外れ値を考慮に入れた線形回帰モデルにおける予測について

知識情報処理研究　　3602B049-3　仲川文隆
指　　導　松嶋敏泰教授

A Note on Prediction from Linear Regression Model with Outliers

by Fumitaka Nakagawa

1　はじめに

回帰分析の応用領域は, 自然科学, 工学, 人文社会科学に至るまで, 非常に広範囲に及んでいる. 特に経営工学分野においては, 統計的品質管理や需要予測などの分野で, 制御, 予測等に有効な手法として, 広く活用されている.

また, 一般に統計的な解析をする場合, データに外れ値が含まれることは多々ある. その際の, 外れ値の扱い方については従来から広く研究されている. 外れ値の扱い方は, 主に外れ値の発生にモデルを置く方法と置かない方法に分けられる [2]. 本研究では前者の外れ値にモデルを置く場合について扱う. 従来, 外れ値の発生を考慮に入れた線形回帰モデルが考案されている [1][6]. [1] では, このモデルに対し, 回帰パラメータの事後確率を求める方法が提案されている.

他方, 線形回帰モデルに対し, ベイズ基準のもとで最適な予測を行う研究がされている [4]. ベイズ基準のもとで最適な予測とは有限個のデータのもとで, 予測による損失を平均的に最小にする意味で最適性が保証される.

本研究ではまず, 外れ値を考慮に入れた線形回帰モデルに対し, ベイズ基準のもとで最適な予測法を示す. しかし, これを求めるには, データ数の指数オーダーの計算量がかかってしまい, 実用上困難である. そこで計算量を削減するアルゴリズムを提案する. 最後に, シミュレーションによって提案法の有効性について考察を行う.

2　モデル

2.1　外れ値を考慮に入れた線形回帰モデル

説明変数が p 個, データが n 組ある場合を考える. 説明変数のベクトル列 $X = (x_1, \ldots, x_n)^t$ $(x_i = (x_{i1}, \ldots, x_{ip}) \in \mathcal{R}^p, (i = 1, \ldots, n))$ ($\mathcal{R}^p$ は p 次元ユークリッド空間) に対する目的変数のベクトル列を $y = (y_1, \ldots, y_n)^t \in \mathcal{R}$,$\theta$ を p 次元パラメータとする. この条件のもとで,y_i の従う確率モデルは,

$$y_i = x_i^t \theta + \epsilon_i, \tag{1}$$

とする. ここで $\epsilon^n = (\epsilon_1, \ldots, \epsilon_n)$ は誤差を表し, 誤差が従う確率モデルは,

$$(1-\alpha)N(0, \sigma^2) + \alpha N(0, k^2\sigma^2), \tag{2}$$

であるとする. ここで,$N(a, b)$ は平均 a, 分散 b の正規分布を表し,$N(0, \sigma^2)$ は通常のデータが従う分布,$N(0, k^2\sigma^2)$ を外れ値のデータが従う分布を表している. また,α は外れ値が発生する確率を表している. ここで,k は 1 よりも十分大きい値であり, 外れ値は通常のデータより大きい分散を持つ分布により発生することを仮定している.

2.2　パラメータの事後分布を求める計算 [1]

パラメータの事後分布は,

$$p(\theta|X, y) = \frac{p(\theta)p(y|X, \theta)}{\int p(\theta)p(y|X, \theta)d\theta}, \tag{3}$$

によって求められる. しかし, このモデルにおいて (3) 式の分母はパラメータ空間上での積分が含まれるため直接解析的に求めることができない.

そこで, パラメータの事後分布を計算するにあたって,z_i として i 番目のデータが外れ値であるとき 1, それ以外のとき 0 をとる隠れ変数を導入する. このとき z_i の系列 $z^n = (z_1, \ldots, z_n)$ によって,n 個のデータが外れ値か外れ値ではないかを表すことができる. この z^n の値によって ϵ を外れ値にあたる部分 $\epsilon_{(r)}$ と通常のデータの部分 $\epsilon_{(s)}$ に,y

を $y_{(r)}$ と $y_{(s)}$,X を $X_{(r)}$ と $X_{(s)}$ に分けることができる. これを用いてパラメータの事後分布は,

$$p(\theta|X,y)=\sum_{Z^n} p(z^n|X,y)p(\theta|z^n,X,y), \quad (4)$$

によって求めることができる. ここでパラメータの事前分布は無情報事前分布で, $p(\theta,\sigma)\propto\frac{1}{\sigma}$ とすると, $p(\theta|z^n,X,y)$ は,

$$p(\theta|z^n,X,y) \propto \left\{1+\frac{(\theta-\hat{\theta})^t(X^tX-\phi X_{(r)}^tX_{(r)})(\theta-\hat{\theta})}{S_{(r)}(\hat{\theta})}\right\}^{-\frac{n}{2}}, \quad (5)$$

となり, 多変量 t 分布に従うことがわかる. ここで $\phi=1-k^{-2}$. 今 $\hat{\theta}$ は z^n が与えられたもとでの最小二乗推定量であり,

$$\hat{\theta}=\left(X_{(s)}^tX_{(s)}+\frac{1}{k^2}X_{(r)}^tX_{(r)}\right)^{-1} \left(X_{(s)}^ty_{(s)}+\frac{1}{k^2}X_{(r)}^ty_{(r)}\right), \quad (6)$$

$S_r(\hat{\theta})$ は,

$$S_r(\hat{\theta})=\left(y_{(s)}-X_{(s)}^t\hat{\theta}\right)^t\left(y_{(s)}-X_{(s)}^t\hat{\theta}\right) +\frac{1}{k^2}\left(y_{(r)}-X_{(r)}^t\hat{\theta}\right)^t\left(y_{(r)}-X_{(r)}^t\hat{\theta}\right), \quad (7)$$

である. また,

$$p(z^n|X,y)\propto\alpha^r(1-\alpha)^{(n-r)}k^{-r} |X^tX-\phi X_{(r)}^tX_{(r)}|^{-\frac{1}{2}}\{S_{(r)}(\hat{\theta})\}^{-\frac{(n-p)}{2}}, \quad (8)$$

となる.(5),(8) よりパラメータの事後分布は多変量 t 分布の 2^n 個の混合として求めることができる.

3 予測問題

3.1 ベイズ最適な予測法

本研究で扱う予測問題は入力と出力の n 個の組 (X,y) に基づいて,x_{n+1} が与えられたときの y_{n+1} の予測値 $\hat{y}_{n+1}$ を得ることとする.

予測値 $\hat{y}_{n+1}$ に対して損失関数を $Loss$ で表すと以下のようになる.

$$Loss=(y_{n+1}-\hat{y}_{n+1})^2. \quad (9)$$

また, 損失関数に対し,y の条件付分布 $p(y|X,\theta)$ と X の条件付分布 $p(X|\nu)$ に関して平均をとった期待損失を危険関数 $Risk$ とよび, 以下で定義する.

$$Risk=\iint Loss\times p(y|X,\theta)p(X|\nu)dy\,dX. \quad (10)$$

本来であれば, この危険関数を最小にする $\hat{y}$ を求めたいのだが, パラメータによって危険関数を最小にする決定関数は異なるため, 任意の θ すべてについてを最小にする $\hat{y}$ は存在しない [5]. そこで危険関数を事前分布で期待値をとったベイズリスクを以下で定義する.

$$BR=\int_\theta Risk\times p(\theta)d\theta. \quad (11)$$

ベイズ最適な予測とは (11) を最小にする予測である. このとき, ベイズ最適な予測は,

$$\hat{y}_{n+1}^*=\int y_{n+1}\int_\theta p(y_{n+1}|x_{n+1},\theta) p(\theta|X,y)d\theta dy_{n+1}. \quad (12)$$

(12) を用いて計算することによって得られる. ここで,

$$p(y_{n+1}|x_{n+1},X,y) =\int p(y_{n+1}|x_{n+1},\theta)p(\theta|X,y)d\theta \quad (13)$$

を予測分布とよぶ.

外れ値を考慮に入れた線形回帰モデルにおいて (12) 式はパラメータ空間上での積分が直接的には解析的に求めることができない. しかし, [1] のパラメータの事後分布を求める方法と同様, 外れ値の出方のパターンの重み付けと考えることで解析的に計算することができる. 今,$w_{(r)}=p(z^n|X,y)$

とおくと, 予測分布は以下のようになる.

$$p(y_{n+1}|x_{n+1},X,y) = \sum_{Z^n}\iint w_{(r)}(\sigma)^{-(n+2)}\alpha \exp\left\{-\frac{S_{(r)}(\theta)+(y_{n+1}-x_{n+1}^t\theta)^2)}{2\sigma^2}\right\}d\theta\,d\sigma + \sum_{Z^n}\iint w_{(r)}(\sigma)^{-(n+2)}(1-\alpha) \exp\left\{-\frac{S_{(r)}(\theta)+\frac{(y_{n+1}-x_{n+1}^t\theta)^2}{k^2}}{2\sigma^2}\right\}d\theta\,d\sigma. \quad (14)$$

(14) の各項の積分は解析的に計算することができ,

$$\iint \sigma^{-(n+2)} \exp\left\{-\frac{S_{(r)}(\theta)+(y_{n+1}-x_{n+1}^t\theta)^2)}{2\sigma^2}\right\}d\theta\,d\sigma \propto \left\{S_{(r)}(\hat{\theta})+y_{n+1}^2\left(1-x_{(n+1)}B^{-1}x_{n+1}^t\right)\right\}^{-\frac{(n-p)}{2}}, \quad (15)$$

$$B = \left(X^tX-\phi X_{(r)}^tX_{(r)}\right)+x_{n+1}^tx_{n+1}, \quad (16)$$

であり, これは t 分布となる. よって予測分布は外れ値の全パターンで重み付けた t 分布の混合として表すことができる.

以上から, 二乗誤差損失を最小にする, すなわちベイズ最適な予測を行うには,t 分布の期待値の事後確率による重み付け和を用いればよいことがわかる. しかし, 実際には 2^n 回の和計算が必要となり, データ数が増えると, この計算は計算量的に困難になる.

3.2 計算量削減法

本節では, 事後分布を近似することで, 和の計算量を減らすことを考える. 和の計算量を減らす方法として,$p(z^n|X,y)$ をすべて計算し, 大きいものから順にいくつかをとって重み付けるという方法が考えられる. しかし,$p(z^n|X,y)$ 全通りの計算は 2^n 通りあるので計算量的に困難である. そこで,$p(\theta|X,y)$ の極大値 $\tilde{\theta}$ を用いて,$p(z^n|X,y,\tilde{\theta})$ を計算し, その値を用いて重み付ける順番を近似的に求めることを考える.

アルゴリズム

step1: $p(\theta|X,y)$ の極大値 $\tilde{\theta}$ を求める. EM アルゴリズム [7] を用いて, 収束先を極大値とする.

step2:

$$\hat{z}_i = \begin{cases} 1 & \text{if } p(z_i|x_i,y_i,\tilde{\theta}) > 0.5 \\ 0 & \text{else} \end{cases} \quad (17)$$

$$\hat{z}_0^n = (\hat{z}_i,\ldots,\hat{z}_n) \quad (18)$$

step3: 確信度 $r_i = |0.5 - p(z_i|x_i,y_i)|$ の計算
確信度の小さいものから順に $\hat{z}_i$ を a 個選び, 反転させた系列 $\hat{z}_1^n,\ldots,\hat{z}_{2^a}^n$ を 2^a 個つくる.

step4:$p(\hat{z}_i^n)$ の計算

$$p(\hat{z}_i^n) = \frac{p(\hat{z}_i^n|X,y)}{\sum_{i=0}^a p(\hat{z}_i^n|X,y)} \quad (19)$$

step5: 予測分布の計算

$$p(y_{n+1}|x_{n+1},X,y) \approx \sum_{i=0}^{2^a} p(\hat{z}_i^n)p(y_{n+1}|x_{n+1},X,y,\hat{z}_i^n) \quad (20)$$

4 シミュレーション

4.1 シミュレーションの目的

提案法の有効性をシミュレーションによって調べる. 評価基準としては, 二乗損失での予測精度を取り上げる. テスト用の未学習データに対する提案アルゴリズムと最適な計算の場合の予測値と真値の二乗誤差の平均で評価を行うこととする. そして, 計算量について考察する. また, 和の計算量と二乗誤差とのトレードオフをみる.

4.2 シミュレーション条件

まず, 説明変数 x_i を二次元とし, それぞれ独立に正規分布 $N(0,1)$ に従って 15 個作成する.

次に,θ を, それぞれ一様に作成し, テスト用のデータを 20 個作成する. テスト用のデータに対する予測値と真値の二乗誤差の 20 回の平均を計る.
混合比 $\alpha = 0.2$,$k^2 = 5$
z の分布 $p(z^n|X, y)$ の大きさによって大きい順に全パターンを重み付けて二乗誤差を比べる. 提案法では $a = 1, \ldots, 13$ までの 2^{13} 通りの重み付けまで考える.

4.3 シミュレーション結果

シミュレーション結果を図 1 に示す.

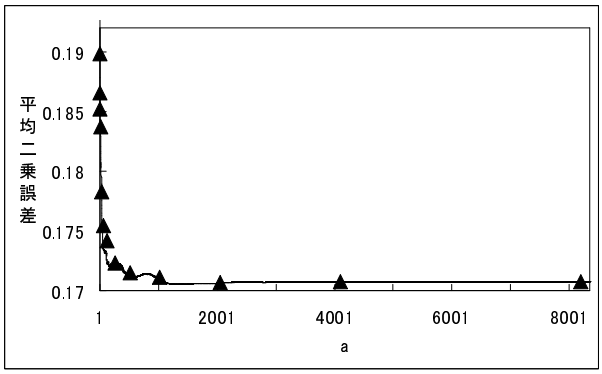

図 1: 提案法と従来法　二乗誤差平均

実線部が $p(z^n|X, y)$ の大きさによって大きい順に並べて重み付けたもの. 三角の点が提案アルゴリズムによって計算したものを表している.

5 考察

提案アルゴリズムは計算量的に計算が困難な $p(z^n|X, y)$ を $p(z|X, y, \tilde{\theta})$ で近似し, 確率の大きな外れ値の出方のパターンのみを考えることで和の計算を減らしている. 実際にアルゴリズムの **step4** で計算している $p(\hat{z}_i^n)$ の値は近似値ではなく, 正確な確率に比例した値となるため, 近似となっているのは, 和の計算で重み付けなかった部分だけとなっている.

計算量については, データが外れ値かを表すパターンすべてで和をとるための計算量は $O(2^n)$ となり, データ数が多い時にはこの部分の計算量が主要項となる. しかし, 提案アルゴリズムにおいては,$\hat{\theta}$ を求めるのにかかる繰り返し回数が平均 2〜3 回程度であるため, 毎回のステップにかかる計算量がデータ数 n の線形オーダーとなる, 大幅に計算量を減らすことができていることがわかる. グラフより,2^{11} 程度重み付けるとそれ以降に関してはほぼ二乗誤差には変化は無くなる. よって近似の精度の面からみても 2^{11} 程度まで重み付ければ十分であることがわかる.

$p(z^n|X, y)$ を $p(z|X, y, \tilde{\theta})$ によって近似しその値の大きい順に重み付けているが, 実験よりおもに最初の方の重み付けでよい値をとることができている. このことから, パラメータの推定値 $\tilde{\theta}$ がよい推定量であることが予想される.

6 まとめ

本研究では, 線形回帰モデルで, 外れ値を考慮にいれたモデルにおいてベイズ基準のもとで最適な予測を示した. また, データ数の指数オーダーの計算量の削減アルゴリズムを作成した.

今後の課題としては,k の値,α の値の変化と二乗誤差の関係を見ていく. また,α の値が未知の場合の予測, パラメータの事前分布に共役な事前分布を置いたときの予測について考えていく.

参考文献

[1] G.E.P.Box , G.C.Tiao : *A Baysian approach to some outlier problems*, Biometrika.1968,pp.119-129

[2] Barnett Vic(1994) : *Outliers in statistical data*,chichester,Wiley&Sons.

[3] 松嶋 敏泰,"帰納・演繹推論と予測 -決定理論による学習モデル-" , 第 1 回 情報論的学習理論ワークショップ予稿集, 1998,pp.1-8.

[4] J.M.Bernardo , A.F.M.Smith : (1994) *Bayesian theory*,wiley.

[5] James.O.Berger(1985):*Statistical Desicion Theory and Bayesian Analysis*, Springer.

[6] D. Pena And G.C.Tiao:*Bayesian Robustness Functions for Linear Models*, Bayesian Statistics 4,1992,pp.365-388.

[7] Dempster,A.P.,Laird,N.M.,Rubin,D.B. : *Maximum likelihood from incomplete data via the EM algorithm*, J.O.R.S.S B, 1977, pp.1-38.