線形回帰の幾何学的特徴に関する研究

2002MM014 服部 慎吾
指導教員　木村 美善

1　はじめに

本研究の目的は, Brian J. McCartin の 2 つの論文 [3][4] の内容を理解すると共にその幾何学的な特徴について, 統計ソフト R による自作プログラムを用いてグラフから探っていくことである. また, 上記の論文で触れられていない特徴についても考察する.

2　線形単回帰

観測データ, $\{(x_i, y_i)\}$ $(i = 1, 2, \cdots n)$ について考える. ただし $x_i = X_i + u_i,\ y_i = Y_i + v_i$ である. ここで, (X_i, Y_i) は正確な値を表し, (u_i, v_i) は対応する確率誤差である. 次の仮定をおく. (i) $E(u_i) = E(v_i) = 0$, (ii) 誤差 u_i と v_iは無相関 , (iii) $u_1, \cdots, u_n$ および $v_1, \cdots, v_n$ は互いに独立. これらの仮定のもとで, 線形単回帰モデル $aX + bY = c$ を考える. ここで a, b は $a^2 + b^2 = 1$ を満たすか, または他の基準化がなされているとする.

3　最小 2 乗法 (OLS) と加重最小 2 乗法 (WLS)

線形単回帰モデル $aX + bY = c$ の a, b, c を最小 2 乗法によって推定するとは

$$\sum_{i=1}^{n}(ax_i + by_i - c)^2 \tag{1}$$

を最小にする a, b, c を選ぶということである. しかし, 一般的なデータでは分散の不均一性の問題が発生することがある. その問題を回避する方法として加重最小 2 乗法がある. 加重最小 2 乗法とは, 最小 2 乗法に重みをつけて

$$\sum_{i=1}^{n}\frac{k}{Var(ax_i + by_i - c)}(ax_i + by_i - c)^2 \tag{2}$$

を最小にする a, b, c を選ぶ方法である.

4　λ 回帰法

まず, $\lambda = \frac{Var(v)}{Var(u)} = \frac{\sigma_v^2}{\sigma_u^2}$ とする. また, この節以降の仮定として誤差は等分散ではないものとする. このとき, WLS における最小化問題は,

$$\min_{a,b,c}\frac{k}{a^2\sigma_u^2 + b^2\sigma_v^2}\sum_{i=1}^{n}(ax_i + by_i - c)^2$$

となる. ここで, k は a, b, c に関係しないので, $k = \max(\sigma_u^2, \sigma_v^2)$ とすれば最小化問題は λ を用いて, $\min_{a,b,c}\frac{\max(1,\lambda)}{a^2+\lambda b^2}\sum_{i=1}^{n}(ax_i + by_i - c)^2$ とすることができる. つぎに, データの平均を $\bar{x}, \bar{y}$ とすると, a, b がどのような値を取ったとしても, $c = a\bar{x} + b\bar{y}$ と表すことで式 (1) は常に最小になる. このことから, 回帰直線は $L : a(x - \bar{x}) + b(y - \bar{y}) = 0$ となり, データの重心 $(\bar{x}, \bar{y})$ を通ることが保証される. ここで, 回帰直線 L の傾き $m = -\frac{a}{b}$ を代入すると, 最小化問題は

$$\min_m \frac{\max(1,\lambda)}{\lambda + m^2}\sum_{i=1}^{n}[(y_i - \bar{y}) - m\cdot(x_i - \bar{x})]^2 \tag{3}$$

となる ([1][5] 参照). この λ 回帰法において, $\lambda = 0$ のとき, x の y への回帰となり, $\lambda = 1$ のときは, 直交回帰そして, $\lambda \to \infty$ のときには, y の x への回帰となる. このように, λ の値を変化させることで, 通常の回帰や主成分分析に対応する直交回帰が得られることがわかる.

λ 回帰直線の傾き $m(\lambda)$ は, 直交回帰の場合を解くことで得ることができる. 傾き $m(\lambda)$ は, x の標本分散 σ_x^2, y の標本分散 σ_y^2, x と y の標本共分散 p_{xy} を用いると,

$$m(\lambda) = \frac{(\sigma_y^2 - \lambda\sigma_x^2) + \sqrt{(\sigma_y^2 - \lambda\sigma_x^2)^2 + 4\lambda p_{xy}^2}}{2p_{xy}} \tag{4}$$

となる. ここで, $p_{xy}\cdot m(\lambda)$ の 1 次の導関数を求めることにより, $m(\lambda)$ の値は単調変化することがわかる.

5　(λ, μ) 回帰法

λ 回帰法は誤差の分散の不均一があると仮定した上での理論であったが, (λ, μ) 回帰法はその上に誤差間に相関があることを仮定したものである. ゆえに, $Cov(u, v) = p_{uv} \neq 0$ を仮定する. このとき,WLS の最小化問題は $\min_{a,b,c}\frac{k}{a^2\sigma_u^2 + 2ab p_{uv} + b^2\sigma_v^2}\sum_{i=1}^{n}[ax_i + by_i - c]^2$ となる. k は線形パラメータ a, b, c に関係がないので, 任意に定めてもよい. ここで, $k = \max(\sigma_u^2, \sigma_v^2)$ とし, $\mu = p_{uv}/\sigma_u^2$ とすると最小化問題は, 回帰直線の傾き $m = -a/b$ を用いて,

$$\min_m \frac{\max(1,\lambda)}{\lambda - 2\mu m + m^2}\sum_{i=1}^{n}[(y_i - \bar{y}) - m\cdot(x_i - \bar{x})]^2 \tag{5}$$

となる ([1][5] 参照). λ と μ の取りうる値の範囲は,$-\infty < \mu < +\infty$ かつ $\mu^2 < \lambda < \infty$ である. 最小値を与える $m(\lambda, \mu)$ もまた, 直交回帰法から導出することができ,

$$m(\lambda, \mu) = \frac{\sigma_y^2 - \lambda\sigma_x^2}{2(p_{xy} - \mu\sigma_x^2)} + \frac{\sqrt{(\sigma_y^2 - \lambda\sigma_x^2)^2 + 4(p_{xy} - \mu\sigma_x^2)(\lambda p_{xy} - \mu\sigma_y^2)}}{2(p_{xy} - \mu\sigma_x^2)} \tag{6}$$

となる. この $m(\lambda, \mu)$ は $m(\lambda)$ と異なる変化をする. $(p_{xy} - \mu\sigma_x^2)\cdot m(\lambda, \mu)$ において, λ についての 1 次の偏微分を行

うことで, μ を固定し λ を変化させるとき, $m(\lambda,\mu)$ は $(p_{xy}-\mu\sigma_x^2)>0$ のとき単調減少し, $(p_{xy}-\mu\sigma_x^2)<0$ のとき単調増加することがわかる.

6 幾何学的性質

式 (6) を変形すると, $(\bar{x},\bar{y})$ を中心とした集積楕円の式になり, λ と μ を変化させることで回帰直線の傾き $m(\lambda)$ と $m(\lambda,\mu)$ はその値を変化させる. ゆえに, 回帰直線は λ と μ を変化させると集積楕円の中心を軸に回転する. $-\infty<\mu<+\infty$ かつ $\mu^2<\lambda<\infty$ であるので, μ の値を μ_0 と固定し, λ の取りうる範囲 $\mu_0^2<\lambda<\infty$ における (λ,μ) 回帰直線の変化を幾何学的に見る. また, 図 1 は $p_{xy}>0$ のときのものであるが, どんな値の場合も楕円は相似になる. 次の 4 つの範囲に分けて考える. I: $\mu<s_\perp$, II: $s_\perp<\mu<0$, III: $0<\mu<m_y$, IV: $m_y<m$. この, 4 つの範囲は図 1 に対応するものである.

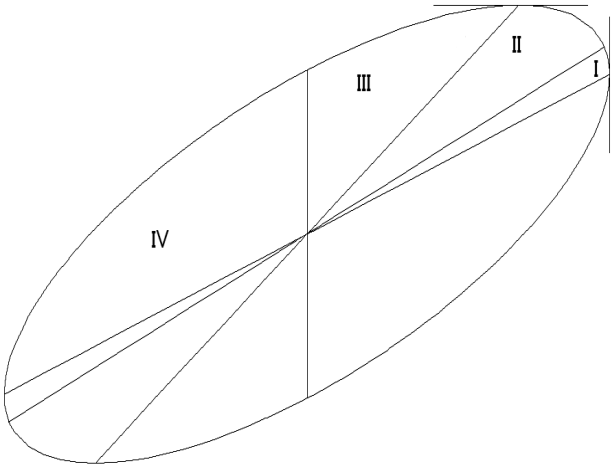

図 1 集積楕円:$p_{xy}>0$ のとき

ここで, $m(\lambda,\mu_0^2)$ の λ ついての極限を求めると, $\lambda\to\infty$ のとき, $m\to m_y^{\pm}$, $s\to\mp\infty$ となり, $\lambda\to\mu^2$ のとき, $m\to\frac{\sigma_y^2-\mu p_{xy}}{p_{xy}-\mu\sigma_x^2}$, $s\to\mu$ となる. ゆえに, λ が μ^2 から ∞ に変化するに当たって, 以下の特徴があることがわかる.

- $p_{xy}-\mu\sigma_x^2>0$ の場合, $\lambda\to\mu^2$ のとき回帰直線は I,II,III の範囲にある. λ が増加するにつれて時計回りに回転し, $\lambda\to\infty$ のとき, y の x への回帰へと近づく.
- $p_{xy}-\mu\sigma_x^2<0$ の場合, $\lambda\to\mu^2$ のとき回帰直線は IV の範囲にある. λ が増加するにつれて反時計回りに回転し, $\lambda\to\infty$ のとき, y の x への回帰へと近づく.
- $p_{xy}-\mu\sigma_x^2$ が ±0 に近ければ近いほど, $\lambda\to\mu^2$ のとき回帰直線は垂直に近くなる.

この結果から, (λ,μ) 回帰直線は λ と μ の値を変化させることで, どんな傾きも取りうることがわかる.

7 シミュレーション

R による自作プログラムを用いて, λ 回帰と (λ,μ) 回帰の変化の様子を視覚的に見れるようにした (プログラムは [2] を参考に作成した). [3] と [4] の理論から得られた特徴に基づいてシミュレーションを行った結果を以下に示す.

1. r_{xy} と σ_x^2/σ_y^2 の値が変化しなければ, 取りうる回帰直線の範囲は変わらない.
2. $\lambda=\mu^2$ に近いとき, λ の小さな変化に対して回帰直線は大きな変化をするが, λ の値が大きくなるにつれて, 回帰直線の変化は小さくなる.

性質 1 は, 楕円が相似に変化するので, 回帰直線の変化する範囲が変わらないからである. 性質 2 は, 回帰直線が楕円の接線の変化に対応する共役直径から得られる点から, λ が小さければ小さいほど接線が急な変化をし, λ が大きければ大きいほど接線は緩やかな変化をすると考えられる. また, 楕円の性質から長軸と短軸の差が大きくなると, この傾向は強くなることがわかる.

8 おわりに

本研究で導出した λ 回帰法と (λ,μ) 回帰法の回帰直線は λ と μ を変化させることでデータの平均を中心とした楕円の中心を軸に回転することがわかる. R を用いることで直線の動く範囲と幾何学的な特徴や変化の様子について知ることができた. しかし, 当初目標としていた最適な λ,μ の値は, データ間の誤差の分散の不均一性と相関性が仮定されて求められた回帰直線であるので, 回帰の良さを測る指標がない. そのため, 最適な変数の値を求めることができなかったのが残念である. しかし, プログラムで線形回帰の幾何学的特長から λ 回帰と (λ,μ) 回帰の性質についての理解を深めることができたことは満足している.

謝辞

本論文を書くにあたり, 御意見, ご指導を下さった木村先生, 安藤先生, 並びにゼミの友人に深く感謝いたします.

参考文献

[1] Cramer, H.: Mathematical Methods of Statistics, Princeton University Press, (1974).
[2] 船尾 暢男 著: The R Tips データ解析環境 R の基本技 · グラフィックス活用集, 九天社 (2005).
[3] McCartin, B.J.: A geometric characterization of linear regression, Statistics Vol.37, (2003).
[4] McCartin, B.J. : The geometry of linear regression with correlated errors, Statistics Vol.39, (2005).
[5] Salmon, G.: A Treatise on Conic Sections, sixth edition Amer Mathematical Society, (1954).