は誤差項の想定について頑強（robust）な推定量であった。

   トービット・モデルでは、正規性が成立しない場合や不均一分散の場合には、最尤推定値は**一致性**を持たない。しかし古典的回帰モデルのように誤差項の想定に対して頑健な推定法はない。これは従属変数の分布をあらゆる情報を使って適切に推測して、それに基づいて密度関数の分布を設定するしかないということを意味している。

   さらに Heckman(1990) は以下で指摘するような問題を解決する目的で次のような一般モデルを提示した。[2]

$$y_{1i} = \mathbf{X}_{1i}\beta_1 + \varepsilon_{1i} \tag{5.11}$$
$$y_{2i} = \mathbf{X}_{2i}\beta_2 + \varepsilon_{2i} \tag{5.12}$$
$$T_i = 1(\mathbf{Z}_i\gamma + \varepsilon_{0i} > 0) \tag{5.13}$$
$$y_i = T_i y_{1i} + (1 - T_i)y_{2i} \tag{5.14}$$

   ここで T_i は $\mathbf{Z}_i\gamma + \varepsilon_{0i} > 0$ ならば 1、$\mathbf{Z}_i\gamma + \varepsilon_{0i} \leq 0$ ならば 0 をとる選択変数である。
   このモデルの要点は以下の２つである。

2. **変数効果の多様性（heterogeneity）**

   変数にかかる係数は個人の属性に応じて決まる。例えば、労働組合が賃金に与える効果について考えてみよう。一般的に、組合員の賃金は非組合員の賃金より高い。(5.11) は組合員の賃金関数とし、(5.12) は非組合員の賃金関数だとしよう。非組合員は学歴による賃金上昇率は低いことが知られている。X_i の k 番目の変数は学歴だとすると、$\beta_2^k < \beta_1^k$ ということである。組合員であるかどうかの効果は切辺だけではなく、係数の違いにも出てくるのである。すなわち、

$$\text{組合員効果 } i = \mathbf{X}_i(\beta_1 - \beta_2) \tag{5.15}$$

   である。

3. **選別性バイアス（selectivity bias）**

   選別性の問題は、選別された（管理）グループの属性が選別の方法自体に影響を受けており、かつ結果に影響を与えるとすれば、選別された変数が結果に影響を与えているという因果関係は、誤っていることがある。例えば、私立学校の教育の質の評価について考えてみよう。教育の効果は卒業後の所得で測れるとする。経済学者は学生の家庭環境に関する情報はないものとする。もしこの私立学校が家庭の資産が大きい学生を優先的に選別しており、また裕福な家庭の学生は教育の質に関わりなく裕福になりやすいという事実があるとすれば、この選別のメカニズムを知らずに計量分析すると、本当は家庭環境の違いに帰属する効果を教育の効果と間違えてしまうことになる。

[2] 本節は Johnston and Dinardo (1997), pp.446-452 を参照。

Heckman(1976) は上述の選別性バイアス問題を二段階の修正を施すことによって解決できることを示した（Heckman's two-step estimator)。

選別性バイアスに関する古典的事例は Gronau(1974) によって提示されている。この場合、女性賃金が結果であり、労働市場への参入を選別効果としよう。

次のような女性賃金関数を考える（これは (5.11) に相当する)。

$$w_i = \mathbf{X}_i\beta + \varepsilon_{1i} \tag{5.16}$$

ここで w_i は対数表示の賃金、$\mathbf{X}_i$ は労働経験や学歴などの属性ベクトルである。ここで問題になるのは労働市場に参入してくる女性はランダムに選ばれたわけではない。この選別によって係数 β はバイアスを生じるだろう。

労働市場への参加関数を次のように書く（これは (5.13) に相当)。

$$T_i = 1 \quad (\mathbf{Z}_i\gamma + \varepsilon_{0i} > 0) \tag{5.17}$$

ここで $\mathbf{Z}$ は女性の労働市場への参加を説明する変数のベクトルである。$\mathbf{Z}_i$ と $\mathbf{X}_i$ には共通の変数も含まれていいが、Gronau は $\mathbf{Z}$ に子供の数も含めた。これは女性の労働市場への参加に影響を与えるかもしれないが、賃金には影響しない。

労働市場に参加している女性に関する賃金関数 (5.17) の期待値をみれば、選別問題は明らかである。

$$E\left[w_i \mid \mathbf{X}_i \quad T_i = 1\right] = \mathbf{X}_i\beta + E\left[\varepsilon_{\gamma i} \mid \varepsilon_{0i} > -\mathbf{Z}_i\gamma\right] \tag{5.18}$$

ε_0 と ε_1 はともに正規分布し、次のような関係がある。

$$\varepsilon_{1i} = \frac{\sigma_{0,1}}{\sigma_0^2}\varepsilon_{0i} + \upsilon_i = \frac{\sigma_{0,1}}{\sigma_0} \cdot \frac{\varepsilon_{0i}}{\sigma_0} + \upsilon_i \tag{5.19}$$

ここで υ_i は ε_{0i} に無相関、$\sigma_{0,1}$ は ε_{0i} と ε_{1i} の共共分散、ε_{1i} は σ_0^2 は ε_{0i} の分散である。

(5.20) を使って (5.19) の誤差項を書き換える。

$$\begin{aligned} E\left[\varepsilon_{1i} \mid \varepsilon_{0i} > -\mathbf{Z}_i\gamma\right] &= \frac{\sigma_{0,1}}{\sigma_0} E\left[\frac{\varepsilon_{0i}}{\sigma_0} \mid \frac{\varepsilon_{0i}}{\sigma_0} > \frac{-\mathbf{Z}_i\gamma}{\sigma_0}\right] \\ &= \frac{\sigma_{0,1}}{\sigma_0} \frac{\phi(\mathbf{Z}_i\gamma/\sigma_0)}{\Phi(\mathbf{Z}_i\gamma/\sigma_0)} \end{aligned} \tag{5.20}$$

ここで $\phi(\cdot)$ は標準正規密度関数、$\Phi(\cdot)$ はその累積密度関数である。選別バイアスは $\sigma_{0,1}$ がゼロでない場合に生じると解釈できる。

ところで (5.17) を OLS で推計すると、β はバイアスを持つが、次のような変数（ここで $\frac{\phi(\mathbf{Z}_i\gamma/\sigma_0)}{\Phi(\mathbf{Z}_i\gamma/\sigma_0)}$ は inverse Mills ratio、あるいはハザード比と呼ばれる）を加えることでこのバイアスを除去することができる。

$$w_i = \mathbf{X}_i\beta + \frac{\phi(\mathbf{Z}_i\gamma/\sigma_0)}{\Phi(\mathbf{Z}_i\gamma/\sigma_0)}\tilde{\sigma} \tag{5.21}$$

Heckman はこのようなモデルを次のような 2 段階修正で推計することを提示した。

1. $\mathbf{Z}_i$ をプロビット推計することで γ/σ_0 を求める。
2. the inverse Mills ratio を計算する。

3. $T_i = 1$ のデータに関して (5.21) を OLS で推計する。この推定量は一致推定量である。

実際には、このような手続きは既存の計量パッケージ（例えば STATA）のコマンド（Heckman）を用いることで Heckman 二段階推計は容易に行うことができる。

もちろん、この推計方法は必ずしも最適とはいえないし、(5.17) を単純に推計したほうがより効率的である場合もある。Davidson and Mackinnon (1993) が指摘しているように、２段階推計は選別性バイアスがあるかどうかのテストに用いるべきであるという考え方もある。

References

[1] Amemiya, T. (1984) "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3-63.

[2] Barnow, B., Cain, G. and Goldberger, A. (1976) "Issues in the Analysis of Selectivity Bias," *Evaluation Studies Review Annual*, 5, 43-59.

[3] Davidson, R. and Mackinnon, J. (1993) *Estimation and Inference in Econometrics*, Oxford: Oxford University Press.

[4] Gronau, R. (1974) "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, 82, 1119-1155.

[5] Heckman, J. (1976) "The Common Structure of Sttistical Models of Truncation, Sample Selection, and Limited dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 175-92.

[6] Heckman, J. (1990) "Varieties of Slection Bias," *American Economic Review*, 80, 313-18.

[7] Johnston, J. and DiNardo, J. (1997) *Econometric Methods*, 4th ed, New York: McGraw-Hill.