成蹊大学理工学研究報告
J. Fac. Sci. Tech., Seikei Univ.
Vol.43 No.1 (2006) pp.17-21

定量限界，検出限界を持つデータによる回帰パラメータの推定

阿部　貴行*[1]，岩崎　学*[2]

Estimation of Regression Parameters from Data with Detection and Quantitation Limits

Takayuki ABE *[1]　and　Manabu IWASAKI *[2]

ABSTRACT： A regression model with detection limit (DL) d and quantitation limit (QL) q is examined. Specifically, our goal is to estimate the regression parameters in a linear regression model $Y = \alpha + \beta x + \varepsilon$, under the condition that the actual values of Y will be observed only if Y is greater than q. For observations less than q, only the numbers of observations m_d and m_q fallen in the intervals $(0, d)$ or (d, q), respectively, will be reported. First, we show that a conventional method that gives the value q for the m_q observations in (d, q) and d for the m_d observations in $(0, d)$ produces inadequate estimates. This type of data is, in the context of incomplete data analysis, called NMAR (Not Missing At Random), and hence we have to take the missingness mechanism into account for the estimation. The maximum likelihood estimation procedure using the EM algorithm is given. This procedure is statistically sound but numerically intractable in practice. Therefore a simplified method is given, which is quite simple but can give adequate estimates. A numerical example is shown to illustrate the usefulness of the procedure proposed.

Keywords：censoring, chemical constituents, EM algorithm, maximum likelihood estimates

(Received February 25, 2006)

１．はじめに

環境問題および健康への関心が高まる中，工業製品に含まれ人体に影響を与える懸念のある化学物質の量を把握し，客観的な数値データとその影響の大きさを示すことが，特に近年，日本を含む世界各国の規制当局側から各種企業あるいは学術機関などに対して求められている。もちろん当局からの要請の有無にかかわらず，その種の分析は科学的にきわめて重要で，企業や各団体はそれを義務として捉える必要がある。しかし，製品中に含まれる化学物質の多くはきわめて微量であり，その計測・定量のためには特殊な計測装置および環境を必要とし，結果として多くのコストがかかってしまう。そこで，計測が比較的容易な別の物質の値 x から当該化学物質の値 y を回帰式 $y = a + bx$ などにより予測する試みがなされている。

ここにひとつ問題がある。それは，微量な化学物質の計測では，検出限界（detection limit = DL もしくは limit of detection = LOD）d および定量限界（quantitation limit = QL もしくは limit of quantitation = LOQ）q がある点である。検出限界と定量限界は次のように説明される（厚生労働省の分析法バリデーションに関するテキスト www.nihs.go.jp/g/ich/quality/q2a/q2a_jp. html 参照）：

(i) 検出限界 d:試料中に存在する分析対象物の検出可能な最低の量のことである。ただし，このとき必ずしも定量できる必要はない。

(ii) 定量限界 q：適切な精度と真度を伴って定量できる，試料中に存在する分析対象物の最低の量のことである。定量限界は，試料中に存在する低濃度の物質を定量する場合の分析能パラメータであり，特に，不純物や分解生成物の定量において評価される。

この種の限界がある場合，データは完全な計測値としては得られず，回帰パラメータの推定に工夫が必要となる。

本論では，次の第 2 節で典型的な実際例および説明のための人工例を示し，問題の所在を明らかにする。第 3

*[1]：万有製薬株式会社，情報処理専攻博士学生
*[2]：情報処理専攻教授 (iwasaki@st.seikei.ac.jp)

節では回帰モデルを定式化し，パラメータ推定のための妥当な方法論（EM アルゴリズム）を与える。この推定法は統計的には妥当であるものの，実際の計算は反復を必要とする複雑なものであり，多くの実際家にとってパラメータの推定値の導出はかなり難しい。そこで，第 4 節でそれに代わる簡便法を示し，その妥当性を吟味する。そして最後の第 5 節で簡単なまとめを行なう。

２．実例と典型的な数値例

Counts, et al. (2004) は，紙巻たばこ中の化学物質の予測モデルを扱っている。たばこの煙中には 4000 種類以上の化学物質が同定されていて，それらの物質の計量には専門のラボを必要とし多額の費用もかかる。そこで，計測が容易なたばこの煙中のタールやニコチン量を用いて種々の化学物質の量を予測しようという試みがなされている。Counts, et al. (2004) では主としてタールの量から各化学物質の量を予測する単回帰モデルを扱っている。論文中では 48 種類の紙巻たばこが選ばれ，種々の化学物質の計測値が示されている。表 1 は化学物質とその単位，および DL と QL の値とその個数の一部である。化学物質の中にはきわめて微量しか含有されていないため，定量限界，検出限界を超える測定値が得られないものもあった（Selenium，Nickel はその例）。

表 1.　DL と QL の例（48 種中）

Constituent	unit	DL	QL	#(DL)	#(QL)
Selenium	ng/cig	2.21	7.37	48	−
Nickel	ng/cig	6.47	21.6	46	2
Chromium	ng/cig	5.94	19.8	37	11
Arsenic	ng/cig	1.13	3.75	8	37
Lead	ng/cig	3.85	12.8	4	19
Resorcinol	μg/cig	0.158	0.526	3	11
Crotonaldehyde	μg/cig	0.98	3.29	1	12

以下では，人工的な数値例を基に具体的な議論を行なう。人工例ではあるが，Counts, et al. (2004) における鉛 (lead) に類したものである。図 1 は乱数により生成した全データ 50 個とそれにあてはめた回帰直線である。回帰直線の式は

$$y = 2.9174 + 1.5326\,x$$

であった（散布図の作成ならびに回帰直線の導出には MS EXCEL を用いた。表示は EXCEL のデフォルトの設定によったので必要以上に桁数が多い。以下同様）。それに対し，検出限界を $d = 3.8$，定量限界を $q = 12.8$ と設定し，観測値が d 以下のときは d と置き，d より大きく q 以下のときは q として散布図を作成し回帰直線をあてはめたのが図 2 である。このときの回帰直線は

$$y = 5.6485 + 1.2895\,x$$

であった。図 2 では全 50 個中 $d = 3.8$ となった観測値が 4 個，$q = 12.8$ となったものが 15 個あった。

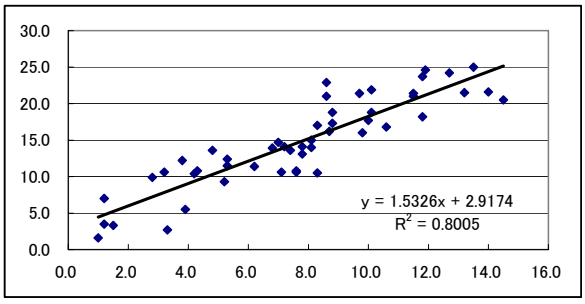

図 1.　人工的に生成した全データと回帰直線

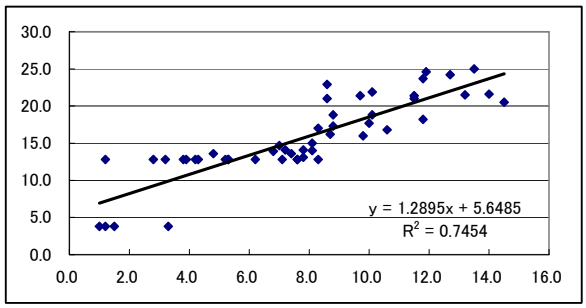

図 2.　検出限界，定量限界のあるデータ

図 2 は，検出限界，定量限界の値をそのままデータとして用い，それらが限界値であることを考慮しなかったときの計算で，全データでの回帰式と比較すると次の 2 つの特徴が読み取れる。

（ア）回帰直線の傾き b が 1.5326 から 1.2895 と小さくなっている。

（イ）決定係数 R^2 が 0.8005 から 0.7454 と減少している。

これらは両方とも，限界を考慮しないパラメータ推定では不適切な結果を生み出すことを意味する。そもそもの分析の目的は目的変数 y（この場合は鉛を想定）の予測であり，あまり大きくないとはいえ（イ）の予測精度の低下は好ましいものではない。問題は（ア）である。y は人体への有害物質であり，その過小評価は我々の健康に対し重大な影響を及ぼしかねない。回帰直線の傾きが小さいと，大きな x に対する y の値は本来の予測値より小さくなってしまい，悪影響が懸念される y の大きな部分での過小評価につながる。図 3 は図 1 の全データと図 2 の限界を無視した回帰分析での残差をプロットしたものである。説明変数 x の大きな部分で，限界を無視したときの残差 (R-CEN) の方が全データの残差 (R-ALL) よりも大きく，

$$y_i - \hat{y}_i^{(CEN)} > y_i - \hat{y}_i^{(ALL)}$$
$$\Downarrow$$
$$\hat{y}_i^{(ALL)} > \hat{y}_i^{(CEN)}$$

より，限界を無視すると過小評価につながることが実証されている。

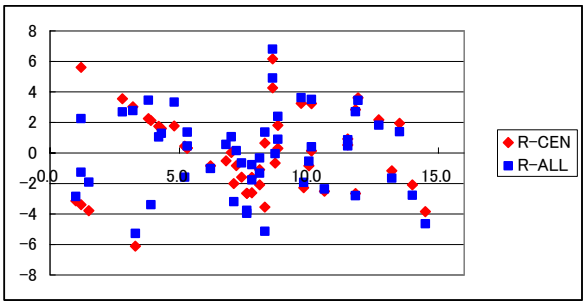

図 3. 全データと限界を無視した場合の残差

３．EM アルゴリズムによるパラメータの最尤推定

まず，回帰モデルおよび種々の仮定を述べておく。説明変数の値が x_i のときの目的変数を表わす確率変数を Y_i とし，それらの間に単回帰モデル

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \ldots, n), \qquad (3.1)$$

を想定する。ここで，α および β が推定対象となる未知パラメータであり，誤差項 ε_i には

$$\varepsilon_i \sim N(0, \sigma^2), \quad R[\varepsilon_i, \varepsilon_j] = 0 \quad (i \neq j) \qquad (3.2)$$

を仮定する。目的変数は化学物質の量であるので必然的に非負であり，厳密な意味では誤差項の正規性は仮定できないが，ここでは近似的なモデルとして正規分布を仮定し，モデルでの Y_i が負となる確率は無視できるほど小さいとする。目的変数 Y_i の実現値を y_i とし，y_i の計測には検出限界 d および定量限界 q が存在すると仮定する $(d < q)$。目的変数 y の値が定量限界 q 以下の場合には，その測定値そのものは分からないものの，それが d 以下であるかもしくは $d < y \le q$ であるかは分かるものとする。

ここで扱っている問題は，不完全データ解析の言葉では，目的変数 y の値に基づく打ち切りであり，欠測メカニズムは NMAR (not missing at random) で，無視できない (nonignorable) 欠測となる（Little and Rubin (2002) もしくは岩崎 (2002) を参照)。したがって，欠測メカニズムを考慮した推測を行なわなければならない。

ここでは，回帰パラメータ α と β の最尤推定値を EM アルゴリズム (Dempster, et al., 1977) により求める。そのため補題を用意する（Johnson and Kotz (1970) もしくは岩崎 (2002) 参照)。

補題 3.1 $Y \sim N(\mu, \sigma^2)$ のとき，

$$a^* = \frac{a-\mu}{\sigma}, \quad b^* = \frac{b-\mu}{\sigma}$$

と置くと，以下が成り立つ：

$$E[Y \mid a \le Y \le b] = \mu + \frac{\varphi(a^*) - \varphi(b^*)}{\Phi(b^*) - \Phi(a^*)}\sigma$$

$$V[Y \mid a \le Y \le b] = \sigma^2\left[1 + \frac{a^*\varphi(a^*) - b^*\varphi(b^*)}{\Phi(b^*) - \Phi(a^*)} - \left\{\frac{\varphi(a^*) - \varphi(b^*)}{\Phi(b^*) - \Phi(a^*)}\right\}^2\right],$$

ここで，$\varphi(y)$ および $\Phi(y)$ はそれぞれ標準正規分布 $N(0, 1)$ の確率密度関数および累積分布関数である。

この補題より次の定理が成り立つ。

定理 3.1 モデル (3.1) および (3.2) の下で，

$$E[Y \mid x; a \le Y \le b] = \alpha + \beta x + \frac{\varphi(a^*) - \varphi(b^*)}{\Phi(b^*) - \Phi(a^*)}\sigma$$

であり，$V[Y \mid x; a \le Y \le b]$ は補題 1 の $V[Y \mid a \le Y \le b]$ と同じとなる。

推定すべきパラメータ α, β, σ の対数尤度関数は

$$l = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - (\alpha + \beta x_i)\}^2$$

である。若干の計算の結果，同時十分統計量は

$$w_y = \sum_{i=1}^{n} y_i, \quad w_{yy} = \sum_{i=1}^{n} {y_i}^2, \quad w_{xy} = \sum_{i=1}^{n} x_i y_i$$

であることが分かる。n 組の全データが観測されたときの回帰パラメータの最尤推定値は

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{w_y}{n},$$

$$s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$s_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}(w_{yy} - n\bar{y}^2)$$

および

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(w_{xy} - n\bar{x}\bar{y})$$

として

$$b=\frac{s_{xy}}{s_{xx}},\quad a=\bar{y}-b\bar{x},\quad s^2=s_{yy}-\frac{{s_{xy}}^2}{s_{xx}} \qquad (3.3)$$

と求められる。

n 組の測定において，検出限界 d 以下の観測値が m_d 個，d より大きく定量限界 q 以下の観測値が m_q 個あり，q 以上で値が観測された観測値が $y_1, \ldots, y_m$ と m 個あるとき（すなわち $n = m_d + m_q + m$），最尤推定値の計算のための EM アルゴリズムは以下のようになる（M ステップの計算は容易であるが E ステップは，多くの EM アルゴリズムがそうであるように，計算が複雑となる）。

・パラメータ推定のための EM アルゴリズム

(1) 初期値 $a^{(0)}$，$b^{(0)}$，$s^{(0)}$ を適当に定める。

(2) E ステップ：第 t ステップ目でのパラメータ値が $a^{(t)}$，$b^{(t)}$，$s^{(t)}$ のとき，説明変数の値が x_i のデータに対し

$$d_i^{(t)}=\frac{d-(a^{(t)}+b^{(t)}x_i)}{s^{(t)}},\quad q_i^{(t)}=\frac{q-(a^{(t)}+b^{(t)}x_i)}{s^{(t)}}$$

とする。そして，補題 3.1 および定理 3.1 における a^* と b^* を，m_d 個の検出限界以下のデータについては

$$a^*=-\infty,\ b^*=d_i^{(t)}$$

とし，m_q 個の検出限界より大きく底流限界以下のデータについては

$$a^*=d_i^{(t)},\quad b^*=q_i^{(t)}$$

として y_i および y_i^2 の期待値を求め，それらにより十分統計量 w_y，w_{yy}，w_{xy} の期待値を計算する。

(3) M ステップ：E ステップで求めた十分統計量を用いて，完全データの公式 (3) から $a^{(t+1)}$, $b^{(t+1)}$, $s^{(t+1)}$ を求める。

(4) 収束判定により，終了するか (2) に戻る。

第 2 節の人工例（図 2 のデータ）に上記の EM アルゴリズムを適用した結果，推定された回帰直線は

$$y = 1.1511 + 1.7019\,x$$

であった。傾きの推定値 1.7019 は限界を考慮しない場合の値 1.2895 よりも大きく，傾きの過小評価が解消されている。

4. 簡便法とその性能

EM アルゴリズムによる最尤推定は，統計理論的には妥当性を持つものであるが，通常の（市販の）ソフトウェアでは対処できず，計算ははなはだ面倒である。そこでここでは，計算が容易でしかも妥当な結果を与える簡便法を考える。

まず，目的変数の値が得られた m 個のデータのみから求めた回帰直線は妥当でないことに注意する。欠測が，x には依存するが y には依存しないいわゆる MAR (Missing At Random) の場合には欠測データを無視した回帰分析は妥当な結果を与えるが（Little and Rubin (2002) もしくは岩崎 (2002) を参照），ここでの欠測は NMAR であるので，欠測は無視可能でない。

そこで，妥当な結果を与える EM アルゴリズムと不適切な結果となった検出限界および定量限界の値をそのまま用いた方法を今一度見直し，妥当と思われる方法を見出す。検出限界 d より大きく定量限界 q 以下となった m_q 個のデータに対応した目的変数 Y_i^* の条件付き確率密度関数を $f_i^*(x)$ とし，累積分布関数を $F_i^*(x)$ とすると $(i = 1, \ldots, m_q)$，Y_i^* の条件付き期待値は

$$E[Y_i^* \mid d < Y_i^* \le q] = \frac{1}{F_i^*(q)-F_i^*(d)}\int_d^q x f_i^*(x)\,dx \qquad (4.1)$$

となる。当然 $d < E[Y_i^* \mid d < Y_i^* \le q] \le q$ であり，この条件付き期待値を最大値の q とした図 2 に基づく解析は妥当ではない。また，EM アルゴリスムは (4.1) の条件付き期待値を逐次計算するものとなっている。十分統計量のひとつ w_y（第 3 節参照）における m_q 個の寄与分は

$$\sum_{i=1}^{m_d} E[Y_i^* \mid d < Y_i^* \le q] = \int_d^q x \sum_{i=1}^{m_d} \frac{f_i^*(x)}{F_i^*(q)-F_i^*(d)}\,dx \qquad (4.2)$$

である。そしてこの値は $m_d(d + q)/2$ に近いことが期待される。その理由は 2 つある。第 1 に，$f_i^*(x)$ が区間 (d, q) で左右対称に近ければその期待値は区間の中点 $(d + q)/2$ に近い。第 2 に，区間 (d, q) で歪度が正のものと負のものとが相殺し $\sum_{i=1}^{m_d} \frac{f_i^*(x)}{F_i^*(q)-F_i^*(d)}$ は全体として左右対称に近くなる。したがって，(4.2) 式の値が $m_d(d + q)/2$ となるよう各 y_i^* の値を定めればよいことになる。

求めようとする回帰直線は左下がりであり，x が小さくなるほど y も小さくなる傾向にある。そこで，そのように y の値を決めればよいと思われるが，それは，y の条件付き分散の過小評価につながり，得策とはいえない。実際，EM アルゴリズムでも，Y の期待値と同時に Y^2 の期待値の推定値を用いて条件付き分散の過小評価を回避している。これらを考え合わせ，最も簡単な方策として，「定量限界以下のデータの値をすべて $(d + q)/2$ とする」ことが考えられる。

次に，検出限界 d 以下の m_d 個のデータについて考える。第 3 節のモデルの定式化の際に述べたが，正規分布は負の値も取り得るが y は非負であり，厳密には矛盾

が生じている。そこで，定量限界以下のデータのアナロジーとして，便宜的に「検出限界以下のデータの値をすべて $d/2$ とする」とする（この部分の理論的な正当化は困難である）。

上記の簡便法「検出限界 d 以下の y の値は $d/2$ とし，定量限界 q 以下で検出限界より大きな y の値は $(d+q)/2$ とそれぞれ置き直して回帰分析にかける」を適用する。第 2 節の人工例に簡便法を適用した結果を図 4 に示す。

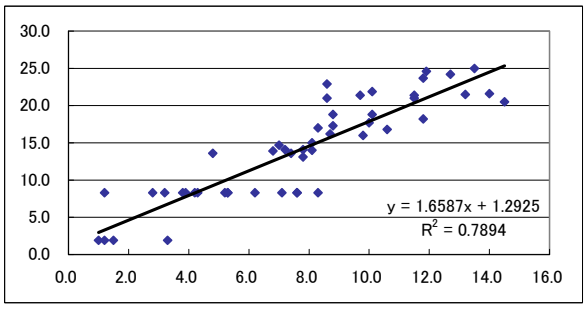

図 4. 簡便法による回帰分析

求められた回帰直線は

$$y = 1.2925 + 1.6587\,x$$

であり，第3節のEMアルゴリズムによる推定値に近く，限界を無視した場合の傾きの過小評価も解消されている。この簡便法は，まさに簡便でありしかも上記の例では best possible と思われる EM アルゴリズムによる最尤推定値に近い推定値を与えている。

５．おわりに

不完全データに対する EM アルゴリズムによる最尤推定や多重代入法による対処などは，近年の統計理論の整備ならびにソフトウェアの開発により実用に供されつつある（Little and Rubin (2002), 岩崎 (2002), Rubin (1987) などを参照）。しかし，本研究で取り上げたような NMAR となる欠測（打ち切り）では，通常のソフトウェアでは対処できず，そのためのプログラムを個別に書かなければならない。それは，科学的な研究には不可欠ではあるが，現実問題では実行が困難である。そこで，それらは理想的な対処法であることは知った上で，なるべくそれに近い結果を与えるより簡便な方法があればそれを採用するのが現実的な選択となるであろう。本研究の視点のひとつはそこにあった。

実は厄介な問題がある。検出限界や定量限界が存在する問題では，y と x の値のいずれかもしくは両方がきわめて 0 に近いことが予想され，生の観測値に対して (3.1) の直線回帰式および (3.2) の仮定がそのまま成立しない恐れがある。特に y_i の分散は，x_i の値が 0 に近いときは小さく，x_i が大きくなるに連れて大きくなることが多い。

このときの対処法として y に適当な変数変換（たとえば対数変換 $\log y$）を施し，モデル (3.1) および仮定 (3.2) の両方が成立するようにする。そして第 4 節で述べた簡便法を適用する。この方法も，定量限界値および検出限界値そのものを用いるのに比べより妥当な結果を与えると推察されるが，本論ではこれに言及する余裕がない。場を改めて議論したい。

本論の第 3 節で用いた回帰パラメータの最尤推定値導出のための EM アルゴリズムは SAS を用いて書いた。興味をもたれた読者にはそれを提供する用意があるので，著者まで連絡されたい。

謝　辞

情報処理専攻修士学生の中戸川智彦君には本論文での計算をお手伝いいただいた。また，本研究には科学研究費基盤研究 (A) No. 16200022 の援助を受けた。ここに感謝する。

参考文献

Counts, M. E., Hsu, F. S., Laffoon, S. W., Dwyer, R. W. and Cox, R. H. (2004) Mainstream smoke constituent yields and predicting relationships from a worldwide market sample of cigarette brands: ISO smoking conditions. *Regulatory Toxicology and Pharmac- ology*, **39**, 111-134.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38 (with discussion).

岩崎 学 (2002) 不完全データの統計解析，エコノミスト社.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data, Second Edition*. John Wiley & Sons, New York.

Johnson, N. L. and Kotz, S. (1970) *Distributions in Statistics. Continuous Univariate Distributions* – 1. John Wiley & Sons, New York.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.