〈論　文〉

打ち切り・切断があるデータに対する密度推定を利用した回帰直線の推定

中　村　永　友

概　　要

本報告は，従属変数に途中打ち切りや切断があるデータに対して線形回帰モデルをあてはめるとき，合計 6 種類のモデルを示しその違いについて言及する。2 種類の既存のモデルと 4 種類のモデルを提案し，これらを比較する。第 1 の方法は n 組の完全データと m 組の不完全データを使う途中打ち切り回帰モデルで，これは残差項に切断正規分布を仮定して，切断点より大きくなる確率をモデルに入れる方法である。打ち切り点以上の確率を推定することで，不完全データの情報として付加する方法である。第 2 は n 組の完全データのみを使う切断回帰モデルで，前項と同じく切断正規分布を利用するが，完全データと切断点の情報を使うモデルである。第 3，4 のモデルは，全変数の同時密度に対して多変量正規分布を仮定し，欠測データ数 m の情報を用いるものと，それが未知であるときに推定するモデルである。第 5，6 は全変数の同時密度に対して本論文で提案する一様・多変量正規分布を仮定し，m を用いるものと未知であるときに推定するモデルである。

提案モデルの妥当性を数値実験で検証し，欠測個数の情報を有効に使うことで十分有効な推定ができることが示された。

1　はじめに

1.1　従属変数に欠測がある回帰モデル

欠測を伴うデータに対して線形回帰モデルをあてはめるとき，途中打ち切り回帰モデル（Censored Regression Model）や切断回帰モデル（Truncated Regression Model）が利用される。これらのモデルは，計量経済学ではスタンダードな統計モデルとなっていて，種々の統計ソフトウェア（TSP など）で実装されている。これらのモデルの本質的な違いは不完全データの扱いである。途中打ち切り回帰モデルは独立変数が観測され，従属変数の観測が欠測している場合のモデルで，切断回帰モデルは両変数が観測されない場合のモデルである。計量経済学の論文やテキスト（Amemiya 1985, Greene 2002, Maddala 1984 など）では，途

中値打ち切り回帰モデルはトービットモデル（Tobit model; Tobin 1958）として紹介されている。

本論文では，独立変数と従属変数の同時密度に多変量の密度関数を仮定する密度推定の視点を入れたモデリングを行う。これは打ち切り回帰モデルや切断回帰モデルでは扱うことの出来ない欠測データ数の情報を取り入れたり，その数が未知のときに推定できるモデルである。

密度推定の視点を入れることの根拠は以下の通りである。通常の線形回帰モデルは，独立変数 $\boldsymbol{X}$ や従属変数 Y に分布の仮定はしないが，$\boldsymbol{X}$ と Y の同時分布 $f(\boldsymbol{X}, Y)$ には $f(\boldsymbol{X}, Y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ という暗黙の仮定をしていることと同値である。本論文はこの暗黙の仮定を前面に出して，独立変数と従属変数の同時密度関数を適当な密度関数を想定し，その密度関数のパラメータの推定値から回帰パラメータ等を導出する方法を提案する。

ここで用いる密度関数は多変量正規分布と，独立変数が $[\alpha, \beta]$ 区間で一様に分布している $X \sim U(\alpha, \beta)$ を仮定した，一様・正規分布モデルを用いる。後者の密度関数によるモデルは，欠測の状況がない場合には通常の線形回帰モデルと同値となる。

欠測データ数が既知の場合のパラメータ推定方法と，未知の場合のパラメータ推定法と欠測数を推定する方法を提案する。これらの回帰モデルをまとめると次の 6 種類となる。

ケース 1．途中打ち切り回帰モデル：従属変数が欠測し，欠測データ数が既知
ケース 2．切断回帰モデル：独立・従属変数共に欠測
ケース 3．密度推定型切断回帰モデル（多変量正規）：独立･従属変数共に欠測し，欠測データ数が既知
ケース 4．密度推定型切断回帰モデル（多変量正規）：独立･従属変数共に欠測し，欠測データ数が未知
ケース 5．密度推定型切断回帰モデル（一様・正規）：独立・従属変数共に欠測し，欠測データ数が既知
ケース 6．密度推定型切断回帰モデル（一様・正規）：独立・従属変数共に欠測し，欠測データ数が未知

以上の状況は取り扱うデータセットにより表 1 次のように表すことができる。

ケース 1 は途中値打ち切り回帰モデル，ケース 2 は切断回帰モデルと呼ばれ，共にトービットモデルとして知られている。ケース 3～6 の欠測値の個数が既知・未知の状況は，これまでモデルとして存在しなかったものである。ケース 3，5 は厳密な意味では欠測数が既知なので，途中打ち切り回帰モデルの 1 種と見なせるであろう。さらにケース 4，6 の欠測値の

表 1 ：データの欠測状況

ケース	モデル	観測データ 完全データ	観測データ 不完全データ
1	途中打ち切り回帰モデル	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$,	$\underbrace{(\boldsymbol{x}_{n+1}, \cdot), \ldots, (\boldsymbol{x}_{n+m}, \cdot)}_{m\text{個}}$
2	切断回帰モデル	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$	
3	密度推定型切断回帰モデル（多変量正規分布）	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$,	$\underbrace{(\cdot, \cdot), \ldots, (\cdot, \cdot)}_{m\text{個：既知}}$
4	密度推定型切断回帰モデル（多変量正規分布）	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$,	$\underbrace{(\cdot, \cdot), \ldots, (\cdot, \cdot)}_{m\text{個：未知}}$
5	密度推定型切断回帰モデル（一様・正規分布）	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$,	$\underbrace{(\cdot, \cdot), \ldots, (\cdot, \cdot)}_{m\text{個：既知}}$
6	密度推定型切断回帰モデル（一様・正規分布）	$\underbrace{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)}_{n\text{個}}$,	$\underbrace{(\cdot, \cdot), \ldots, (\cdot, \cdot)}_{m\text{個：未知}}$

n 組が正常に観測され（完全データ），? 個または m 組が不完全な形（不完全データ）で記録されていることを表す。

便宜上完全データ n 個を前に，不完全データ m 個を後ろに並べているだけで，添え字の並びに意味のある時系列的なデータではない。

個数が未知のモデルは，ケース 2 の切断回帰モデルと一見同一に見えるが，これは欠測数を推定することを考慮するモデルであるので，この点で異なる。

切断回帰モデルのパラメータ推定値は，打ち切り回帰モデルほどうまく補正されないことが報告されている（縄田 1997）。これからもわかるように，切断回帰モデルは打ち切り部分の情報がないために，十分なパラメータ推定の補正ができないモデルである。また，途中打ち切り回帰モデルや切断回帰モデルは，欠測データ数のみが既知であってもそれをモデルの中に考慮していない。そこで，本論文はこれらのモデルの弱点を克服するために，欠測データ数が既知・未知のときに密度推定の方法を使って，できるだけ偏りのない回帰モデルのパラメータ推定の方法を提案する。提案モデルは途中打ち切り回帰モデルと切断回帰モデルの中間的な位置付けとなる。また，ケース 3， 4 は独立変数が正規分布を，ケース 5， 6 は独立変数が一様分布を仮定することになる。

1.2　モデルの考え方

本論文で扱うデータの観測状況は，(1)従属変数 y_i がある値 c より大きな値で観測できない，(2)独立変数が正規分布，または一様分布に従う，という条件を想定する。したがって本論文で扱う欠測は無視できない欠測であることが前提となる。

通常の最小自乗法による（重）回帰モデルの回帰パラメータの推定は，残差のみに正規分布を仮定し，独立変数や従属変数には分布を仮定しない。しかし，その回帰パラメータの導

出には，各変数の分散や共分散の推定値を用いているため，結果として独立変数と従属変数の同時密度関数が

$$f(x, y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

と仮定したときのパラメータ推定値 $\hat{\boldsymbol{\mu}}$ と $\hat{\boldsymbol{\Sigma}}$ を用いるものになる。つまり図１で説明すると，２次元正規分布の確率密度関数を表す楕円をデータに対して想定して，直線 l を推定することになる。そのとき，仮に，〝X～正規分布〟でなかったり，あるいは〝X～一様分布〟という仮定があったとしても，最小自乗法では(1)式を仮定してパラメータ推定してもよいことになる。

しかし，独立変数が正規分布でなく，一様分布に従うと考えられ，従属変数が途中打ち切りや切断がある場合には，考慮しなければいけない点がある。それは図１の領域Ａのデータの存在を無視してはいけない点である。途中打ち切り回帰モデルの場合には，独立変数の情報から領域Ａを考慮することはできるが，切断回帰モデルや密度推定型切断回帰モデルＡで(1)式を仮定したとき，領域Ａを考慮することはできない。

ケース２の切断回帰モデルでは，独立変数がどのような分布にしたがうかは関係なく，残差分布に切断正規分布を単に仮定するモデルであり，完全データのみで完結しているモデルである。したがって，図１の領域ＡやＢは直接的に考慮していないモデルである。

本論文は切断回帰モデルに密度推定の視点を入れて，(1)独立変数が一様分布を想定し，(2)欠測数が既知の場合と未知の場合のモデル構築し，(3)欠測数が未知の場合にはその欠測数を推定する方法を議論することで，結果としてできるだけ偏りのないパラメータ推定の方法を提案する。

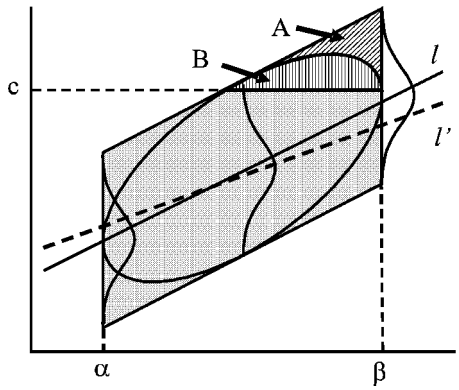

図１：データのモデリング

推定したい直線は l。欠測を考慮しない場合は l' のように回帰パラメータが偏って推定される。

通常の線形回帰モデルのパラメータ推定は，多変量正規分布を仮定した密度推定結果のパラメータを使って推定することになる。しかし，独立変数が一様分布を仮定するときには，図のような平行四辺形の領域にデータが分布していると仮定し，領域Ａがあることをモデリングしなければいけない。

以下，2 節では，本論文で扱うモデルとそのパラメータ推定方法について説明する。3 節では，提示したモデルの数値実験を通してモデルの比較を行う。

2　不完全データによる回帰モデルの構築

本節では，6 つのモデルとそれぞれのパラメータ推定方法について紹介する。各パラメータ推定方法は EM 法（Dempster et. al 1977）を前提にしたものである。各パラメータで解いた式は〝新しい推定値＝h（1 つ前の推定値)〞という形式で示している。

2.1　途中打ち切り回帰モデル（ケース 1）

途中打ち切り回帰モデルは，従属変数がある値以上（以下）で観測できないという状況のデータを扱う。それは以下のように表現できる。

$$y_i=\begin{cases} y_i^* & y_i^*\le c \\ \cdot & y_i^*>c \end{cases} \tag{2}$$

このとき

$$y_i^*=ax_i+b+e_i,\quad e_i\sim N(0,\sigma^2) \tag{3}$$

のモデルを想定するとき，以下のように尤度関数が構成され，パラメータ推定ができる。

(2)，(3)式を考慮するとき，ある点 x において残差が閾値 c 以上になる確率は，

$$Pr(Y>c)=1-\boldsymbol{\Phi}\left(\frac{c-ax-b}{\sigma}\right)=\boldsymbol{\Phi}\left(-\frac{c-ax-b}{\sigma}\right)$$

で求められる。$\boldsymbol{\Phi}(\cdot)$は標準正規分布の分布関数である。

このとき尤度関数は次式となる。

$$\begin{aligned} L_1&=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{e_i^2}{2\sigma^2}\right\}\times\prod_{i=n+1}^{n+m}\int_{(c-ax_i-b)/\sigma}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{t^2}{2}\right\}dt \\ &=\prod_{i=1}^{n}\frac{1}{\sigma}\phi\left(\frac{e_i}{\sigma}\right)\times\prod_{i=n+1}^{n+m}\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right). \end{aligned} \tag{4}$$

ここで$\phi(\cdot)$は標準正規分布の密度関数である。

$\lambda=\sigma^2$ とおいて，対数尤度関数を以下のように書くことができる：

$$l_1=\log L_1=-\frac{n}{2}\log 2\pi-\frac{n}{2}\log\lambda-\frac{1}{2\lambda}\sum_{i=1}^{n}(y_i-ax_i-b)^2+\sum_{i=n+1}^{n+m}\log\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sqrt{\lambda}}\right). \tag{5}$$

未知パラメータを推定するために，極値を求めると以下のようになる：

$$\frac{\partial l_1}{\partial \lambda}=-\frac{n}{2\lambda}+\frac{1}{2\lambda^2}\sum_{i=1}^{n}(y_i-ax_i-b)^2+\frac{1}{2\lambda^{3/2}}\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sqrt{\lambda}}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sqrt{\lambda}}\right)}(c-ax_i-b)=0,$$

$$-n\lambda+B\sqrt{\lambda}+R=0,$$

$$-n\sigma^2+B\sigma+R=0.$$

このように $\sigma=\sqrt{\lambda}$ に関する 2 次方程式となるので，残差分散を以下のように解くことができる：

$$\sigma=\sqrt{\lambda}=\frac{1}{2n}\left\{B\pm\sqrt{B^2+4nR}\right\}. \tag{6}$$

ここで，

$$B=\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right)}(c-ax_i-b),\quad R=\sum_{i=1}^{n}(y_i-ax_i-b)^2$$

である。複合の負号では$\sigma<0$ となるため無効である。

同様に回帰係数 a と定数項 b について極値を求めると以下のようになる：

$$\frac{\partial l_1}{\partial a}=\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i-ax_i-b)+\frac{1}{\sigma}\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right)}x_i=0,$$

$$\frac{\partial l_1}{\partial b}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)+\frac{1}{\sigma}\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right)}=0.$$

これらを a，b について解くと以下の式を得る：

$$a=\frac{1}{\sum_{i=1}^{n}(x_i-\tilde{x})^2}\left\{\sigma\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right)}(x_i-\tilde{x})+\sum_{i=1}^{n}(x_i-\tilde{x})(y_i-\tilde{y})\right\}, \tag{7}$$

$$b=\frac{\sigma}{n}\sum_{i=n+1}^{n+m}\frac{\phi\left(-\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(-\frac{c-ax_i-b}{\sigma}\right)}+\tilde{y}-a\tilde{x}. \tag{8}$$

ここで，$\tilde{x}=n^{-1}\sum_{i=1}^{n}x_i$, $\tilde{y}=n^{-1}\sum_{i=1}^{n}y_i$ は，完全データのみで計算される擬平均である。

このように，回帰モデルのパラメータを推定することができる。

2.2　切断回帰モデル（ケース 2）

切断回帰モデルは，閾値 c 以下のみの観測データのみを用いる回帰モデルである。それは

(2)式の仮定の下に，独立変数も観測できないという状況である。その場合の回帰モデルを

$$y_i^*=ax_i+b+e_i,\quad y_i^*\leq c,\quad e_i\sim N_t(0,\sigma^2)$$

とおくことができる。ここでは c 以上で y は観測できないので，残差項が切断正規分布 $N_t(\ )$ にしたがっていると仮定する統計モデルである。切断正規分布はある値 c 以上で分布が切断されている正規分布で，その密度関数は

$$N_t(0,\sigma^2):N_t(t|0,\sigma^2;t<c)=\frac{\sigma}{\boldsymbol{\Phi}(c/\sigma)}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{t^2}{2\sigma^2}\right\}$$

で定義される。これより尤度関数を構成すると，次式のようになる：

$$\begin{aligned}L_2&=\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{e_i^2}{2\sigma^2}\right\}\times\left\{\int_{-\infty}^{(c-ax_i-b)/\sigma}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{t^2}{2}\right\}dt\right\}^{-1}\right]\\&=\prod_{i=1}^{n}\frac{1}{\sigma}\phi\left(\frac{e_i}{\sigma}\right)\times\prod_{i=1}^{n}\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)^{-1}.\end{aligned}\tag{9}$$

$\nu=\sigma^2$ とおいて，対数尤度関数を以下のように書くことができる：

$$l_2=\log L_2=-\frac{n}{2}\log 2\pi-\frac{n}{2}\log\nu-\frac{1}{2\nu}\sum_{i=1}^{n}(y_i-ax_i-b)^2-\sum_{i=1}^{n}\log\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sqrt{\nu}}\right).\tag{10}$$

まず，残差分散σ^2 について解く：

$$\begin{aligned}\frac{\partial l_2}{\partial\nu}=-\frac{n}{2\nu}+\frac{1}{2\nu^2}\sum_{i=1}^{n}(y_i-ax_i-b)^2-\frac{1}{2\nu^{3/2}}\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sqrt{\nu}}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sqrt{\nu}}\right)}(c-ax_i-b)=0,\\-n\nu-B'\sqrt{\nu}+R'=0,\\-n\sigma^2-B'\sigma+R'=0.\end{aligned}$$

このように $\sigma=\sqrt{\nu}$ に関する 2 次方程式となるので，以下のように解くことができる：

$$\sigma=\sqrt{\nu}=\frac{1}{2n}\left\{-B'\pm\sqrt{B'^2+4nR'}\right\}.\tag{11}$$

ここで，

$$B'=\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)}(c-ax_i-b),\quad R'=\sum_{i=1}^{n}(y_i-ax_i-b)^2$$

である。複合の負号では $\sigma<0$ となるため無効である。

同様に a，b について極値を求めると以下のようになる：

$$\frac{\partial l_2}{\partial a}=\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i-ax_i-b)+\frac{1}{\sigma}\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)}x_i=0,$$

$$\frac{\partial l_2}{\partial b}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-ax_i-b)+\frac{1}{\sigma}\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)}=0.$$

これらを a，b について解くと以下の式を得る：

$$a=\frac{1}{\sum_{i=1}^{n}(x_i-\tilde{x})^2}\left\{\sigma\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)}(x_i-\tilde{x})+\sum_{i=1}^{n}(x_i-\tilde{x})(y_i-\tilde{y})\right\}, \tag{12}$$

$$b=\frac{\sigma}{n}\sum_{i=1}^{n}\frac{\phi\left(\frac{c-ax_i-b}{\sigma}\right)}{\boldsymbol{\Phi}\left(\frac{c-ax_i-b}{\sigma}\right)}+\tilde{y}-a\tilde{x}. \tag{13}$$

ここで，$\tilde{x}=n^{-1}\sum_{i=1}^{n}x_i$，$\tilde{y}=n^{-1}\sum_{i=1}^{n}y_i$ は，完全データのみで計算される擬平均である。

このように σ, a, b に関して解いた式を漸化式として見なして，右辺の推定したいパラメータに現在の推定値を代入して，新たな推定値を計算する。これを繰り返して収束した値をパラメータを推定することができる。

2.3 楕円密度推定型切断回帰モデル（ケース 3， 4）

回帰モデルをあてはめるようなデータに対して，密度推定の視点でデータに確率密度関数をあてはめて，回帰式を求めることができる。つまり，独立変数 $\boldsymbol{x}$ と従属変数 y の同時密度関数 $f(\boldsymbol{x}, y)$に多変量正規分布を想定し，同時密度関数の平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$を求めると回帰パラメータが推定できる。この方法による切断回帰モデルのパラメータを推定する方法を以下に示す。

独立変数 $\boldsymbol{x}=(x_1, ..., x_d)$と従属変数 y をまとめて $\boldsymbol{z}=(x_1, ..., x_d, y)^T$ と表したとき，$\boldsymbol{z}$ の密度関数（独立変数と従属変数の同時密度関数）は，$d+1$ 次元正規分布となる：

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right\}$$

ここで，

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_j\\ \vdots\\ \mu_d\\ \mu_y\end{pmatrix},\quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{21} & \cdots & \sigma_{j1} & \cdots & \sigma_{d1} & \sigma_{y1}\\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{j2} & \cdots & \sigma_{d2} & \sigma_{y2}\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots\\ \sigma_{1j} & \sigma_{2j} & \vdots & \sigma_j^2 & \vdots & \sigma_{dj}^2 & \sigma_{yj}\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots\\ \sigma_{1d} & \sigma_{2d} & \cdots & \sigma_{jd} & \cdots & \sigma_d^2 & \sigma_{yd}\\ \sigma_{1y} & \sigma_{2y} & \cdots & \sigma_{jy} & \cdots & \sigma_{dy} & \sigma_y^2\end{pmatrix}$$

であり，添字の $1, ..., j, ..., d$ は d 次元の独立変数 x_i を，y は従属変数を表す。

単回帰モデル $y=ax+b$ のとき，

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{pmatrix}$$

となり，これらの推定値を $\hat{}$ で表すとき，回帰パラメータ $\hat{a}$，$\hat{b}$ と残差分散 $\hat{\sigma}^2$ は以下のように推定できる。

$$\hat{a}=\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2},\quad \hat{b}=\hat{\mu}_y-\hat{a}\hat{\mu}_x,\quad \hat{\sigma}^2=\hat{\sigma}_y^2-2\hat{a}\hat{\sigma}_{xy}+a^2\hat{\sigma}_x^2,$$

$$V(e)=E[e^2]-E[e]^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i-b)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{(y_i-\bar{y})^2-2a(x_i-\bar{x})(y_i-\bar{y})+a^2(x_i-\bar{x})^2\right\}=\sigma_y^2-2a\sigma_{xy}+a^2\sigma_x^2.$$

図 1 に示すように従属変数が値 c 以上で観測されず，同時に独立変数も欠測してる状況のとき，以下の手続きで密度関数のパラメータを推定することができる。データを $\boldsymbol{z}_i=(\boldsymbol{x}_i, y_i)$，$\boldsymbol{Z}_n=\{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n\}$ とする。（添字 i はデータの個体番号を表し，j は独立変数の変数の番号を表す。）

⑴欠測数 m が既知のときのパラメータ推定手順（ケース 3 ）

欠測数 m が既知の場合，欠測領域の確率を

$$p=\int_R f(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{t}$$

として，対数尤度関数

$$\log L_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Z}_n, m, p)=\log\left[\prod_{i=1}^{n}f(\boldsymbol{z}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\times p^m\right]=\sum_{i=1}^{n}\log f(\boldsymbol{z}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})+m\log p$$

の最大化

$$\underset{\boldsymbol{\mu},\,\boldsymbol{\Sigma}}{\operatorname{argmax}} \quad \log L_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Z}_n, m, p)$$

によって未知パラメータの推定ができる。

⑵欠測数 m が未知のときのパラメータ推定手順（ケース 4 ）

欠測数 m が未知のときは，以下の方法でパラメータ推定を行う。

1．現在のパラメータの推定値 $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ を固定して，次式で欠測数の推定を行う。

$$\hat{m}=\frac{n\hat{p}}{1-\hat{p}}.$$

これは欠測領域の確率 p が $\frac{m}{n+m}$ で推定できることを使ったものである。

2．$m=\hat{m}$ として，対数尤度

$$\log L_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Z}_n, \hat{m}, \hat{p})=\sum_{i=1}^{n}\log(\boldsymbol{z}_i|\boldsymbol{\mu}, \Sigma)+\hat{m}\log\hat{p}$$

の最大化

$$\underset{\boldsymbol{\mu},\,\boldsymbol{\Sigma}}{\operatorname{argmax}} \quad \log L_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Z}_n, \hat{m})$$

によって $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ を得る。

3．1.と 2.を収束条件を満足するまで繰り返す。

ここで，$\hat{p}=p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$は欠測領域における確率密度関数の確率である。独立変数の定義域を $\boldsymbol{X}\in[-\infty, \infty]^d$，従属変数の欠測領域を $y\in[c, \infty]$とするとき，欠測欠損領域は $R=[-\infty, \infty]^d \cap[c, \infty]$となり，次の式で与えられる。

$$\hat{p}=p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})=\int_R f(\boldsymbol{z}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})d\boldsymbol{z}.$$

単回帰モデルのとき独立変数の定義域は $x\in[-\infty, \infty]$，従属変数の欠測領域は $y\in[c, \infty]$となるので，欠測領域は $R=[-\infty, \infty]\cap[c, \infty]$となり，以下のように書くことができる。

$$p=\int_c^{\infty}\int_{-\infty}^{\infty} f(x, y|\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})dxdy=1-\boldsymbol{\Phi}\left(\frac{c-\mu_y}{\sigma_y}\right)=\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right).$$

以上の準備の下で，平均ベクトルと分散共分散行列の EM 法の第 $\ell+1$ ステップの更新式は以下のようになる（中村他，2005）。

$$\begin{aligned}\boldsymbol{\mu}^{(\ell+1)}&=\frac{n}{n+m^{(\ell)}}\tilde{\boldsymbol{\mu}}+\frac{m^{(\ell)}}{n+m^{(\ell)}}\frac{\int_R \boldsymbol{t}f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})d\boldsymbol{t}}{\int_R f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})d\boldsymbol{t}}\\&=\frac{n}{n+m^{(\ell)}}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix}+\frac{m^{(\ell)}}{n+m^{(\ell)}}\frac{1}{\int_R f(u, t|\mu_x^{(\ell)}, \mu_y^{(\ell)}, \sigma_x^{2(\ell)}, \sigma_y^{2(\ell)}, \sigma_{xy}^{(\ell)})dtdu}\end{aligned}$$

$$\times\begin{pmatrix}\int_R uf(u,\,t|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu\\ \int_R tf(u,\,t|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu\end{pmatrix},$$

$$\boldsymbol{\Sigma}^{(\ell+1)}=\frac{n}{n+m^{(\ell)}}\tilde{\boldsymbol{\Sigma}}^{(\ell)}+\frac{m^{(\ell)}}{n+m^{(\ell)}}\frac{\int_R(\boldsymbol{t}-\boldsymbol{\mu}^{(\ell)})(\boldsymbol{t}-\boldsymbol{\mu}^{(\ell)})^T f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})d\boldsymbol{t}}{\int_R f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})d\boldsymbol{t}}$$

$$=\frac{n}{n+m^{(\ell)}}\begin{pmatrix}\tilde{\sigma}_x^2 & \tilde{\sigma}_{xy}\\ \tilde{\sigma}_{xy} & \tilde{\sigma}_y^2\end{pmatrix}$$

$$+\frac{m^{(\ell)}}{n+m^{(\ell)}}\frac{1}{\int_R f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})d\boldsymbol{t}}$$

$$\times\begin{pmatrix}\int_R u^2 f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu & \int_R utf(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu\\ \int_R utf(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu & \int_R t^2 f(\boldsymbol{t}|\boldsymbol{\mu}^{(\ell)},\,\boldsymbol{\Sigma}^{(\ell)})dtdu\end{pmatrix}.$$

ここで，$\tilde{\boldsymbol{\mu}}=n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$，$\tilde{\boldsymbol{\Sigma}}^{(\ell)}=n^{-1}\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu}^{(\ell)})(\boldsymbol{x}_i-\boldsymbol{\mu}^{(\ell)})^T$ は，観測されたデータから計算される擬平均ベクトルと擬分散共分散行列，$(\ )^T$ はベクトルの転置である。

単回帰モデルの場合の種々の期待値は以下のように書き下せる：

$$E[X|R]=\int_c^\infty\int_{-\infty}^\infty xf(x,\,y|\mu_x,\,\mu_y,\,\sigma_x^2,\,\sigma_y^2,\,\sigma_{xy})dxdy=\frac{\sigma_{xy}}{\sigma_y}\phi\left(\frac{c-\mu_y}{\sigma_y}\right)+\mu_x\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right),$$

$$E[Y|R]=\int_c^\infty\int_{-\infty}^\infty yf(x,\,y|\mu_x,\,\mu_y,\,\sigma_x^2,\,\sigma_y^2,\,\sigma_{xy})dxdy=\sigma_y\phi\left(\frac{c-\mu_y}{\sigma_y}\right)+\mu_y\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right),$$

$$E[X^2|R]=\int_c^\infty\int_{-\infty}^\infty x^2f(x,\,y|\mu_x,\,\mu_y,\,\sigma_x^2,\,\sigma_y^2,\,\sigma_{xy})dxdy$$

$$=\frac{\sigma_{xy}(2\mu_x\sigma_y^2+(c-\mu_y)\sigma_{xy})}{(\sigma_y^2)^{3/2}}\phi\left(\frac{c-\mu_y}{\sigma_y}\right)+(\mu_x^2+\sigma_x^2)\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right)$$

$$E[Y^2|R]=\int_c^\infty\int_{-\infty}^\infty y^2f(x,\,y|\mu_x,\,\mu_y,\,\sigma_x^2,\,\sigma_y^2,\,\sigma_{xy})dxdy$$

$$=\sigma_y(\mu_y+c)\phi\left(\frac{c-\mu_y}{\sigma_y}\right)+(\mu_y^2+\sigma_y^2)\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right),$$

$$E[XY|R]=\int_c^\infty\int_{-\infty}^\infty xyf(x,\,y|\mu_x,\,\mu_y,\,\sigma_x^2,\,\sigma_y^2,\,\sigma_{xy})dxdy$$

$$=\frac{\mu_x\sigma_y^2+c\sigma_{xy}}{\sigma_y}\phi\left(\frac{c-\mu_y}{\sigma_y}\right)+(\mu_x\mu_y+\sigma_{xy})\boldsymbol{\Phi}\left(-\frac{c-\mu_y}{\sigma_y}\right).$$

ここで，$\sigma_y=\sqrt{\sigma_y^2}$，$\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$，$\boldsymbol{\Phi}(z)=\int_{-\infty}^z\phi(t)dt$ である。

2.4　一様・正規密度推定型切断回帰モデル（ケース 4，5）

独立変数 X に一様分布を仮定し，残差項に切断正規分布を仮定するモデルを以下のように作ることができる。それは密度関数を

$$f(x, y|a, b, \sigma^2, \alpha, \beta)=\frac{1}{(\beta-\alpha)\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-ax-b)^2}{2\sigma^2}\right\}$$

としてデータにモデルのあてはめを行い，パラメータ a，b，σ^2 を推定することである。打ち切りや切断がない場合には，この密度推定法で推定されたパラメータは，通常の回帰モデルと同値の結果を与える。

独立変数 X が一様分布を仮定したときに，打ち切り領域 R における確率は，

$$p=p(a, b, \sigma^2, \alpha, \beta)=\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2, \alpha, \beta)dtdu$$

となる。密度推定の観点から，パラメータ a，b，σ^2 が何らかの形で与えられた（推定された）とき，打ち切り領域 R の確率 p と観測されているデータ数 n を使って，R における欠測数を

$$\hat{m}=\hat{m}(\hat{a}, \hat{b}, \hat{\sigma}^2)=\frac{n\hat{p}}{1-\hat{p}}$$

で推定することができる。

m を欠測数，または推定された欠測数とするとき，尤度関数と対数尤度関数は以下のようになる。

$$L_4=L_4(a, b, \sigma|X, Y)=\prod_{i=1}^n f(x_i, y_i|a, b, \sigma^2)\times\left\{\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu\right\}^m,$$

$$\begin{aligned}\log L_4(a, b, \sigma|X, Y)&=\sum_{i=1}^n \log f(x_i, y_i|a, b, \sigma^2)+m\log\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu\\&=\sum_{i=1}^n\{\log(\beta-\alpha)-\frac{1}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(y_i-ax_i-b)^2\}\\&\quad+m\log\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu.\end{aligned}$$

対数尤度関数をパラメータ a，b，σ^2 で偏微分して 0 とおき，当該パラメータで解くと以下のようになる：

$$\frac{\partial\log L_4}{\partial a}=\sum_{i=1}^n\frac{x_i}{\sigma^2}(y_i-ax_i-b)+\frac{m}{\sigma^2}\frac{\int_\alpha^\beta\int_c^\infty u(t-au-b)f(u, t|a, b, \sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu}=0,$$

$$\frac{\partial\,\boldsymbol{log}\,L_4}{\partial b}=\sum_{i=1}^n\frac{1}{\sigma^2}(y_i-ax_i-b)+\frac{m}{\sigma^2}\frac{\int_\alpha^\beta\int_c^\infty (t-au-b)f(u, t|a, b, \sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu}=0,$$

$$\frac{\partial\,\boldsymbol{log}\,L_4}{\partial\sigma^2}=-\frac{n+m}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^n(y_i-ax_i-b)^2+\frac{m}{2\sigma^4}\frac{\int_\alpha^\beta\int_c^\infty (t-au-b)^2f(u, t|a, b, \sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u, t|a, b, \sigma^2)dtdu}=0.$$

以上より，a，b，σ^2 を次のように解くことが出来る。

$$a=\frac{n\tilde{w}+mW-\frac{1}{n+m}(n\tilde{y}+mT)(n\tilde{x}+mU)}{n\tilde{x}_2+mU_2-\frac{1}{n+m}(n\tilde{x}+mU)^2},\quad b=\frac{1}{n+m}\{n\tilde{y}+mT-a(n\tilde{x}+mU)\},$$

$$\sigma^2=\frac{1}{n+m}\{n\tilde{\sigma}^2+mQ\}.$$

ここで，各記号は以下の通りである：

$$\tilde{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \tilde{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\quad \tilde{x}_2=\frac{1}{n}\sum_{i=1}^{n}x_i^2,\quad \tilde{y}_2=\frac{1}{n}\sum_{i=1}^{n}y_i^2,\quad \tilde{w}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i,\quad \tilde{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i-b)^2,$$

$$T=\frac{T'}{V}=\frac{\int_\alpha^\beta\int_c^\infty tf(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu},\quad U=\frac{U'}{V}=\frac{\int_\alpha^\beta\int_c^\infty uf(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu},$$

$$T_2=\frac{T_2'}{V}=\frac{\int_\alpha^\beta\int_c^\infty t^2f(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu},\quad U_2=\frac{U_2'}{V}=\frac{\int_\alpha^\beta\int_c^\infty u^2f(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu},$$

$$W=\frac{W'}{V}=\frac{\int_\alpha^\beta\int_c^\infty tuf(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu},$$

$$Q=\frac{Q'}{V}=\frac{\int_\alpha^\beta\int_c^\infty (t-au-b)^2f(u,t|a,b,\sigma^2)dtdu}{\int_\alpha^\beta\int_c^\infty f(u,t|a,b,\sigma^2)dtdu}=a^2U_2+2abU-2aW-2bT+T_2+\frac{b^2}{V},$$

$$\begin{aligned}V=\frac{1}{a(\beta-\alpha)}\Big\{&-\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)+\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)\\&+(c-a\alpha-b)\mathbf{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)-(c-a\beta-b)\mathbf{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},\end{aligned}$$

$$\begin{aligned}T'=\frac{1}{2a(\beta-\alpha)}\Big\{&-(c+a\alpha+b)\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)+(c+a\beta+b)\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)\\&-\Big((c+a\alpha+b)(-c+a\alpha+b)+\sigma^2\Big)\mathbf{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)\\&+\Big((c+a\beta+b)(-c+a\beta+b)+\sigma^2\Big)\mathbf{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},\end{aligned}$$

$$\begin{aligned}U'=\frac{1}{2a^2(\beta-\alpha)}\Big\{&-(c+a\alpha-b)\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)+(c+a\beta-b)\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)\\&+\Big((c+a\alpha-b)(c-a\alpha-b)+\sigma^2\Big)\mathbf{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)\\&-\Big((c+a\beta-b)(c-a\beta-b)+\sigma^2\Big)\mathbf{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},\end{aligned}$$

$$T_2'=\frac{1}{3a(\beta-\alpha)}\Big\{-\Big(b^2+bc+c^2+a\alpha(2b+c+a\alpha)+2\sigma^2\Big)\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$+\Big(b^2+bc+c^2+a\beta(2b+c+a\beta)+2\sigma^2\Big)\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)$$
$$-\Big(b^3-c^3+a^3\alpha^3+3(b+a\alpha)(ab\alpha+\sigma^2)\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$+\Big(b^3-c^3+a^3\beta^3+3(b+a\beta)(ab\beta+\sigma^2)\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},$$

$$U'_2=\frac{1}{3a^3(\beta-\alpha)}\Big\{-\Big((b-c)^2-a\alpha(b-c-a\alpha)+2\sigma^2\Big)\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$+\Big((b-c)^2-a\beta(b-c-a\beta)+2\sigma^2\Big)\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)$$
$$-\Big((b-c)^3+a^3\alpha^3+3(b-c)\sigma^2\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$+\Big((b-c)^3+a^3\beta^3+3(b-c)\sigma^2\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},$$

$$W'=\frac{1}{6a^2(\beta-\alpha)}\Big\{\Big(b^2+b(c-a\alpha)-2(c^2+ac\alpha+a^2\alpha^2-\sigma^2)\Big)\sigma\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$-\Big(b^2+b(c-a\beta)-2(c^2+ac\beta+a^2\beta^2-\sigma^2)\Big)\sigma\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)$$
$$+\Big(b^3+2(c^3-a^3\alpha^3)-3b(c^2+a^2\alpha^2-\sigma^2)\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)$$
$$-\Big(b^3+2(c^3-a^3\beta^3)-3b(c^2+a^2\beta^2-\sigma^2)\Big)\boldsymbol{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\},$$

$$Q'=b^2V+\frac{1}{a(\beta-\alpha)}\Big[(b^2-2\sigma^2)\sigma\Big\{\phi\Big(-\frac{c-a\alpha-b}{\sigma}\Big)-\phi\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\}$$
$$-(b^2-\sigma^2)\Big\{(c-a\alpha-b)\boldsymbol{\Phi}\Big(-\frac{c-a\alpha-b}{\sigma}\Big)-(c-a\beta-b)\boldsymbol{\Phi}\Big(-\frac{c-a\beta-b}{\sigma}\Big)\Big\}\Big].$$

3　数 値 実 験

3.1　実験条件

提案したモデルのふるまいを数値実験で検証した。

設定 1：平均ベクトル：$\boldsymbol{\mu}=(\mu_x,\mu_y)^T=(5.5,7.4)^T$，分散共分散行列：

$$\boldsymbol{\Sigma}=\begin{pmatrix}S^2x & \cdot\\ Sxy & S^2y\end{pmatrix}=\begin{pmatrix}2.9 & \cdot\\ 2.9\times0.8 & 2.9\end{pmatrix}$$

の 2 次元正規分布に従う乱数を生成。データ数 $N=500$，繰り返し数 $S=10000$，閾値を $c=10$ として，従属変数がこの値より大きいとき条件に合わせて欠測させる。

設定 2：$x\in[10,80]$で一様乱数を生成し，回帰直線 $y=0.225x+12$ に誤差項 $e_i\sim N(0,5^2)$を加えた。データ数 $N=500$，繰り返し数 $S=10000$，閾値を $c=30$ として，従属変数がこの値より大きいとき条件に合わせて欠測させる。

3.2　結果

実験の結果を表 2 に示す。

設定 1 の実験結果から以下のことがわかる。(1)打ち切り回帰モデルは独立変数の情報があるため，回帰係数，定数項，残差分散はほぼ真値で推定されている。(2)切断回帰モデルは提案モデルも含めて，思うほどうまく推定されていない。真値から遠いということ。(3) 4 種類の提案モデルを真値への近さを基準とした場合に，以下の序列ができる。

打ち切り回帰＞正規密度推定 2 ＞正規密度推定 1 ＞一様・正規密度推定 2 ＞切断回帰≫一様・正規密度推定 1

(4)一様・正規密度推定 1 のモデルの推定結果が悪い理由は，図 1 で説明すれば，領域 $\boldsymbol{A}$ に本来データが存在しないのに，そこに「ある」という仮定をおいたモデルだからである。推定した欠測領域確率を見れば明らかである。同様に一様・正規密度推定 2 の確率も多めに推定されている。

設定 2 の実験結果から以下のことがわかる。(1)打ち切り回帰モデルは独立変数の情報があるため，回帰係数，定数項はほぼ真値で推定されているが，残差分散は小さめである。(2)切断回帰モデルは提案モデルも含めて，いちばんよくない。(3) 4 種類の提案モデルを真値への近さを基準とした場合に，以下の序列ができる。

表 2 ：数値実験の結果

設定 1：楕円形データを生成。データ生成モデル：$\{\mu_x, \mu_y\}=\{5.5, 7.4\}$，$\Sigma=\{\{2.9,\cdot\},\{2.9\times0.8, 2.9\}\}$，$N=500$，$S=10000$，$c=10$。

モデル	データの利用状況	回帰係数(s.e.)	定数項(s.e.)	残差分散(s.e.)	欠損領域確率(s.e.)
データ生成モデル		0.8	3.0	1.044	0.0634
線形回帰	データ全使用	0.800 (0.0269)	2.999 (0.156)	1.012 (0.0625)	—
線形回帰	$c>y$の(x,y)使用	0.727 (0.0274)	3.313 (0.159)	0.925 (0.0590)	—
打ち切り回帰	$c<y$のx使用	0.800 (0.0276)	3.002 (0.158)	1.015 (0.0637)	—
切断回帰	$c>y$の(x,y)使用	0.787 (0.0329)	3.055 (0.176)	0.884 (0.0534)	—
正規密度推定 1	$c>y$の(x,y)使用, m推定	0.791 (0.0315)	3.045 (0.170)	0.960 (0.0642)	0.0597 (0.0116)
正規密度推定 2	$c>y$の(x,y)使用, m既知	0.794 (0.0277)	3.030 (0.159)	0.962 (0.0621)	0.0615 (0.00793)
一様・正規密度推定 1	$c>y$の(x,y)使用, m推定	0.897 (0.0294)	2.549 (0.161)	1.036 (0.0699)	0.203 (0.0291)
一様・正規密度推定 2	$c>y$の(x,y)使用, m既知	0.789 (0.0263)	3.031 (0.153)	0.949 (0.0611)	0.153 (0.0227)

設定 2：線形データを生成。データ生成モデル：$y=0.225x+12$，$x=U(10, 80)$，$N=500$，$S=10000$，$c=30$。

モデル	データの利用状況	回帰係数(s.e.)	定数項(s.e.)	残差分散(s.e.)	欠損領域確率(s.e.)
データ生成モデル		0.225	12.000	25.000	0.127
線形回帰	データ全使用	0.225 (0.0109)	12.010 (0.542)	24.878 (1.570)	—
線形回帰	$c>y$の(x,y)使用	0.176 (0.0106)	13.168 (0.536)	19.562 (1.320)	—
打ち切り回帰	$c<y$のx使用	0.225 (0.0112)	12.013 (0.548)	24.873 (1.711)	—
切断回帰	$c>y$の(x,y)使用	0.207 (0.0143)	12.371 (0.609)	18.076 (1.212)	—
正規密度推定 1	$c>y$の(x,y)使用, m推定	0.208 (0.0122)	12.513 (0.572)	20.483 (1.500)	0.102 (0.0163)
正規密度推定 2	$c>y$の(x,y)使用, m既知	0.214 (0.0107)	12.427 (0.560)	20.921 (1.449)	0.117 (0.0114)
一様・正規密度推定 1	$c>y$の(x,y)使用, m推定	0.225 (0.0142)	11.998 (0.601)	24.960 (2.113)	0.127 (0.0172)
一様・正規密度推定 2	$c>y$の(x,y)使用, m既知	0.225 (0.0118)	12.008 (0.579)	24.914 (1.849)	0.127 (0.0114)

一様・正規密度推定２＞一様・正規密度推定１＞打ち切り回帰＞正規密度推定２＞正規密度推定１＞切断回帰

(4)一様・正規密度推定モデル１，２では，欠測確率はほぼ真値を推定している。

4　おわりに

一般の回帰モデルとは違い，従属変数に強い分布の仮定を強いるモデルである。しかし，欠測データの個数の情報があれば，しかるべきモデルの仮定の下でうまく推定できることがわかった。

残されている課題として，他のモデルとの比較の意味で，提案モデルの対数尤度のバイアス，欠測しているデータ数の数え方と情報量規準でどのように扱うか，といったことが挙げられる。

謝　辞

本研究は札幌学院大学研究促進奨励金 SGU-S05-198007-02 の補助を受けています。

参考文献

[1] Amemiya T. (1985). *Advanced Econometrics*, Harvard University Press.
[2] Dempster A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, **39**, 1-38.
[3] Greene, William H. (2002). Econometric Analysis, 5th ed., Prentice Hall.
[4] Maddala, G. S.(1983). Limited-Dependent and Qualitative Variables in Econometrics, Cam-bridge University Press.
[5] 縄田和満（1997）.「Probit, Logit, Tobit」,「応用計量経済学Ⅱ」第４章，牧厚志，浪花貞夫，宮内環，縄田和満著，多賀出版.

（なかむら　ながとも　統計科学・計量分析学専攻）

（2008 年 10 月 21 日受理）