回帰分析とその応用
— 愛・地球博データを中心として —

2002MM099　田中　千尋
指導教員: 木村　美善

1　はじめに

回帰分析では最小2乗法による推定が通常よく用いられる.しかし,最小2乗法は1つの外れ値によって大きな影響を受けてしまう. 本研究では,愛・地球博入場者数データを用いて回帰診断と,代表的なロバスト法の一つであるLMS推定量による外れ値の考察を行う.

2　回帰分析

2.1　線形回帰モデル

応答変数yとk個の説明変数$(x_1, x_2, \cdots, x_k)$に関するn個の観測値が与えられているとき,次の線形重回帰モデルを考える.

$$y_i = \theta_0 + \theta_1 x_{1i} + \cdots + \theta_k x_{ki} + \epsilon_i \qquad (i = 1, 2, \cdots, n)$$

ここでϵ_iは誤差項でy_iの変動のうち$(x_{1i}, x_{2i}, \cdots, x_{ki})$では説明しきれない部分を表している. ([1],[3],[6]参照)

2.2　誤差

誤差ϵ_iに対して次の仮定をおく.

1. $E(\epsilon_i) = 0 \qquad (i = 1, 2, \cdots, n)$ [不偏性]
2. $V(\epsilon_i) = \sigma^2 \qquad (i = 1, 2, \cdots, n)$ [等分散性]
3. $Cov(\epsilon_i, \epsilon_j) = 0 \quad i \neq j$ [無相関性]
4. ϵ_iは互いに独立に正規分布$N(0, \sigma^2)$に従う. [正規性]

2.3　最小2乗法

線形回帰モデルにおいて,実測値y_iと予測値$\hat{y_i}$の差$\hat{\epsilon_i}(\theta) = y_i - \hat{y_i}$を残差といい,最小2乗法とは残差平方和が最小になるように, θの推定値を定める方法であり,得られる$\hat{\theta}$を最小2乗(LS)推定量という:

$$\sum_{i=1}^{n} \hat{\epsilon_i}^2(\hat{\theta}) = \min_{\theta} \sum_{i=1}^{n} \hat{\epsilon}_i^2(\theta)$$

2.4　LMS推定量

LMS推定量$(\hat{\theta}_{LMS})$とは,残差の2乗の中央値を最小にする推定量である([5]参照):

$$med_i \hat{\epsilon_i}(\hat{\theta}_{LMS}) = \min_{\theta} med_i \hat{\epsilon}_i^2(\theta)$$

3　回帰診断

3.1　回帰診断とは

回帰診断とは,回帰モデル式の妥当性や,誤差に関する仮定の正当性をチェックしたり,外れ値の影響を調べたりする方法である. ([2],[7]参照)

3.2　残差

独立・同一正規誤差という標準的仮定の是非を確認するために残差プロットと正規Q-Qプロットがよく用いられる. 正規Q-Qプロットでは点がほぼ直線上に並べば与えられたデータが正規分布に近い分布をしていることになる. W統計量（$Shapiro - Wilk$検定）は残差の正規性を検定するために用いられる:

$$W = \frac{1}{s^2} \left\{ \sum_{i=1}^{[\frac{n}{2}]} a_{n,m} (x_{(m)} - x_{(i)}) \right\}^2, m = n - i + 1$$

3.3　ハット行列

最小2乗推定量$\hat{\theta} = X(X'X)^{-1}X'y$を用いると,予測値ベクトル$\hat{y} = X\hat{\theta}$は,以下のように表せる.

$$\begin{aligned} \hat{y} &= X\hat{\theta} = X(X'X)^{-1}X'y \\ &= Hy \end{aligned}$$

ハット行列$H = X(X'X)^{-1}X'$はyから$\hat{y}$への変換をするので,ハット行列又は射影行列と呼ばれる.ハット行列Hは対称かつベキ等である.

3.4　梃子比

行列Hの第(α, α)要素を$h_{\alpha\alpha}$とおくと,個々の残差の推定値の分散は$Var(\hat{\epsilon}) = \sigma^2(1 - h_{\alpha\alpha})$となる. $h_{\alpha\alpha}$を梃子比と呼ぶ.梃子比$h_{\alpha\alpha}$については

$$\begin{cases} 0 \leq h_{\alpha\alpha} \leq 1 \\ \sum h_{\alpha\alpha} = p' = p + 1 \end{cases}$$

が成り立ち,これより$h_{\alpha\alpha}$の平均的な値はp'/nである.

3.5　スチューデント化残差

$Var(\hat{\epsilon}) = \sigma^2(1 - h_{\alpha\alpha})$を使って,平均0,分散1を持つ標準化残差$\hat{\epsilon_i}/\sigma\sqrt{1 - h_{\alpha\alpha}}$が得られる. そして$\sigma$と$s$ $(= \sqrt{(y'(I - H)y)/(n - p - 1)})$を取り替えることで得られた

$$r_i = \frac{\hat{\epsilon_i}}{s\sqrt{1 - h_{\alpha\alpha}}}$$

を標準化残差(内的スチューデント残差)と呼び,sの代わりに$s_{(i)}$を用いたものをスチューデント化残差(外的スチューデント化残差)と呼ぶ. $s_{(i)}$とは(y_i, x_i')を省いた後,残ったn-1個の観測値で計算したものである.

3.6　Cookの距離

1つのデータが推定された回帰モデルパラメータに大きく影響を与えることがある. i番目の観測値(y_i, x_i')を削除して回帰分析を行うことによって得られた推定値$\hat{\theta}_{(i)}$と全データを用いた推定値$\hat{\theta}$との違いがi番目のデータの回帰推定値への影響の大きさを示す. Cookの距離によって,$\hat{\theta}_{(i)}$と$\hat{\theta}$を比較することができる.

4　データによる分析

4.1　データについて

2005年2月から2006年9月まで行われた愛地球博入場者数のデータを扱う.このデータは185日間,18変数のデータである.　入場者数を曜日,連休,最高気温,最低気温,降水量,日照時間,天気,台風,イベント有無を用いて分析した.([4],[8],[9]参照)

4.2　分析結果

AICによる変数選択を行い,土曜日,連休,最高気温,最低気温,日照時間,雨を説明変数とした. 最小2乗法により回帰式, $\hat{y} = 9.91991 + 0.14258x_8 + 0.06687x_{10} + 0.32565x_{11} + 0.20333x_{12} + 0.01275x_{14} - 0.15858x_{17}$ が得られる.

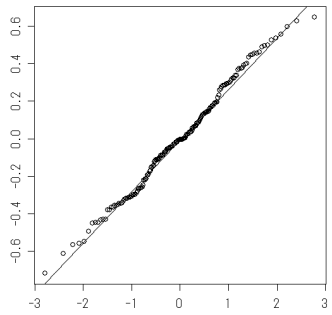

図 1: LS正規Q-Qプロット

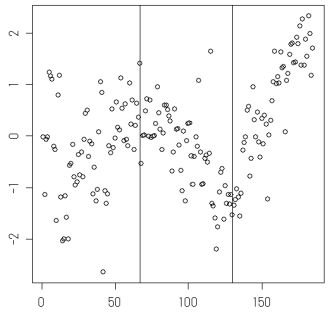

図 2: LS残差プロット

Q-QプロットとW検定のp値が0.329から,正規性は否定されない.決定係数は0.4543,自由度調整済み決定係数は0.4359となる.　LMS推定量を用いて分析を行い,LSの残差プロットと比較する.図2,3の横軸は,開幕からの日数である.

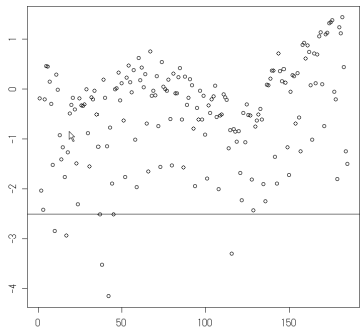

図 3: LMS残差プロット

LMSの残差プロットから観測値10,17,37,38,42,45,116の7個の外れ値が検出された. LMS推定量の外れ値を除いた決定係数は0.4982,自由度調整済み決定係数は0.4806となり先ほどの値より良い.

次に,回帰診断から外れ値を検出する方法を用いた.回帰係数への影響を与えるCook距離が0.5を越える目立った値は観測値は検出されなかった.　しかし,スチューデント化残差より観測値14,15,18,42,119,176,178,182,183の9個が外れ値と考えられる. 梃子比を調べた結果,観測値1,3,12は危険とみなされたが,その他の分析から外れ値とは考えにくい. これらの観測値9個を除いた決定係数は0.494,自由度調整済み決定係数は0.476となり,回帰診断を繰り返す. 最終的に外れ値として検出された観測値は16個となり決定係数は0.5194,自由度調整済み決定係数は0.5016となった. 外れ値と見なされた観測値を詳しく調べてみる. 観測値42は5月5日の連休で晴れにも関わらず入場者数が伸びなかったために外れ値となり,4月5,7,8日は共通し開幕後の伸び悩みが原因と考えられる. そして,中盤は7月17日は連休の急増,7月21,22,26日は急激な減少から外れ値になってしまった. 9月10,11,14,15,16,18,20,22,23日は,終盤の混雑がまとまって外れ値になったと思われる.

LS残差プロットから3つの時期で異なった傾向がみられる.そこで,開幕から67日目と130日目で区切り,ダミー変数を追加して分析を行った. その結果,各残差プロットはランダムに分布し,残差にみられたパターンは解消された.

5　おわりに

回帰診断やロバスト回帰を用いて外れ値を検出する際に,外れ値を安易に取り除くことは危険である. そしてロバスト回帰だけでは不十分であり,回帰診断だけでも判断は不十分である.本研究では回帰診断とロバスト回帰の比較を行ったが, 個々の分析ではなく総合的な研究も大切だ. さらに外れ値とみなされた観測値の詳細を調査することも重要である.

参考文献

[1] 圓川隆夫: 多変量のデータ解析,朝倉書店(2006)

[2] Chatterjee,S and Price,B: Regression Analysis Example,Wiley,New York(1977) (佐和隆光・加納悟訳: 回帰分析の実際,新曜社,1980)

[3] 神谷美紀・水戸藍・竹内愛希代: ロバスト回帰の研究,南山大学数理情報学部数理科学科卒業論文(2006)

[4] 間瀬茂・神保雅一・鎌倉稔成・金藤浩司: 工学のためのデータサイエンス入門,数理工学社(2004)

[5] Rousseeuw,P.J.and Leroy A.M.: Robust Regression and Outlier Detection,Wiley,New York(1986)

[6] 佐和隆光: 回帰分析,朝倉書店(2002)

[7] 田中豊・脇本和昌: 多変量統計解析法,現代数学社(1998)

[8] 気象庁: http://www.jma.go.jp/jma/index.html

[9] 財団法人2005年日本国際博覧会協会: http：//www.expo2005.or.jp/