電力需要分析のための新しいデータ解析手法

キーワード：需要分析，回帰診断，探索的データ解析，リサンプリング手法

小野賢治　大屋隆生

〔要　旨〕

本報告は回帰診断，探索的データ解析，リサンプリング手法といった新しいデータ解析手法を電力需要の分析に適用して得られた成果をまとめたものである。

回帰診断は電力需要モデルにおいて，異常データや推定量の質を落としている原因を見いだすのに有用である。また，探索的データ解析は，電力需要分布の特徴の比較において新しい視点を与える。さらに，リサンプリング手法によって電力需要モデルの妥当性の評価や，徒来の方法では推定できなかった統計量の推定が可能となった。

1. はじめに

電力需要の変動の主な要因は気象，人口，景気等である。しかしこの他にも数多くの要因があり，そのすべてをとらえることはできない。それゆえ，電力需要の分析には確率モデルをもとにしたデータ解析の手法を用いることが必要となる。

現在，電力需要の分析には数多くの手法が用いられているが，それらの大部分は，単に，理論分布である正規分布を仮定して平均や分散を求める，回帰直線をあてはめるある，分布を仮定してパラメータを推定する，というように古典的な統計理論に基づいた分析法である。これらの手法は，分析の対象となるデータが仮定した分布に厳密にしたがっているときは正しい推定結果が得られる。また，推定のための計算は比較的容易である。しかし，データが仮定した分布にしたがっていないときは，分析結果が正しいという保証はない。

当然のことながら，現実のデータである理論分布にしたがうものは皆無といってよい。現実

のデータには，分布形にゆがみやひずみがあったり，他の多くのデータから離れた値（外れ値あるいは異常値）があったりする。したがって，データの背後にある現象を的確にとらえ，適切な行動に結びつく分析を行うためには，古典的な分布理論あるいは漸近理論に基づくデータ解析だけでは不充分であり，以下のことが必要となる。

i) データのようすを細かくみる。

ii) データが仮定した分布からはずれているために，モデルの推定結果がどのような影響を受けるかを評価する。

iii) 異常値の影響を受けにくい分析法，あるいは特定の分布にとらわれない分析法を用いる。

近年の電算機技術の急速な発展にともなって，上の i)，ii)，iii) の観点によるデータ解析の手法がしだいに広まりつつある。

本報告はこのような，いわば「新しいデータ解析手法」のいくつかを紹介し，あわせて電力需要分析に適用して得られた成果について述べたものである。

2. 回帰診断による電力需要モデルの検討

2.1 背　景

電力需要の分析においては，電力需要の変動を気象，人口，景気等の変量で説明するモデルを仮定し，そのパラメータを推定するために回帰分析が用いられることが多い。回帰分析は電力需要分析に限らず，一般のデータ解析において最も広範に用いられる分析手法の一つである。

通常の線形回帰モデルにおける仮定は，説明変数を与えられたスコアとみなし，被説明変数の値は，説明変数の線形結合に，データごとに独立で同一分布の誤差項が加わったとするものである。しかし1章でも述べたように現実のデータはこのようなモデルに厳密にしたがっているわけではない。またそれとは別に，説明変数の構造に問題がある場合もある。

電力需要分析における回帰分析の適用に際しての主な問題点は以下の3つである。

1) データの中に，予期せぬ原因あるいは原因不明の異常値（あるいは外れ値）があることが多い。異常値が存在すると，回帰係数の推定値はその異常値にひきずられてしまう。

2) 電力需要の変動を説明するための変量には似たような動きをするものが多く，多重共線性（説明変数の一部または全部に線形関係がある状態）に近い状態になることがある。この場合には回帰係数の推定値が計算できないか，できたとしても非常に不正確になる。

3) 電力需要のデータはそのほとんどが時系列データであり，誤差項が時間の経過にともなってある一定のパターンで変動し（系列相関），独立の仮定が成り立たない場合が多い。このような場合には，誤差分散の推定量にバイアス（偏り）が生ずることが知られている。

データにこのような問題点があることを見落したまま分析を行うと，誤った結論を導くことになりかねない。本章では，特に上の（1）と（2）について，現在行っている分析がそのような原因によってどの程度の影響を受けているかを「診断」する方法，すなわち「回帰診断の方法とその対策について述べ，あわせてその方法を電力需要と気象との関係に適用した結果を

示す。尚，本章における手法は〔1〕によるものである。また，分析計算は独自に開発したりフトウェアを用いて行った(注)。

2.2 有影響データの診断

ここで用いる有影響データ診断の方法は，1つのサンプルを除いたときの残差や回帰係数などの変化の大きさを調べるものである。表 2.1 に回帰モデルの一般式を示し，表 2.2 に有影響データ診断の諸統計量の定義式を示す。

次に，実際の電力需要の モデル分析 に 対 し て，この有影響データ診断を行った結果を示す。対象としたモデルは，ある電力会社の夏季（7月・8月）の平日における，1日の発受電端電力量（日量）と1日の平均気温との関係をあらわす以下のモデルである[2]。

$$y_i=a+bt_i+\varepsilon_i \qquad (2.1)$$

ただし，y_i：日量

t_i：平均気温

a, b：回帰係数

ε_i：誤差項

表 2.3 は，上のモデルについて，表 2.2 で示した有影響データ診断の諸統計量を計算した結果である。

有影響データに対する一般的な対策としては，以下のことが挙げられる。

1) もとのデータに記入ミス等がないかどうか調べる。
2) 多重共線性のチェックを行う。
3) 異常値とみなしてとり除く。
4) モデルが適切であるかどうか検討する。

ここでは，この有影響データ診断を行うことによって，以下の点でモデル分析に役立った。

i) この年の8月2日のデータが異常値であると判断され，サンプルから除外することにより，r^2 の値が 0.1 近く上昇した。

ii) (2.1) 式のモデル（1次式モデル）が不適切であると判断され，折線モデルあるいは2次式モデル[2]のあてはめを検討するに

表 2.1 回帰モデルの一般式

モデル $y=X\beta+\varepsilon$		推定式 $y=Xb+e$	
y	：被説明変数（$n\times1$）	y	：同 左
X	：説明変数 （$n\times P$）	X	：同 左
β	：回帰係数 （$P\times1$）	b	：βの最小2乗推定量
ε	：誤 差 （$n\times1$）	e	：残 差
σ^2	：誤差分散	s^2	：誤差分散の推定量
x_i	：Xの第 i 行	$b(i)$	：第 i サンプルを除いたときの β の最小2乗推定量
X_j	：Xの第 j 行	$s^2(i)$	：第 i サンプルを除いたときの誤差分散の推定量
$X(i)$	：第 i 行を除いたX行列		

(注) ここで述べる回帰診断の諸統計量の実際の計算は，筆者が開発した回帰分析システム AREAS (Advanced REgression Analysis System) を用いた。このシステムは，当所において経済分析に広く用いられている TSP (Time Sevies Processor) の専用データバンクからデータを読み込み，OLS 推定および回帰診断を行うシステムである。

尚，回帰診断の諸統計量の計算は，汎用統計パッケージ SAS で行うことができる。

表 2.2 有影響データ診断の諸統計量

(1) RSTUDENT

$$\mathrm{RSTUDENT}_i \equiv \frac{e_i}{s(i)\sqrt{1-h_i}} \qquad \text{ただし，} h_i = \mathrm{HATDIAG}_i$$

(2) HATDIAG

$$\mathrm{HATDIAG}_i \equiv h_i \equiv \boldsymbol{x}_i (X'X)^{-1} \boldsymbol{x}_i'$$

(3) COVRATIO

$$\mathrm{COVRATIO}_i \equiv \frac{\det\{s^2(i)[X'(i)\,X(i)]^{-1}\}}{\det[s^2(X'X)^{-1}]} = \frac{1}{\left[\frac{n-P-1}{n-P} + \frac{e_i^{*2}}{n-P}\right]^P (1-h_i)}$$

ただし，$e_i^* = \mathrm{RSTUDENT}_i$

det は行列式をあらわす。

(4) DFFITS

$$\mathrm{DFFITS}_i \equiv \frac{1}{S(i)\sqrt{hi}}[\hat{y}_i - \hat{y}_i(i)]$$

(5) DFBETAS

$$\mathrm{DFBETAS}_{ij} \equiv \frac{b_j - b_j(i)}{S(i)\sqrt{(X'X)^{-1}_{jj}}}$$

ただし，$(X'X)^{-1}_{jj}$ は行列 $(X'X)^{-1}$ の (j, j) 成分

至った。

このように有影響データ診断の統計量を検討することにより，問題となるデータを発見し，その原因を検討して正確な推定に役立てることができる。

2.3 多重共線性の診断

回帰モデルの説明変数群の中に多重共線性が存在するという状態は，ここでは，その変数群のうちの1つの変数を残りの変数に回帰させたときの重相関係数が高くなる状態であるとする。経済時系列データの場合は，いくつかの変量が同じようなパターンで変動することが多いので，多重共線性が起こりやすい。

2.3.1 多重共線性のもたらす悪影響

1) 数値計算上の影響

説明変数間に多重共線性が存在すると，説明変数における誤差が回帰係数の推定量に拡大されて伝わる。

2) 統計上の影響

データに多重共線性が存在すると回帰係数の推定量の分散が大きくなり，推定，検定，予測等の精度が落ちる。

表 2.3　有影響データ診断の諸統計量（A社，昭和57年7月・8月のデータ）

日付	RSTUDENT	HATDIAG	COVRATIO	DFFITS	DFBETAS	
					定数項	平均気温
7/ 1	0.591	0.139 ◆	1.198 ◆	0.237	0.226	−0.217
2	0.665	0.156 ◆	1.217 ◆	0.286	0.274	−0.264
5	−0.399	0.035	1.079	−0.076	−0.051	0.045
6	0.176	0.037	1.088	0.034	0.024	−0.021
7	0.764	0.082	1.112	0.229	0.206	−0.195
8	0.363	0.041	1.087	0.075	0.056	−0.050
9	0.031	0.033	1.085	0.006	0.004	−0.003
12	0.085	0.045	1.098	0.018	−0.012	0.013
13	0.389	0.024	1.067	0.060	−0.005	0.011
14	0.410	0.024	1.066	0.064	0.020	−0.014
15	0.775	0.069	1.094	0.210	0.183	−0.172
16	0.414	0.080	1.131	0.122	0.109	−0.103
19	0.024	0.031	1.083	0.004	0.003	−0.002
20	0.497	0.025	1.063	0.079	0.030	−0.023
21	0.348	0.041	1.088	0.072	0.053	−0.048
22	0.291	0.026	1.073	0.048	0.021	−0.017
23	0.226	0.025	1.074	0.036	0.015	−0.012
26	−0.392	0.061	1.109	−0.100	−0.085	0.079
27	−0.126	0.054	1.109	−0.030	−0.025	0.023
28	−0.514	0.029	1.067	−0.089	−0.050	0.042
29	−0.490	0.023	1.061	−0.075	−0.012	0.004
30	−0.642	0.023	1.053	−0.100	−0.027	0.017
8/ 2	−2.440 ◆	0.047	0.839 ◆	−0.540 ◆	0.347	−0.387 ◆
3	−1.276	0.023	0.994	−0.196	0.001	−0.020
4	−0.920	0.041	1.051	−0.191	0.114	−0.128
5	−0.513	0.034	1.073	−0.096	0.048	−0.056
6	−0.323	0.037	1.084	−0.063	0.034	−0.039
9	−0.144	0.031	1.081	−0.026	0.011	−0.013
10	0.362	0.057	1.106	0.089	−0.063	0.069
11	0.507	0.045	1.085	0.110	−0.069	0.077
12	0.038	0.023	1.074	0.006	0.001	−0.000
13	−1.852	0.024	0.916	−0.292	0.046	−0.075
16	−4.496 ◆	0.023	0.482 ◆	−0.687 ◆	−0.107	0.039
17	−1.534	0.024	0.962	−0.242	0.038	−0.062
18	−0.662	0.040	1.070	−0.135	0.078	−0.088
19	0.294	0.041	1.090	0.061	−0.036	0.041
20	0.896	0.053	1.066	0.211	−0.145	0.159
23	0.690	0.066	1.098	0.184	−0.138	0.149
24	1.569	0.064	0.998	0.410	−0.303 ◆	0.329 ◆
25	1.520	0.064	1.005	0.397	−0.294	0.319 ◆
26	1.483	0.034	0.979	0.278	−0.137	0.160
27	1.205	0.053	1.033	0.284	−0.194	0.214
30	0.325	0.038	1.086	0.065	−0.036	0.041
31	1.073	0.037	1.031	0.210	−0.113	0.129

（注）◆印は influential なデータであることを示す。

2.3.2 多重共線性診断の指標

ここで用いる多重共線性診断の指標は，説明変数行列の特異値分解をもとに計算されるものである。以下で各指標について説明する。

1) 条件指標（Condition Index）η_k

$$\eta_k \equiv \frac{\mu_{\max}}{\mu_k} \quad k=1,\cdots,p \qquad (2.2)$$

ただし，

$\mu_{\max}$：説明変数行列の最大特異値(注)

μ_k：説明変数行列の k 番目に大きい特異値

p：説明変数の数

（注） 説明変数の行列を $X(n\times p)$ としたとき，Xの特異値分解は，

$$X=UDV' \qquad (2.3)$$

で表わされる。ここで，Uは $n\times p$，D と V は $p\times p$ の行列で，$U'U=V'V=I_p$（単位行列であり，Dは対角成分が非負の対角行列である。Dの対角成分を行列Xの特異値と呼ぶ。

2) 回帰係数の分散の分解 π_{ij}

これは，各回帰係数の推定量の分散を，説明変数行列の各特異値に対応した成分に分解したものである(注)。

（注） (2.3) 式を用いると，回帰係数の推定量の分散は次のように表わせる（表2.1の記号を用いる)。

$$\mathrm{cov}(\boldsymbol{b})=\sigma^2(X'X)^{-1}=\sigma^2VD^{-2}V' \qquad (2.4)$$

特に，第 k 回帰係数の推定量の分散は，

$$\mathrm{var}(b_k)=\sigma^2\sum_{j=1}^{p}\frac{v_{kj}^2}{\mu_j^2} \quad k=1,\cdots,p \qquad (2.5)$$

となる。ただし v_{kj} は行列 V の (k,j) 成分である。(2.5) 式で表わされる分散の分解を，次のように，全体の分散を1とした割合で表わす。

$$\phi_{kj}\equiv\frac{v_{kj}^2}{\mu_j^2}$$

$$\phi_k\equiv\sum_{j=1}^{p}\phi_{kj}$$

$$\pi_{jk}\equiv\frac{\phi_{kj}}{\phi_k} \qquad (2.6)$$

図 2.1 は，これらの指標を表形式に並べたものである。

回帰係数の分散の分解

条件指標	var(b_1)	var(b_2)	·····	var(b_P)
η_1	π_{11}	π_{12}	·····	π_{1P}
η_2	π_{21}	π_{22}	·····	π_{2P}
⋮	⋮	⋮		⋮
η_P	π_{P1}	π_{P2}	·····	π_{PP}

図 2.1 多重共線性診断の指標

2.3.3 診断の手順

多重共線性の診断は，図 2.1 の形の表を用いて以下の手順で行う。

〔ステップ1〕 条件指標の値を大きいと判定する基準値を η^* 設定する。

（例） $\eta^*=10$ あるいは 15 あるいは 30

〔ステップ2〕 各条件指標の値を η^* と比較する。η^* を超える条件指標が存在する場合には，次の3つのケースのうちのいずれであるかをみる。

（ケース1） η^* を超える条件指標はただ1つしかない。

（ケース2） η^* を超える条件指標が複数個存在し，かつそれらの条件指標の大きさのオーダがほぼ等しい（競合的関係)。

（ケース3） η^* を超える条件指標が複数個存在し，かつ大きさのオーダの異なる条件指標が混ざり合っている（支配的関係)。

（例） $\eta_1=1$，$\eta_2=3$，$\eta_3=30$，$\eta_4=300$

〔ステップ3〕 大きいと判定された各条件指

標に対応する線形にほぼ近い関係について，その関係にかかわっている変量を，回帰係数の分散の分解から識別する。その際に各分散の構成比 π_{ij} が大きい（すなわち j 番目の変量が π^* に対応する線形にほぼ近い関係にかかわっている）と判定する基準 π^* を設定する（$\pi^*=0.5$ が実際上よい基準である）。上の3つのケースで判定方法が異なる。

（ケース1の場合） η^* を超える条件指標を η_i とすると $\pi_{i1}, \pi_{i2}, \cdots, \pi_{ip}$ のうち，π^* を超えるものが2つ以上ある場合には，それに対応する変量は線形にほぼ近い関係にかかわっており，対応する回帰係数の推定量の質が落ちていると考えられる（π^* を超えるものがただ1つの場合にはどの回帰係数の質も落ちていない）。

（ケース2〔競合的関係〕の場合） η^* を超える条件指標を $\eta_1, \eta_2, \cdots, \eta_q$ とすると，$\sum_{i=1}^{q}\pi_{i1}, \sum_{i=1}^{q}\pi_{i2}, \cdots, \sum_{i=1}^{q}\pi_{in}$ のうち π^* を超えるものに対応する回帰係数の推定量の質が落ちていると考えられる。ただし，どの変量がどの線形関係にかかわっているかを識別することはできない。

（ケース3〔支配的関係〕の場合） この場合には，オーダの小さい方の条件指標に対応する線形関係にかかわっている変量を識別することはできない。したがって，オーダの小さい方の条件指標についてのπの値のうち π^* を超えるものがただ1つであっても，それに対応する回帰係数の推定量の質は落ちている可能性がある。この場合には，補助回帰（ある説明変数をその他の説明変数群で回帰させる）などさらに詳しい分析をする必要がある。

〔ステップ4〕 各々の線形（にほぼ近い）関係について，それにかかわっている変量を用いて補助回帰を行い，その関係を調べる。

〔ステップ5〕 多重共線性の影響を受けていない変量を識別する。

2.3.4 多重共線性への対策

今までに述べた方法によって多重共線性が回帰係数の推定量の質を落としていると診断されたときの対策としては以下のような方法が挙げられる。

1) 新しいデータ（サンプル）を追加する〔実際には不可能な場合が多い〕。

2) ベイズ流の推定法を用いる。

a) 純粋なベイズ推定

b) 混合推定

c) リッヂ回帰

2.3.5 電力需要データへの適用例

ここでは前節と同じく，ある電力会社の夏季における日量と気象変量との関係の分析において，多重共線性の診断を行った結果を示す。この分析で用いた回帰モデルは以下の式で表わされる。

$$y_i=a+b_1t_i+b_2(t_i-t^*)\mathrm{IND}_i+h_i+\varepsilon_i \quad (2.7)$$

ただし，y_i：1日の発受電端電力量

t_i：1日の平均気温

$$\mathrm{IND}_i=\begin{cases}0 & t_i\leqq t^* \text{ のとき}\\ 1 & t_i>t^* \text{ のとき}\end{cases}$$

t^*：回帰係数の変化点

h_i：1日の平均湿度

ε_i：誤差項

このモデルは(2.1)式のモデルに加えて，高温区間では需要量の気温感応度（気温1℃の上昇に対する需要量の増加）が高くなること，および湿度の影響を考慮したモデルである。

図2.2は，ある電力会社の昭和58年のデータに(2.7)式のモデルをあてはめたときの，多

〔特 異 値〕	〔条件指標〕	回帰係数の分散の分解			
		〔定 数 項〕	〔平均気温〕	〔高温区間ダミー〕	〔平均湿度〕
1 1.9063	1.0000	0.0002	0.0003	0.0037	0.0003
2 0.60061	3.1740	0.0006	0.0001	0.1402	0.0019
3 0.58960D−01	32.332	0.0124	0.5068	0.7059	0.6272
4 0.41233D−01	46.233	0.9867	0.4929	0.1502	0.3705

(注) 〔高温区間ダミー〕は，(2.7) 式における $(t_1-t^*)\mathrm{IND}_i$ のことである。

図 2.2 多重共線性診断指標の計算結果

重共線性の診断結果である。条件指標の基準値を $\eta^*=30$ に設定すると，この値を超える条件指標が2つ存在する。この2つの値のオーダはほぼ等しいので，2.2.3 の〔ステップ2〕におけるケース2（競合的関係）に相当する。各々の回帰係数の分散の分解のうち大きな条件指標に対応する2つの値の和をそれぞれについて計算すると，定数項，平均気温，平均湿度については 0.99 を超えており，高温区間のみのダミー変数についても 0.85 とかなり高い値になっている。すなわち，この場合はすべての回帰係数の推定量の質が落ちているといえる。

3. 探索的データ解析による電力需要分布の分析

3.1 背 景

電力需要分析あるいは負荷研究においては，電気の使用量や負荷の大きさの需要家間の散らばりのようすをみるためにヒストグラムを描き，さらに分布の位置，広がり等を知るために標本平均や標本分散などの要約統計量を計算する。しかしこのような方法による分布の検討には以下の問題点がある。

1) 数多くの分布を一度に比較する場合，ヒストグラムでは分布の特徴のとらえ方が明確でなく，比較がしにくい。

2) サンプル数が少ない場合，ヒストグラムはもとの分布の形をあらわすことができなくなってしまう。

3) 異常値がある場合，標本平均はその異常値にひきずられてしまい，分布の位置を正確には反映しなくなる。また，ヒストグラムでは異常値の個数や位置をつかみにくい。

本章では，上の問題点を解消するための新しい分布形の表示法として，探索的データ解析の一手法である箱ヒゲ図 (boxplot)[4] を用いて電力需要分布の比較検討を行った結果と，さらにその分析に続くベキ変換について述べる。

3.2 箱ヒゲ図について

箱ヒゲ図とは，データの分布のようすを略図的に示すものである。図 3.1 に箱ヒゲ図の作図方法を示す。箱ヒゲ図をみることによって，データの分布に関する以下の特徴を読みとることができる。

① 位置 (location)

② 散らばり (spread)

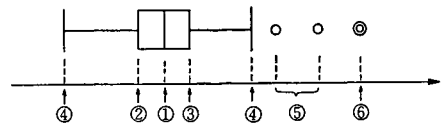
③ ゆがみ (skewness)

④ 尾の長さ (tail length)

⑤ 異常値 (outlying data points)

また，箱ヒゲ図によるデータの表示には以下の利点がある。

1) 箱ヒゲ図を並べることによって，多くのデータの分布のようすを視覚的に比較する



① 中央値（M）
② 下方の4分位値（F_L）［データを昇順に並べたとき，大きさの順位が全体の$\frac{1}{4}$である点］
③ 上方の4分位値（F_U）［データを昇順に並べたとき，大きさの順位が全体の$\frac{3}{4}$である点］
④ 異常値でない最も端のデータ点
⑤ 異常値（区間〔$F_L-\frac{3}{2}d_F$, $F_U+\frac{3}{2}d_F$〕に入らず，〔F_L-3d_F, F_U+3d_F〕に入るデータ点，ただし$d_F=F_U-F_L$）
⑥ 特に離れた異常値（区間〔F_L-3d_F, F_U+3d_F〕に入らないデータ点）

図 3.1　箱ヒゲ図の作図法

ことができる。

2）箱ヒゲ図で表示される統計量は央央値など順序に基づくものであり，異常値の影響を受けにくい（resistant である）。

3.3　箱ヒゲ図による表示

ここでは，ある電力会社の昭和 54 年の電灯需要データを箱ヒゲ図を用いて分析した。分析方法としては，以下の2つを行った。

1）ある契約アンペアについて，12ヶ月の分布を比較する。

2）ある月について，契約アンペアごとの分布を比較する。

図 3.2，3.3 は上の 1），2）のそれぞれの表示のうちの一つを示したものである。この分析の結果，需要量のレベルと散らばり方の季節変動，契約アンペアによる分布のひずみ方のちがい等で従来までの分析では得られないような興味深い情報が得られた（例：各々の分布について，大きい側の異常値は数多く存在するが，箱をみると上方へのひずみはあまりみられない）。

3.4　ベキ変換

図3.3のように分布のレベル（中央値）が大

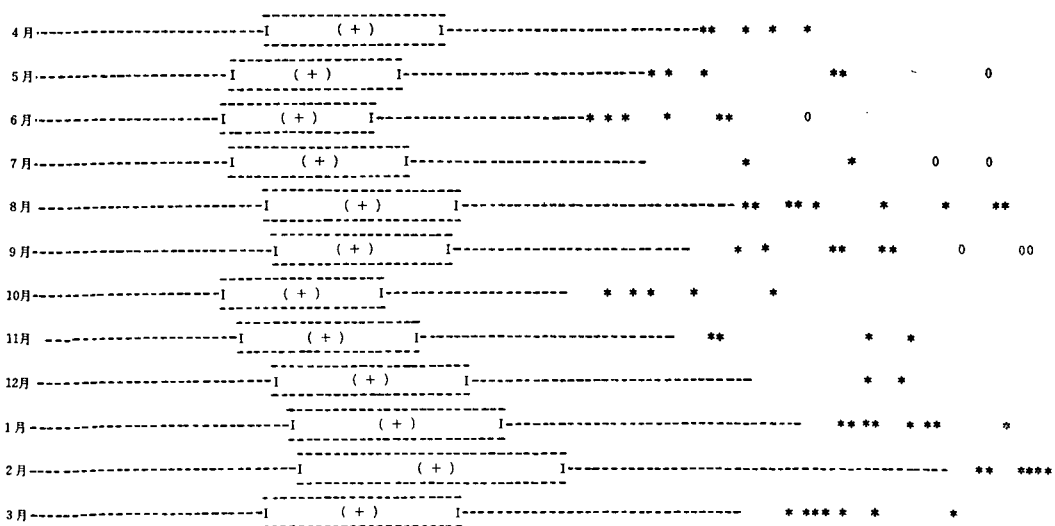


図 3.2　電灯需要分布の箱ヒゲ図（B社，昭和54年度，20A，使用量 10～600 kWh）

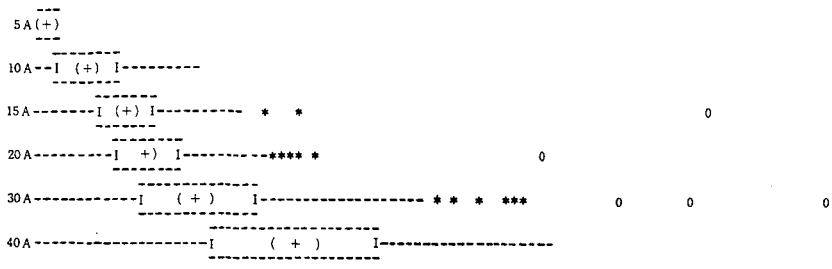


図 3.3 電灯需要分布の箱ヒゲ図（B社，昭和54年4月，使用量 10 kWh 以上）

きくなるにつれて散らばり（箱の長さ）が長くなる傾向がある場合，もとのデータに次のようなベキ変換を施すことにより，散らばりがレベルに依存しなくなり，分析がしやすくなる場合がある[4]。

$$z=\begin{cases} x^{1-b} & (b\neq 1) \\ \log_{10} x & (b=1) \end{cases}$$

最適な b の値は，図 3.4 の散布図（spread-versus-level plot）で回帰直線をあてはめたときの傾きである。ここで用いたデータでは，図 3.4 のように全体としては直線相関に近い傾向があるものの，契約 10 アンペア以下と 15 アンペア以上という2つのグループに分かれているとみることができる。すなわち，この2つのグループで需要構造が異なっていると考えられる。15 アンペア以上のグループについて直線をあてはめて傾きをみると1付近の値であるので対数変換が適切であるといえる。

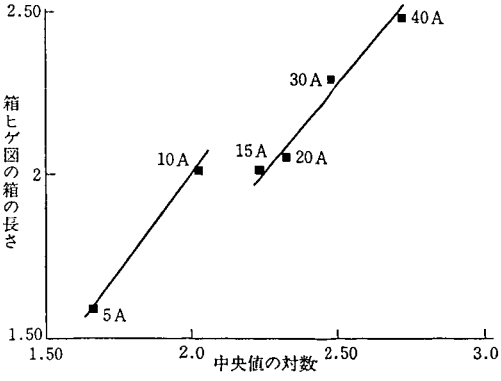


図 3.4 レベルと散らばり方の関係（図 3.3 のデータについて）

4. リサンプリング手法による夏季電力需要分析

4.1 リサンプリング手法の特徴

リサンプリング手法とは，

（1） 既に得られているデータの中から何個かのデータを取り出し仮のリサンプリングデータを作る。

（2） そのリサンプリングデータから統計量を計算する。

という手順を何度か繰り返し，その結果得られた何個かの統計量全体で妥当性を判定する手法である。リサンプリング手法には，

- データの分布を仮定しない
- 厳密な数学的分析により求めることができない統計的属性値を数値的に求めることができる

という利点がある。リサンプリング手法には，Jacknife 法，Bootstrap 法，交差検定法などがある。

本章では，夏季電力需要と気温との関係において直線モデルの回帰係数の分布の推定に Jacknife 法を用いることにより直線モデルが不適切であると判断した例と，折れ線モデルの気温感応変化点の分散の推定に Bootstrap 法を

用いた例を紹介する。

4.2 Jacknife 法による直線モデルの妥当性判定

本節では，夏季における日量（1日の電力使用量）と気温の関係の単回帰（直線）モデルにおける回帰係数の分布を，Jacknife 法を用いて推定する。

Jacknife 法

1947 年に Quenouille [8] は，統計量のバイアスのノンパラメトリックな推定方法として，Jacknife 法を用いた。Tukey [9] は，1958 年に M. Querouille の方法を Jacknife 法と名づけるとともに統計量の分散の推定に Jacknife 法を用い，1970 年代にその考え方を発展させた。

n 個のデータ $X_1, \cdots, X_n$ が与えられた時に統計量 T の属性 A を Jacknife 法により求める手順を以下に示す。

（1） 以下の手順を n 回 ($i=1, \cdots, n$) 繰り返す。

(1—1) n 個のデータから X_i を取り除き $n-1$ 個のデータ $X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n$ を作る。

(1—2) $n-1$ 個のデータ $X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n$ から統計量 T_i を求める。

（2） （1）で求めた n 個の統計量 T_i ($i=1, \cdots, n$) をもとに A を求める。

日量の温度に対する単回帰（直線）モデルの回帰係数の分布

A 電力会社の昭和 58 年 7 月と 8 月の平日（ただし 8 月 15 日～17 日は除く）の 1 日の電力使用量（日量）を被説明変数とし，その日の平均気温を説明変数とした単回帰モデル [2] を考える（図 4.1）。このモデルに Jacknife 法を

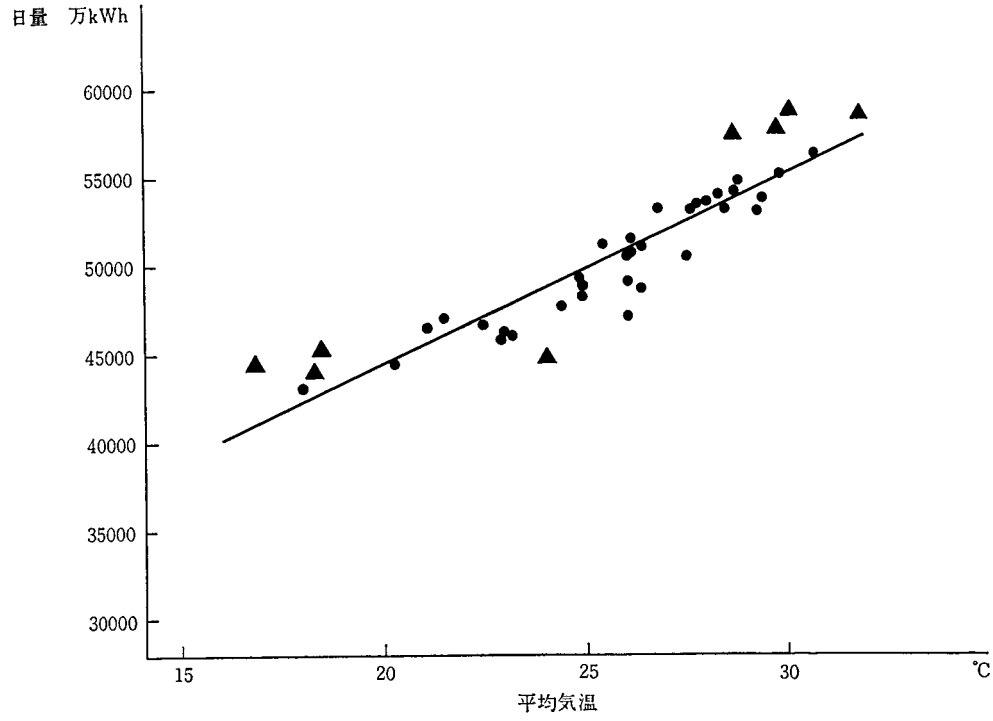


図 4.1 夏季の日量と平均気温の関係（A社，昭和 58 年，平日）

単位＝100kWh

LO 2146, 2223, 2263, 2290 ←⑥

7	229∘	677
8	230*	4
9	230∘	6
12	231*	344
19	231∘	7777899
⑤→(9)	232*	012222234
13	232∘	66899
8	233*	
8	233∘	9
7	234*	3
6	234∘	5

① ② ③ ④

HI ③ 2355, 2356, 2361, 2367, 2383

(1) 定数項

単位＝100kWh/°C

LO 1037, 1044, 1045, 1050

5	105F	4
6	105S	7
8	105∘	89
9	106*	1
17	106T	23333333
(7)	106F	4445555
17	106S	66667
12	106∘	89
10	107*	01
8	107T	333
5	107F	
5	107S	66

HI ⑦ 1086, 1100, 1129

(2) 平均気温

①その行にあるデータ値の深さの最大数（あるデータ値の深さとは、そのデータ値の全データ内での大きい方からの順位と小さい方からの順位のうちの小さい方の数をさす。）
②上位のケタの数字
③上位ケタに＊がつく行には下位1ケタが0、1、2、3、4であるデータが∘がつく行には下位1ケタが5、6、7、8、9であるデータが入る。
④ ②③で決まる行にあるデータ値の下位1ケタの数字の並び（ヒストグラムの棒に対応する∘例えば230∘にある6という数字は2306×100kWhを表わす。）
⑤中央値がある行の深さの欄には、その行にあるデータの個数を記入する。
⑥下方の異常値はLO、上方の異常値はHIと書かれた行に記入する。（異常値は、下方の4分位値をF_L上方の4分位値をF_U、$d_F=F_U-F_L$としたとき、区間〔$F_L-\frac{3}{2}d_F$、$F_U+\frac{3}{2}d_F$〕に入らないデータ値である。
⑦＊、t、f、s、∘の行にはそれぞれ下位1ケタが｛0、1｝、｛2、3｝、｛4、5｝、｛6、7｝、｛8、9｝であるデータが入る。

図 4.2 単回帰（直線）モデルにおける回帰係数の分布の幹葉表示

適用した結果，得られた回帰係数の分布を幹葉表示〔12〕で示したのが図4.2である。そのデータを取り除いた結果，得られた回帰係数が異常値（図4.2の幹葉表示においてHIまたはLOに入っている）となったデータを▲で図4.1に示す。それらの値は温度の低い部分と高い部分に集中しており，単回帰（直線）モデルが夏季の日量を説明するのに不適切であることを示している。

4.3 Bootstrap 法による折れ線モデルの気温感応度変化点の分散推定

本節では，夏季における日量と温度の関係における折れ線モデル〔2〕において気温感応度の変化点の分散をBootstrap法を用いて求めた例を示す。気温感応度の変化点の分散は解析的にその推定式を導くことは困難であるが，Bootstrap法を用いることにより求めることができる。

Bootstrap法

Bootstrap法は1977年にB. Efron〔10〕によって提案された方法であり，Jacknife法に比べてその適用範囲が広い。

n個のデータ $X_1, \cdots, X_n$ が与えられた場合に統計量 T の属性 A を求める手順を以下に示す。

（1） $\hat{F}$を X_i $(i=1, \cdots, n)$ の従う分布Fのノンパラメトリックな最尤推定とする。（この場合 $\hat{F}$はデータ点 $X_1, \cdots, X_n$ にお

いて重み $1/n$ の分布である。）

（2） 以下の手順を適当な回数（M 回とする）繰り返す。

(2—1) $\hat{F}$ に従って n 個のデータ $X_{I_j(1)}, \cdots, X_{I_j(n)}$ を取り出す。

(2—2) $X_{I_j(1)}, \cdots, X_{I_j(n)}$ を用いて統計量 T^{*j} を求める。

（3） （2）によって得られた M 個の統計量 T^{*j} $(j=1, \cdots, M)$ を用いて属性 A を求める。

(2—1) は $X_1, \cdots, X_n$ の中から繰り返しを許して n 個のデータ $X_{I_j(1)}, \cdots, X_{I_j(n)}$ を取り出すのと同等である。つまり，$I_j(k)$ は 1 から n の n 個の整数のうち，1 個の値を j, k に依らず独立に等確率でとると考えればよい。

<u>折れ線モデルにおける気温感応度変化点の分散</u>

日量は平均気温 T_c℃（気温感度変化点と呼ぶ）で傾きの変化する折れ線に回帰すると考える折れ線モデル〔2〕(図 4.3) を考える。つまり，

$$y_i = a + b_1 t_i + b_2 (t_i - T_c)\mathrm{IND}_i + \varepsilon_i$$

ただし，y_i：日量

t_i：平均気温

T_c：気温感応度変化点

$$\mathrm{IND}_i = \begin{cases} 0 & t_i \leqq T_c \\ 1 & t_i > T_c \end{cases}$$

(indicator variable)

と表わすことができるモデルを考える。

このとき回帰係数 a, b_1, b_2 と気温感応度変化点 T_c の推定には最尤推定法を用いる。ここで，IND_i は非解析的な（不連続な）t_i の関数である。従って T_c の分散を求める推定式を導き出すのは困難である。しかし，気温感応度変化点 T_c の分散は Bootstrap 法に従って以下に示す手順で計算できる。

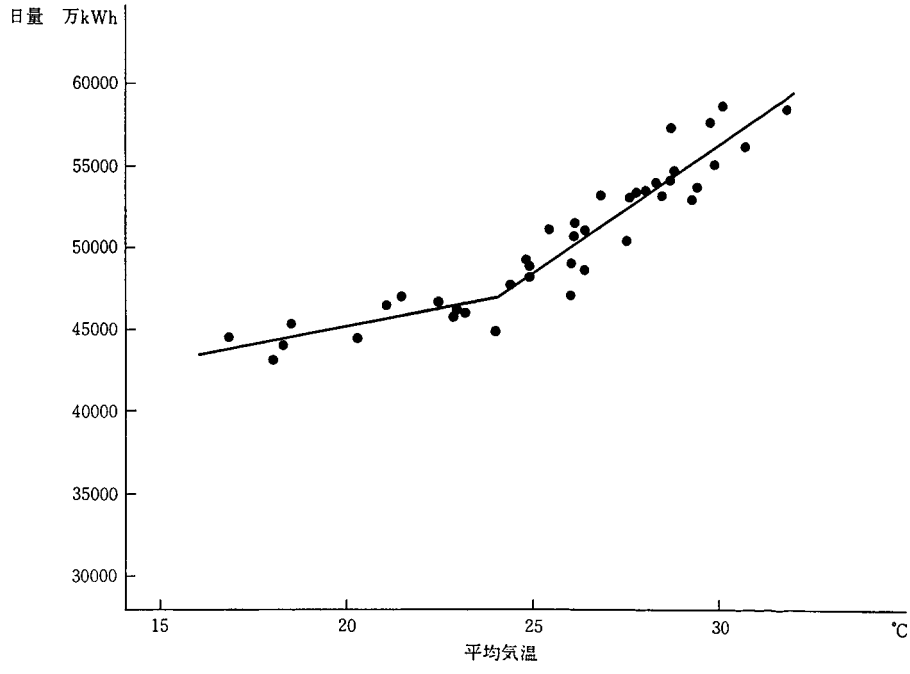


図 4.3 折れ線モデル（A社，昭和 58 年，平日）

（1） $(t_1, y_1), \cdots, (t_n, y_n)$ を用いて a, b_1, b_2, T_c の最尤推定量 a^*, b_1^*, b_2^*, T_c^* を求め，残差

$$\varepsilon_i^* = y_i - (a^* + b_1^* t_i + b_2^*(t_i - T_c^*)\mathrm{IND}_i)$$

を求める。

（2） 以下の手順を M 回 $(j=1, \cdots, M)$ 繰り返す。

(2—1) （1）で求めた推定値 a^*, b_1^*, b_2^*, T_c^*, 残差 $\varepsilon_{I_j}{}^*{}_{(k)}$ $(k=1, \cdots, n)$ を用いて $y_k{}^j = a^* + b_1^* t_k + b_2^*(t_k - T_c^*) + IND_i + \varepsilon_{Ij(k)}$ 作る。ただし $I_j(k)$ は，1から M の整数のうち1つの値を j, k に依らずに独立に等確率で選んだものとする。

(2—2) $(t_1, y_1{}^j), \cdots, (t_n, y_n{}^j)$ を用いて最尤推定法により b^{*j}, b_1^{*j}, b_2^{*j}, T_c^{*j} を求める。

（3） （2）によって得られた M 個の気温感応度変化点 T_c^{*j} $(j=1, \cdots, M)$ の分散を求める。

$$\frac{1}{M-1}\sum_{j=1}^{M}\left(T_c^{*j} - \frac{1}{M}\sum_{j=1}^{M} T_c^{*j}\right)^2$$

が気温感応度変化点 T_c, の推定値 T_c^* の分散の推定値である。

以上の手順をA電力会社の昭和 58 年の7月と8月の平日（8月 15 日～17 日を除く）のデータに $M=500$ として適用した結果，気温感応度変化点 T_c の推定値 T_c^* の分散は 1.342 となった。

5. 今後の課題

今後は，電力需要分析だけでなく，経済分析，工学や生物学におけるデータ解析など，幅広い分野にこれらの手法を適用し，数多くの分析を通してノウハウを蓄積し，さらに正確で質の高い分析に役立てていくことが必要である。

参考文献

〔1〕 Belsley, D. A., Kuh, E., and Welsch, R. E. "Regression Diagnostics, Identifying Influential Data and Sources of Collinearity" Wiley, 1980

〔2〕 小野賢治，森清 堯「夏季電力需要と気象要因」電力中央研究所研究報告 No. 583003, 経済研究所, 1984

〔3〕 井上勝雄「計量経済学の理論と応用」有斐閣, 1983

〔4〕 Hoaglin, D. C., Mosteller, F., and Tukey, J. W. "Understanding Robust and Exploratory Data Analysis" Wiley, 1983

〔5〕 Velleman, P. F. and Hoaglin, D. C. "Applications, Basics, and Computing of Exploratory Data Analysis" Duxbury Press, 1981

〔6〕 森清 堯「電力需要とその分布の統計的分析」電力中央研究所研究報告, No. 679005, 経済研究所, 1980

〔7〕 Efron, B. "The Jackknife, the Bootstrap and Other Resampling Plans" Society for Industrial and Applied Mathematics, 1982

〔8〕 Quenouille, M. "Approximate Tests of Correlation in Time Series" J. R. S. S. Ser. B, 11, 18～84, 1949

〔9〕 Tukey, J. "Bias and Confidence in Not Quite Large Samples" Ann. Math. Statist., 29, p. 614, 1958

〔10〕 Efron, B "Bootstrap Methods: Another Look at Jackknife" Ann. Statist., 7, 1～26, 1979

〔11〕 小野賢治，森清 堯「夏季における電力負荷と気象」電力中央研究所研究報告 No. 584013, 経済研究所, 1985

〔12〕 小野賢治，大屋隆生「電力需要分析のための新しいデータ解析手法とその適用例」電力中央研究所研究報告 No. 584005, 経済研究所, 1985

おの けんじ
おおや たかお
情報システム部
経営情報研究室