日本統計学会誌
第25巻，第3号，1995年
245頁～260頁

あるマーケティング・データのモデル化

柴田里程*，中園美香**

Modelling a Marketing Data

Ritei Shibata* and Mika Nakazono**

An approach to the modelling of a weekly sales data for 13 weeks in 1992, of 9 commodities, from instant coffee to tea bags in 139 various chain stores in Japan, is reported. The logarithm of sales volume of each items per annual sales of each store, are linearly modelled by explanatory variables, logatithm of the price, week effect for each items, location of the store, type of chain stores which the store belongs to, and the prefecture which the store is located in. The fitting of the model is quite satisfactory and the fitted model explains various aspects of the underlying data. However, we found that residuals are highly correlated with the correlation between 0.58 and 0.74, in terms of week, which indicates that some stores are constantly biased upward or downward from the fitted model. This fact suggests that other reasons have to be seeked for. Some of them are found but others are left for further investigation.

1992年後半の13週間について，全国大小のスーパーマーケット139店でのインスタントコーヒー，カレールーなど保存食品9品目の週ごとの売上量を説明するモデルをデータから探索した．最終的に一応満足のいくモデルとして，各週の商品別の平均価格，商品別の週効果，店舗の立地条件，属するチェーンのタイプで説明するモデルが得られた．特に週効果の導入が重要である．ただし，残差を週で層別してみると，週の間隔にかかわらず，常に0.58から0.74の範囲での相関があることがわかり，これをさらに詳しく調べてみると，あてはめたモデルよりも定常的に売上量の多い店，少ない店の存在が判明した．その原因の一つとしては，店の新旧も考えられるが，それ以外にも近辺の人口密度の変化，競合店の存在などさまざまな要因を考慮する必要のあることが明らかになった．本論文は，これら要因をさらに詳しく研究する契機にもなった，探索的データ解析の対象としては比較的大きな実際データのモデル化に関する報告であり，このようなアプローチの有効性を示すものである．

1. はじめに

POS (point of sales) システムの発達により，各種の売上データがあまり人手を煩わさずに瞬時にしかも大量に計算機上に蓄積されるようになってきた．このようにして収集されたデータが現在どのように活用されているかについては [7] に詳しい．当然，そこでは商品管理，売れ筋の把握，小売店の経営支援など日常業務への活用が主である．他方，マーケティング研究のための利用としては，まず記述統計的な利用があり，さまざまなマーケティング理論上の仮説の検証などがある．また特定の店舗を対象に価格効果と商品属性の関連性を数量化I類，II類を用いて確かめる試みも始まっている ([7] の6章)．しかし，統計的代表性，つまりあては

論文受付：1994年6月　改訂受付：1995年9月　受理：1995年9月
* 慶應義塾大学理工学部，〒223 横浜市港北区日吉 3-14-1
** ニールセン・ジャパン（株），〒140 東京都品川区東大井 2-13-8　ケイヒン東大井ビル

めた統計モデルの妥当性の検証や，データの継続性，安定性の検証など，統計的方法を用いる際に不可欠な事柄の検証の必要性は指摘されてはいるものの実際にはいまだに手つかずの状態である．本論文では，このような現状を念頭に，多数の店舗での13週に渡るPOSデータをまず素直に眺め，そこで得られた知見から説明力のあるモデルを探し出し，その妥当性を検証するという，いわゆる探索的データ解析のアプローチをとることにする．特定の経済理論あるいはマーケティング理論から導かれたモデル [3] の妥当性を検証するというアプローチでなくこのようなデータに語らせるといった発見的なアプローチをとることにした1つの理由は，ここで扱うマーケティングデータがごく普通のデータでありながら，その量と多様性からあらじめ適当なモデルを想定するのは困難だと思われたからである．実際，本論文で最終的に得れたモデル (4.2) は，週ごとあるいは店舗ごとのモデル化では不十分で，商品，店ごとの各の価格，週効果，店の立地条件，属するチェーンのタイプ，位置する県などの変量をその交作用まで考慮した同時モデルを考える必要があることを示している．

マーケティング理論の枠組で自己完結的に導かれるモデルが重要であることについてはいまでもないが，一方，データにもとづく探索によって得られたモデルも，新たな枠組での理を展開する手がかりを与えるという点でも同じように重要である．著者たちは，統計科学，るいはもう少し広くデータ科学の独自性を主張しようとするのならば，各応用分野固有のモルとは別に，データから帰納されるモデルを積極的に提示していくことがきわめて重要であと信じている．また，実際にPOSデータを十分に活用しようと思ったら，まずデータから始ことから始める必要があることも確かである．

もちろんこのような探索的なアプローチをとるためには，大量のデータに対してさまざまモデルを簡単にあてはめ，試し，修正できる柔軟な計算環境が必要であり，また，日頃このようなデータを扱っている人間の知識がかかせない．幸い近年の計算機技術の進歩と先進的フトウエアの出現により前者の問題は画期的に改善された．実際ここではAT&Tベル研で開発され日本でも普及してきたS [4, 5, 6] を用いてモデルの探索と評価を行っているのソフトウエアを用いた解析の過程で得られた副産物として，データの変容，名前の管理計算機統計の分野に属するさまざまな知見も得られたが，本論文ではこの点に関しては必範囲で触れるにとどめる．後者に関しては，ニールセン・ジャパン（株）の御理解と，著の間の密接な共同研究体制により実現することができた．データから適切なモデルをどのに構築したらよいかは古くから統計解析の重要な関心事であり，近年人工知能技術によ験を自動的に積み上げようとする試みも数多くなされているが，やはり，対象とするデ多様性に比してまだまだ経験の蓄積が不足し，蓄積の組織だった方法さえも確立されてのが現状であろう．この節目に本論文が1つの組織だった蓄積の一助になり，統計学会姉の今後の研究の助けになれば望外の幸せである．

2. 対象とするデータ

ここで解析の対象とするデータは，全国大小のスーパーマーケット139店における，タントコーヒー，カレールーなど保存食品9品目の週ごとの売上高と売上個数の，1992年40週（9月28日から10月4日まで）から第52週（12月21日から12月27日まで）分のデータである．実際には各商品はパッケージの違い，ブランドの違いなどによりさ分化されており，1112種類のJAN (Japanese Article Number) コードで識別されている論文では，とりあえず個々のブランドによらない，モデルの探索を目的としたいので，ドの違いは無視してデータを集約 (aggregate) した．ただし，商品のパッケージによいは反映した集約を行った．このほかに利用する背景情報としては以下で述べるよう

所在地，タイプ，属するチェーン，食品と日用雑貨品の年間総売上高がある．住宅街か駅前かなどの立地条件，レジの台数，面積，開店時間，定休日，駐車場の収容台数などの背景情報なども利用できたが，店の規模に関しては年間総売上高で正規化するので立地条件以外は結果的にモデルには取り入れる必要がないことがわかったので，ここでは省略する．

ここで用いたPOSデータは，店名，商品のJANコード，週をキーとする売上高と売上個数であるが，集約の段階で，欠損値などのため利用できない部分があり，最終的には17県に渡る139店における9商品（1店だけは8商品）に関する各週1250件，つまり13週で総計16250件の売上高と売上量，その他の説明変量の値のデータとなった．したがって，原データでは，店ごとにまた商品ごとにもデータ数は同一であったが，集約後は各変量について観測数の釣合がとれたデータではなくなっている．ここでは，釣合のとれるようにその1部だけをとりだしたりすることはせず不釣合のまま解析することにする．したがってこのことを，モデルの探索の段階でも，検証の段階でも常に注意する必要がある．なお，これら139店の属するチェーン総数は38である．一口にPOSデータといっても，さまざまな段階があり，日常業務で用いられているPOSデータは最低でも1日単位，場合によってはオンラインで刻々収集されている．このような莫大なデータの処理という問題も，統計的問題として興味深く重要ではあるが，本論文の目的とは離れるので，ここではこれ以上触れないことにする．

3. データの特徴の探索

集約したデータを次のような変量を持つデータ・フレーム [5]，つまり1つの関係形式（relational scheme，ただし各列に行列データも許す）にまとめた．

- pc：以下のような9種類の商品の区分（product class）を表す因子変量．
  インスタントコーヒー　カレールー等　シリアル
  トマト・野菜ジュース　トマトケチャップ　ドレッシング
  マーガリン　マヨネーズ　紅茶
- chain：38のチェーン（chain）を識別する因子変量．
- store. id：139店（store）を識別する因子変量．
- pref：以下のような各店の所在県（prefecture）を表す因子変量．
  愛知　茨城　岐阜　京都　群馬　埼玉　三重　神奈川　静岡　千葉
  大阪　長野　東京　栃木　奈良　兵庫　和歌山
- type：ニールセンでの店のタイプの3分類，GMS (general merchandize store；従業員50人以上で，衣料品，食料品，(家電，家具などの)住居関連品の売り上げ比率が，それぞれ10％以上70％未満)，SML (small marchandize store : large ; GMSでないが，年間総売り上げが10億円以上)，SMS (small marchandize store : small ; GMSでもなくSMLでもない）を表す因子変量．
- acv：(all commodity value)，年間食品・日用雑貨総合売上高（1991年度）．単位100万円．
- sales：該当の商品の各店での一週間の売上高13週分．単位は円
- volume：該当の商品の各店での一週間の売上量13週分．単位はkg．
- price：sales/volumeによって求めた該当の商品の各店での1kgあたりの週平均価格13週分．単位は円/kg．

なお，以下では各因子の水準を上に掲げた順序での水準番号1，2，…，9，a，b，c，…で示すこともある．たとえば，pcの水準1は「インスタントコーヒー」，水準2は「カレールー等」を意味する．これらの変量のうち，数値変量であるsales, volume, acvの分布は右にゆが

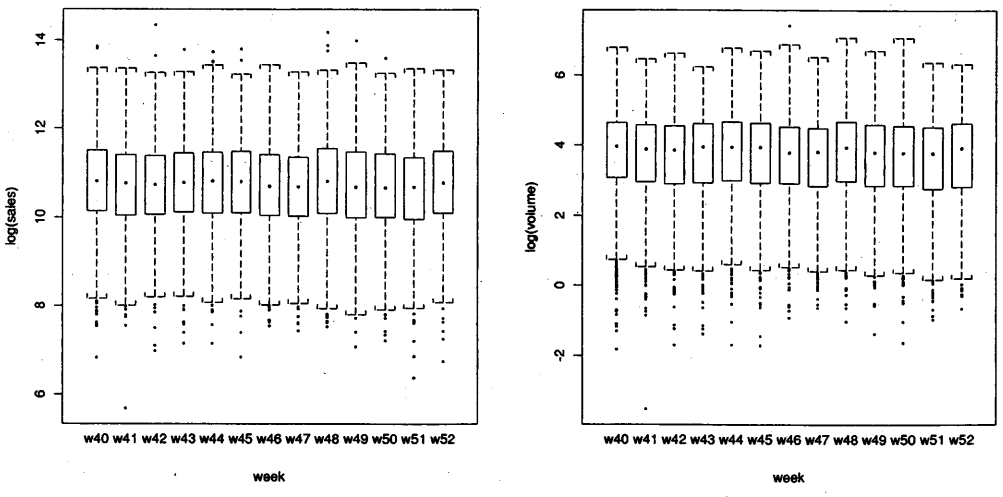

図1　sales と volume の対数変換後の箱型図

んだ分布をしており，このままでは線形モデルを通常の最小二乗法であてはめた際に問題を生ずる．しかし，この場合にもよく知られているように，対数変換でそのゆがみを是正することができる．図1は sales と volume の対数変換後の週別の箱型図で，特に小さな値の側に，いくつかのはずれ値が散見されるものの対数変換の妥当性を示している．ここでは図示を省略するが acv についても同様であり，price についてもその定義から対数変換が妥当であることになる．もちろん，この図はさまざまな説明要因を無視した箱型図であり，週ごとに異なる変換が必要かどうかの大ざっぱなチェックのためだけに用いた．

4. モデルの探索

まず，反応変量として sales に注目するか volume に注目するか選択の余地がある．通常は商品ごとに独立に売上高を反応変量として考え解析することが多い．しかし，ここでは異なる商品の売上に含まれる各店の特徴などの情報も有効に利用したモデル化を行いたいので，単価の異なる商品の売上高である sales をそのまま反応変量とはせず，売上量である volume を基本的な反応変量と考えてモデル化する．ただし，volume そのままではなく，各店の食品に関する売上規模である acv で規格化した log(volume/acv) のモデル化である．したがって変量

$$\log(\text{volume}/\text{acv})=\log(\text{sales}/(\text{acv}*\text{price}))$$

のモデル化ともみなせるので，以下のモデル化は log(sales) を，log(acv)＋log(price) に各説明変量による線形な寄与を加えた形でモデル化したと考えてもよい．

説明変量としては各商品の価格がもっとも基本的であるが，ここでは売上量のうち価格で説明できる部分の絶対的な値に興味があり，週ごとの変化も調べたいので，通常の回帰モデルのような1つの一般平均を導入したモデルではなく，pc の各水準ごとに異なる一般平均を導入する平行回帰（parallel regression）モデル

$$\log(Y_{ijk})=\alpha_i+\beta_i X_{ijk}+\varepsilon_{ijk} \qquad (4.1)$$

を用いた．ただし，Y_{ijk} と X_{ijk} はそれぞれ，第 i 商品の第 j 週の第 k 店での基準化した売上量 volume/acv と，価格の対数 log(price) を pc の水準ごとに，つまり商品ごとの平均を除去して0に中心化（centered）した c. log. price であり，ε_{ijk} は独立な誤差である．ここで用いた S で

のモデルの記述方式［5］を用いれば，このようなモデルは次のように簡単に表せる．

$$\log(\text{volume/acv}) \sim -1+\text{pc}+\text{pc}:\text{c.log.price}$$

～の左辺が反応変量，右辺がそれを説明する変量を表す．記号＋は単に各変量の区切りを意味する．実際には各変量は，因子変量（factor variable，カテゴリー変量）ならばコーディングされた上で，数値変量ならそのままで，係数パラメータが導入され，線形モデルとなる．また，記号：は変量間の交互作用を表す．−1が一般平均を導入することを禁止しているのでpc＋pc：c.log.priceはpcの各水準ごとにc.log.priceの線形関数をあてはめることを指定していることになる．ただしc.log.priceの代わりにlog(price)をそのまま用いて

$$\log(\text{volume/acv}) \sim -1+\text{pc}+\text{pc}:\log(\text{price})$$

とすると，pcは0と1の値をとるダミー変量でコーディングされ，交互作用pc：log(price)は各ダミー変量とlog(price)の積でコーディングされるので，log(price)の変動がその尺度に比べて少ない場合にはpcの係数とpc：log(price)の係数の推定量の間の相関が極めて−1に近くなる．これは，単回帰の場合に，説明変量の値を$x_1, \cdots, x_n$として，定数項と傾きの推定量の相関が$-(\sum x_i/n)/(\sum x_i^2/n)^{1/2}$となることと本質的に同じである．どちらのモデルを用いても，あてはめたモデル自体にはかわりはないが，ここでは，まず各週ごとにモデルをあてはめ，その係数の動きから説明変量の選択を行うので，前者のモデルでないと価格に関する係数の動きの解釈を誤る可能性がある．

さらに各店の特徴をchain, pref, typeで表すことにしてモデル

$$\log(\text{volume/acv}) \sim -1+\text{pc}+\text{pc}:\text{c.log.price}+\text{chain}+\text{pref}+\text{type}$$

を各週ごとにあてはめた．因子変量のコーディングとしては，欠損値などによりもともと観測数がそろっていないので，直交対比を用いる積極的な理由はなく，結果の解釈のしやすさから処理対比[5]を用いることにする．このあてはめ結果のうち，特に変量typeの各水準の係数推定値の週別変化を示したのが次の図2である．Sでは処理対比を指定すると，因子変量は常に第1水準を0にコーディングするので，図では第1水準つまり愛知県の係数が常に0となっている．

図2から水準1のGMS（総合店）と，残りの水準2のSML（総合店ではないが年間総売り

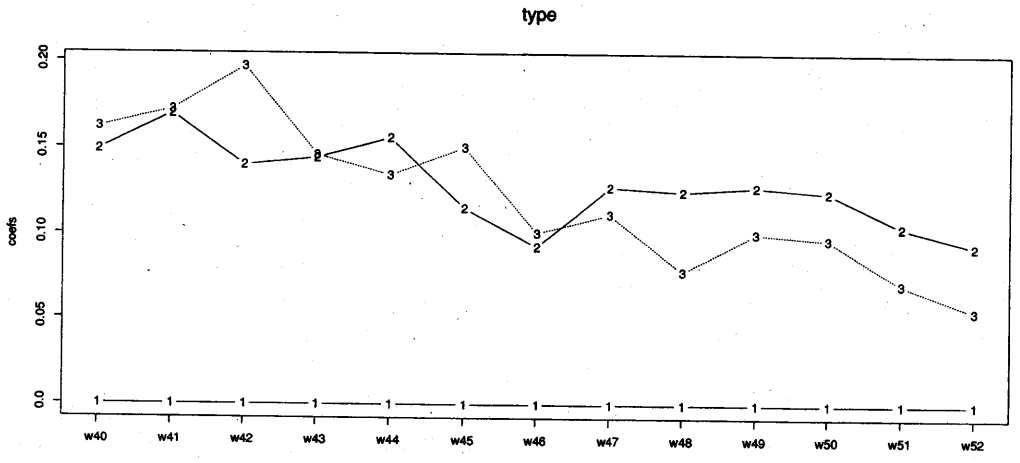

図2　変量typeの各水準の係数推定値の週別変化

上げ10億円以上)，水準3のSMS（その他）との違いがはっきりするが，それがなぜ次第にGMSに近づくのかの説明が困難である．また，ここでは図示を省略するが，変量chainの38水準間の違いもあまりはっきりしない．そこで，変量chainとtypeの代わりに次の2変量を用いることにした．

- location：日本スーパーマーケット名鑑［1］の店舗立地条件から，店舗を
  駅前　商店街　住宅地　郊外
  の4水準に区分する因子変量．
- chain.type：各スーパーマーケット・チェーンを，その店舗数，店舗の展開状況により
  全国小　全国大　地域小　地域大
  の4種類に大分類する因子．ただし，「全国」とは関西と関東にまたがって店舗を展開しているチェーンであり，「地域」はそれ以外のチェーンである．「大」，「小」は「全国」の場合は30店舗，「地域」の場合は20店舗を目安に分類した．

つまり，モデル

$$\log(\text{volume}/\text{acv}) \sim -1+\text{pc}+\text{pc}:\text{c.log.price}+\text{location}+\text{chain.type}+\text{pref}$$

を週ごとにあてはめることにした．ただし，上述の資料からは立地条件の不明な店舗が17店あったので，このモデルではこれが不明な店に関するデータは利用できなくなり，13週間の総計14261件のデータに対してのあてはめを行うことになった．以下に各変量に関するデータの要約を掲げておく．

volume		acv		pc	
Min.	: 0.03	Min.	: 450	紅茶	:1586
1st Qu.	: 18.07	1st Qu.	:1132	マヨネーズ	:1586
Median	: 47.24	Median	:1661	マーガリン	:1586
Mean	: 70.86	Mean	:1892	ドレッシング	:1586
3rd Qu.	: 97.59	3rd Qu.	:2260	トマトケチャップ	:1586
Max.	:1186.00	Max.	:7234	トマト・野菜ジュース	:1586
				(Other)	:4745

c.log.price		location		chain.type		pref	
Min.	:−0.768900	駅前	:3393	全国小	:1989	東京	:3042
1st Qu.	:−0.079330	郊外	:1170	全国大	:2808	大阪	:1755
Median	: 0.004847	住宅街	:7475	地域小	:4680	愛知	:1521
Mean	:−0.001513	商店街	:2223	地域大	:4784	神奈川	:1404
3rd Qu.	: 0.084170					埼玉	:1053
Max.	: 0.725400					静岡	: 936
						(Other)	:4550

この要約では数値変量volume, acv, c. log. priceに関しては，その最小値，最大値，平均値と共に四分位点も示されている．特にvolumeの平均と中央値が大きくずれているのは，すでに注意したように，大きく右にゆがんだ分布をしていることの反映である．それに比べるとacvの分布はそれほど大きくゆがんではいないが，その範囲が4億5千万円から72億3千5百万円にまで及んでいることは，どうしてもacvによる規準化と対数変換が必要なことの裏付けにもなっている．残りの因子変量に関しては，各水準ごとの観測数が示されている．pcの観測数1586を週数13で割った122が店の数であるが，(other)の3商品の計が4758でないのは，すでに触れたように「シリアル」に関してある店舗の売り上げが13週に渡って不明なためである．loca-

tion に関しても週数×商品数＝117 で割った観測数が各水準に属する店舗数である．住宅街の店舗数が多いのはデータの性格上やむを得ないが，結果に大きく影響を与えるほどの違いではない．chain. type に関しても同様である．pref に関しても，(other) は残り 11 県の総計を示しているが，店舗数は最低の長野県でも 17 店，それ以外は 27 店以上あり，極端に店舗数の少ない県はない．なお，さきにも述べたように，これらの変量以外に，各店舗の面積や駐車場の収容台数などの変量も説明変量として加えてみたが，いずれも寄与は低い．反応変量として店の規模で基準化した売上量を考えているため，店舗の規模のもう 1 つの指標である面積の説明力があまりなくなり，駐車場の収容台数も location, chain. type, pref などからある程度定まってしまうのは当然である．

次に各変量の係数の時間的な変化を示す図 3 を与えておく．左側の図に示された各商品ごとの回帰直線の切片の週ごとの変化はあまり大きくなく，値の大きい方から掲げれば，ほぼ，「マヨネーズ」，「カレールー等」，「マーガリン」，…，「インスタントコーヒー」，「シリアル」，「紅茶」の順に並んでいる．これは，価格，店の特徴，立地条件，などに依存しない基本的な各商品の各週での売上量と解釈できるが，売上量は kg 単位の重さであるので，「マヨネーズ」や「カ

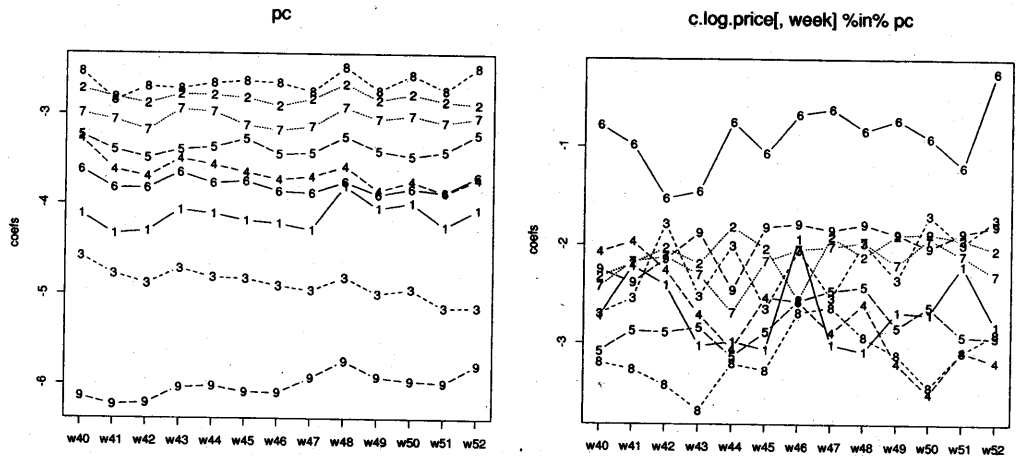

図 3　商品別の平行回帰の切片と傾きの週別変化

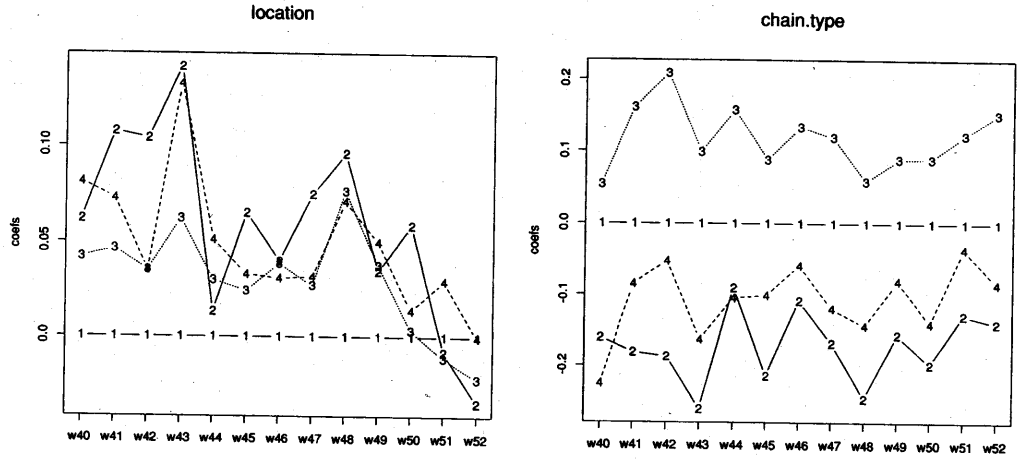

図 4　変量 location と chain.type の各水準の係数推定値の週別変化

レールー」の値が大きく，シリアルや紅茶の値が小さくなるのは当然である．一方，回帰直線の傾きである右側の図を眺めると切片よりも週ごとの変動は大きいものの水準6の「ドレッシング」の値が特に0に近い．これはこの商品の売上量が価格の変動に対してあまり敏感ではないことを示唆している．逆に水準8の「マヨネーズ」などは価格の変化に対して敏感である．なお，図の表題のc.log.price[,week]%in%pcはc.log.price[,week]:pcと同等で，各商品ごとに異なる傾きの回帰直線を導入したことを示している．

図4は変量locationとchain.typeに関する同様な図である．ただし，第1水準の係数は常に0であるので，第1水準に対する相対的な変化とみる必要がある．図4の左側の図から，立地条件に関しては，ほぼ，「商店街」，「郊外」，「住宅地」，「駅前」の順になっている事がわかる．しかし週が進むにつれ，右下がりに落ちており，時間的変化をもこの変量が説明してしまっている．このように4変量のうち，もっぱら立地条件の係数に顕著な時間的変化が現れているのは，それはそれで興味深いが，実は週ごとに独立なモデルをあてはめる解析の限界を示すものでもある．右側の図から，チェーンのタイプに関しては特別な週変化は見られないが，「地域小」，「全国小」，「全国大」，「地域大」の順になり，小規模の，その中でも全国展開していないようなチェーン店のほうが，ここで取り上げたような日常的な保存食品に関しては売上比率が高いことを示している．図5は変量prefの係数の週別変化である．この図では，水準dの東京と水準8の神奈川が高く，逆に水準5の群馬や水準7の三重，さらには水準eの栃木が低く1つの地域性を示している．ここで対象としている商品は，いわゆる工場生産の保存食品であるので，都会化の影響が大きく現れてくるのは当然であろう．週別変化に特定の傾向がないことは，このような影響がきわめて安定的であることを示している．

以上のような予備的な解析によって基本的なモデルが得られたが，変量locationの係数の変化に含まれているような時間効果を分離するには，週を表す変量weekを導入し，13週すべてのデータへの同時あてはめを行う必要がある．以上の週ごとの解析結果も考えあわせて，時間効果は商品別に導入するのが適当と判断し，モデル

$$\log(\text{volume}/\text{acv}) \sim -1+\text{pc}+\text{pc}:\text{c, log. price}+\text{pc}:\text{week}+\text{location}+\text{chain. type}+\text{pref}$$

を13週のデータすべてにあてはめることにした．ただし，このためには新たな変量weekに対

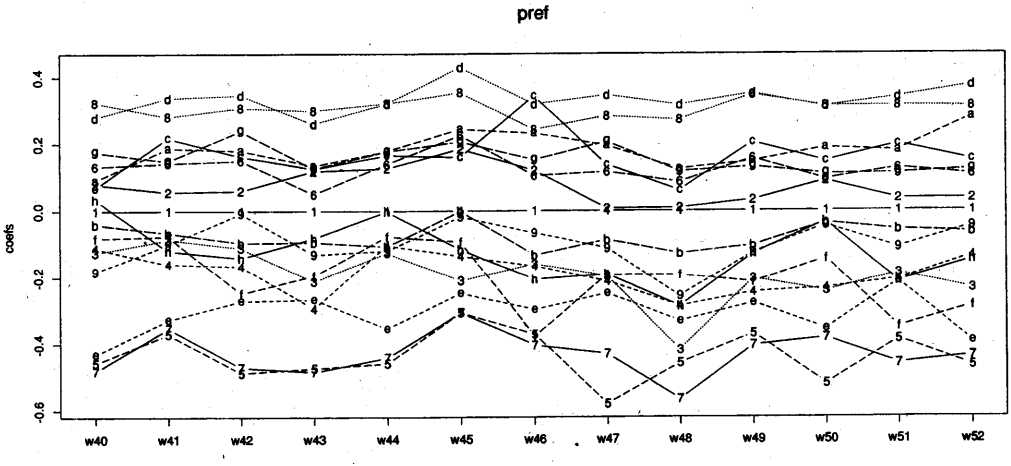

図5　変量prefの各水準の係数推定値の週別変化

応した各変量の値13週分を1列にならべた形式のデータに直す必要がある．念のため，このモデルを伝統的な数式で表現しておくと，(4.1)と同じように Y_{ijk} と X_{ijk} をそれぞれ，第 i 商品の第 j 週の第 k 店での基準化した売上量 volume/acv と価格 c.log.price とすれば，

$$\log(Y_{ijk})=\alpha_i+\beta_i X_{ijk}+\lambda_{ij}+\mu_{l(k)}+\nu_{m(k)}+\xi_{n(k)}+\varepsilon_{ijk} \qquad (4.2)$$

となる．ただし，$l(k), m(k), n(k)$ はそれぞれ第 k 店に関する location, chain. type, pref の水準番号であり，ε_{ijk} は誤差である．ここではモデルのあてはめを最小二乗法で行うので，少なくとも誤差に関して独立性と，正規分布と比べてそれほど裾の重くない同一分布を仮定する必要がある[5]．次節以降では，このモデル (4.2) のあてはめ結果について詳しく検討する．

5. あてはめたモデルの検討

5.1 係数の推定値

前節で述べたモデルのあてはめのよさを議論する前に，推定された係数を眺めてみよう．まず，次の分散分析表からわかるようにデータ総数14261に対して係数は148個で，モデルとして決して大きすぎはしない．実際，残差分散は0.15とかなり小さい．この表の各行は各変量あるいは交互作用項に対応しており，Df は自由度，つまり係数の個数で，Sum of Sq は平方和，Mean Sq はそれを自由度で割った平均平方和，F Value は F 統計量の値，Pr(F) は p 値である．どの変量，項も残差にくらべてきわめて有意であるが，特に pc の水準別の平行回帰部分の寄与が大きい．

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
pc	9	214844.7	23871.63	159132.3	0
location	3	59.5	19.83	132.2	0
chain.type	3	27.2	9.07	60.4	0
pref	16	523.8	32.73	218.2	0
c. log. price %in% pc	9	1660.9	184.55	1230.2	0
pc : week	108	101.5	0.94	6.3	0
Residuals	14113	2117.1	0.15		

モデル (4.2) をあてはめたときの各変量の係数は以下の値からわかるように週ごとのあてはめ結果から得られる係数の週に関する平均とほぼ一致しており，その標準偏差つまり標準誤差は十分小さい．

表1は平行回帰部分の係数をまとめたもので，前節の考察と同じように，とくに傾きに関して「ドレッシング」が例外的に小さな絶対値をもっている．これは，この商品の売上量が価格にあまり影響されないことを示唆している．しかし，これ以外の商品の係数は比較的大きな絶対値をもっており，その差も有意である．これら傾き β は，反応変量の対数変換をもとへ戻せば

$$\text{volume}=\text{acv}\cdot\text{price}^{\beta}\cdots$$

の関係で売上量に関係していることがわかる．その絶対値が大きな商品ほど，売上量が価格からより大きな影響を受けるといえる．以下，他の変量については，その推定係数の大きさの順に並べ換えてその標準誤差とともに示しておく．

表2は立地条件に関する係数である．ただし，NA は欠損値 (Not Available) を表す．先にも述べたように，この因子変量を処理対比でコーディングする際，水準1の「駅前」を0に，

表1　商品別の平行回帰の回帰係数

商品	切片 α_i		傾き β_i	
	係数推定値	標準誤差	係数推定値	標準誤差
インスタントコーヒー	−4.139	0.03805	−2.705	0.05951
カレールー等	−2.778	0.03807	−2.094	0.0691
シリアル	−4.650	0.03818	−2.222	0.1028
トマト・野菜ジュース	−3.352	0.03804	−2.783	0.06890
トマトケチャップ	−3.285	0.03805	−2.811	0.07000
ドレッシング	−3.671	0.03806	−0.9425	0.1108
マーガリン	−3.034	0.03804	−2.200	0.6604
マヨネーズ	−2.657	0.03810	−3.187	0.07974
紅茶	−6.143	0.03808	−2.000	0.05317

表2　変量 location の各水準の係数推定値 μ_l とその標準誤差

	駅前	住宅街	商店街	郊外
係数推定値	0.0	0.02920	0.04846	0.05701
標準誤差	NA	0.008532	0.01190	0.01546
exp(係数推定値)	1.0	1.030	1.050	1.059

その他を1に割り当てているので，これらの係数は「駅前」を基準にした相対値とみる必要がある．また，表の3行目に対数変換を元へ戻したときの変量 volume への各水準の効果を示してある．つまりこの値を γ とすれば

$$\text{volume} = \gamma\,\text{acv}\cdots$$

の形で変量 volume の値に影響し，それぞれの水準によって定まる倍率になっている．これは以下すべての因子変量に関していえることである．表2からわかるように，「商店街」と「郊外」の差は有意ではないが，それ以外の間では有意な差があり，ここで扱ったような商品に関しては，「駅前」よりも「住宅街」，さらには「商店街」あるいは「郊外」の店での売上量のほうが多いことがわかる．ただし，volume に対する効果は高々+5.9％程度であり，以下の他の変量に比べてそれほど売上量に対する影響が大きい変量ではない．

表3は明らかにチェーンの展開形態と規模による違いが−16％から+12％程度存在することを示している．

表4からわかるように，所在県による違いも，−36％から+40％の範囲で，ほぼこの順で売上量に影響する．特に兵庫が+16.7％であるのに対し，大阪が−7.55％であるのは興味深い．図6は，各商品ごとの週効果である項 pc : week の係数推定値の図示である．最初の週を0としてコーディングしているので，値は第40週に対する相対値である．なお，標準誤差はすべての商品，週について 0.0496 から 0.04991 の範囲に収まっている．

図6から，9月末の第40週から年末の第52週にかけて水準9の「紅茶」の効果が増加して

表3　変量 chain. type の各水準の係数推定値 ν_m とその標準誤差

	全国大	地域大	全国小	地域小
係数推定値	−0.1750	−0.1080	0.0	0.1154
標準誤差	0.01258	0.01139	NA	0.01183
exp(係数推定値)	0.8394	0.8976	1.0	1.122

表4　変量 pref の各水準の係数推定値 ξ_n とその標準誤差

	群馬	三重	栃木	奈良	岐阜
係数推定値	−0.4397	−0.4311	−0.3097	−0.1994	−0.1905
標準誤差	0.02849	0.02361	0.02353	0.02831	0.02033
exp(係数推定値)	0.6443	0.6498	0.7336	0.8192	0.8265

	京都	和歌山	静岡	大阪	愛知	茨城
係数推定値	−0.1853	−0.1274	−0.1005	−0.07850	0.0	0.07234
標準誤差	0.02345	0.02408	0.01713	0.01405	NA	0.02102
exp(係数推定値)	0.8308	0.8804	0.9044	0.9245	1.0	1.075

	埼玉	兵庫	長野	千葉	神奈川	東京
係数推定値	0.1239	0.1543	0.1609	0.1835	0.3045	0.3337
標準誤差	0.01605	0.01893	0.02906	0.01809	0.01484	0.01282
exp(係数推定値)	1.132	1.167	1.175	1.201	1.356	1.396

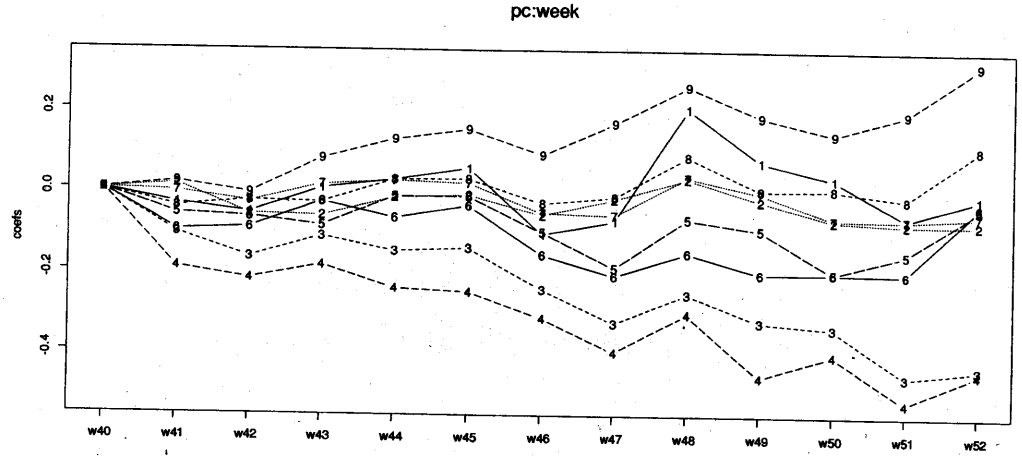

図6　商品ごとの週効果 λ_{ij}

いるのに対し，水準3の「シリアル」と水準4の「トマト・野菜ジュース」の効果は逆に減少していることがはっきり読みとれる．これは，価格，店の特徴などとは無関係な，季節が秋から冬へ移り変わることにともなう時間効果である．これ以外にも暮れの第52週にはすべての商品の売上量が増えていることがわかる．第48週でもすべての商品についてこの効果が増加しているが，この原因はいまのところ不明である．

5.2　あてはめのよさ

残差の様子を探るため，あてはめたモデルにもとづく予測値に対する実際の反応変量の値の散布図を描いたのが図7である．14261点のうち2点ほど極端にあてはまりの悪いデータがあるが，これは次の絶対残差の散布図である図8を眺めるといっそうはっきりする．残差の絶対値が2.5を越える2点について各変量の値を表5に示しておく．

表5の紅茶は，残差が大きくずれている場合で，この週の平均価格が1 kgあたり4983.254円であるのに対しその約倍の値段であり，1週間で30 kgしか売れていない．また，volume/acvのこの週の紅茶に関する平均は0.002838であり，このように規模で規格化した売上量で見ても約1/10である．実際，このデータに関しては，データの提供元に問い合わせ訂正することができた．よく指摘されることではあるが，データに対する適当なモデル化を行ったときの1つの

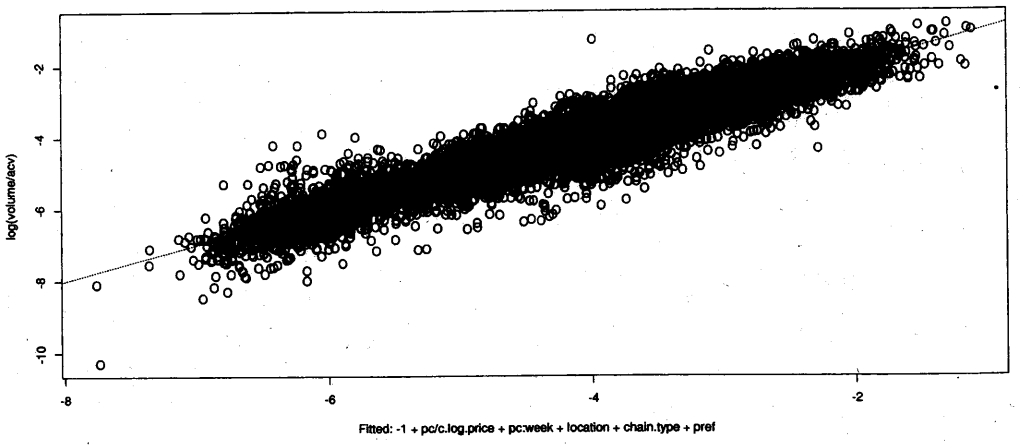

図7 あてはめ値に対する反応変量の値の散布図

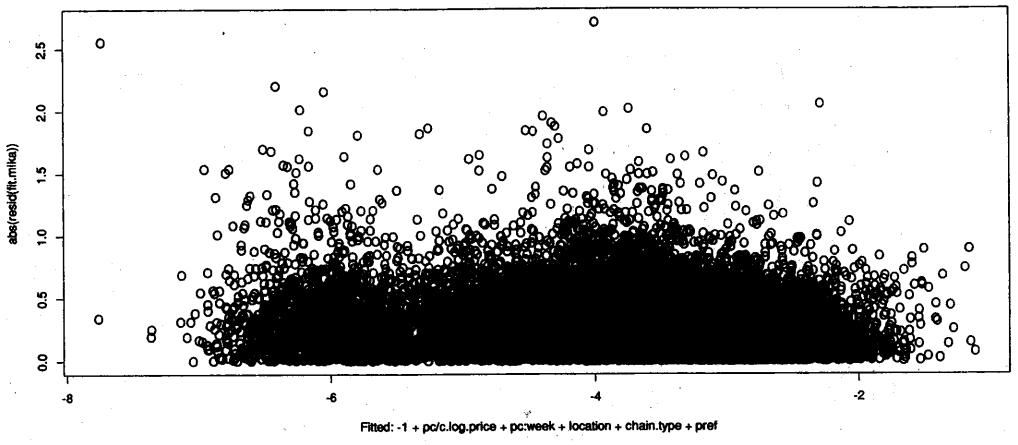

図8 あてはめ値に対する絶対残差の散布図

表5 残差が極端に大きな2つのデータ

pc	week	volume	price	volume/acv	pref	location	chain.type
紅茶	41週	0.0300	9866.667	0.00003410	栃木県	住宅街	地域小
インスタントコーヒー	42週	387.7444	4357.030	0.2746	岐阜県	郊外	地域小

大きな副産物は，このような異常値の発見など，データのチェックや浄化(cleaning)が容易になることがあげられる．次の図9からもこのデータがかなり極端であることがわかる．

逆にインスタントコーヒーの場合は，この週の平均価格が1kgあたり4781.227円であるのに対し，4357.030円は割安であり，387.7444kgも売れている．volume/acvでみてもこの週の平均は0.02100であり，10倍近く売り上げている．よく調べてみるとこの店はときどき大量販売をすることで有名な店であることが判明した．この他の残差の大きなデータについても詳しく検討したが，たとえばその大きさが3σ，つまり今の場合3×0.3873より大きな残差をもつデータは43件，-3×0.3873以下の残差をもつデータは104件にすぎない．両方合わせても総データ数の約1％である．しかもその内訳は，前者に関しては40週の1店での4商品と，46週の

1店での1商品を除いて，「インスタントコーヒー」，「紅茶」，「トマト・野菜ジュース」の3商品であり，後者の場合も，104件のうち7件を除いて「シリアル」が7件あるほかは上記の3商品に限られる．これら3商品はいわゆる目玉商品として取り上げられやすいだけでなく，図10

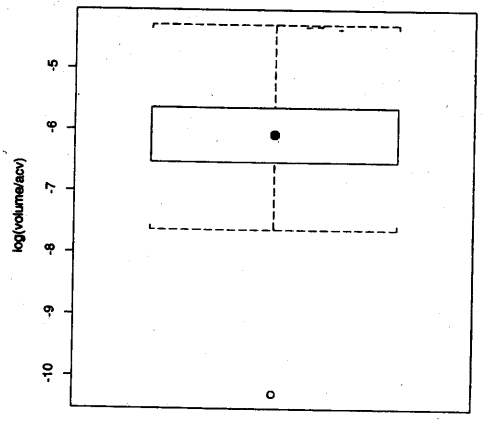

図9　第41週の紅茶に関する log(volume/acv) の箱型図

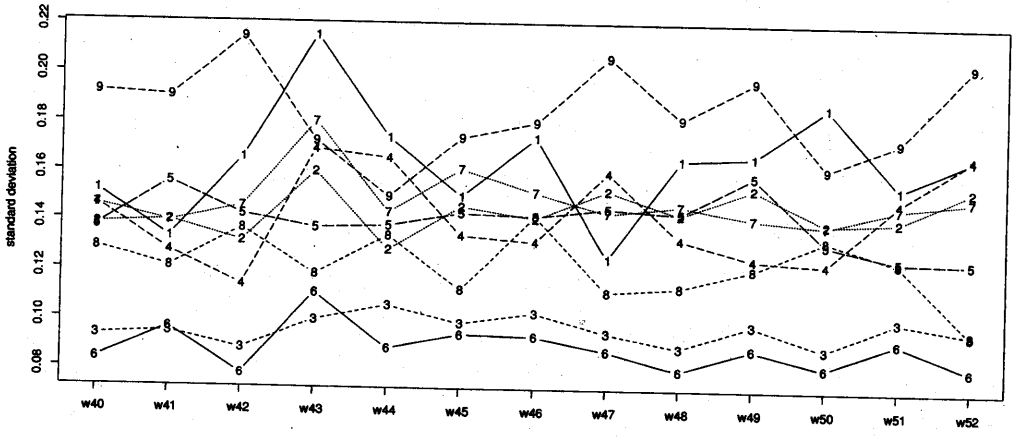

図10　変量 c. log. price の週別の標準誤差

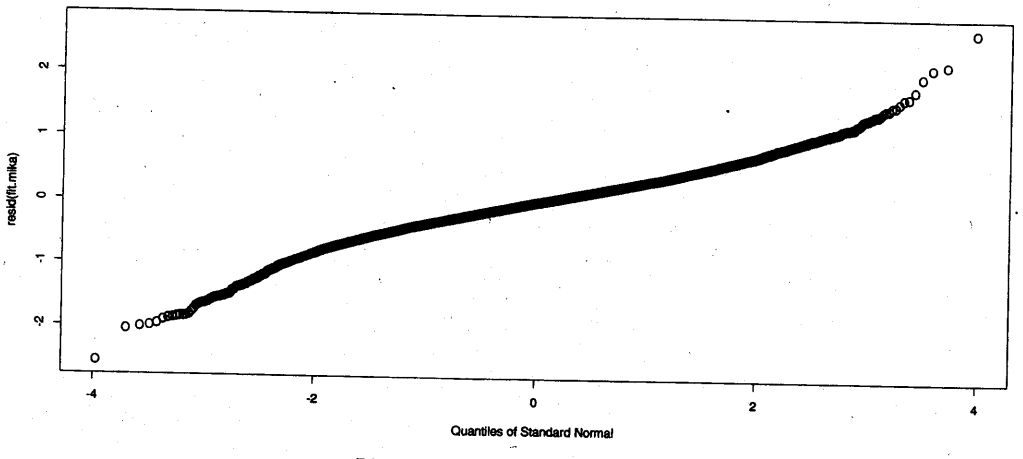

図11　残差の正規 QQ プロット

からわかるようにc. log. priceの標準偏差が大きく，また，その週による変化が大きいことからもこれらの商品の特徴が裏付けられる．

次に，残差の分布のチェックをしておこう．図11からわかるように残差は正規分布より少し裾の重い分布をしている．これに関しては次節でさらに詳しく議論するが，一言でいえば，location, chain.type, prefなどでは説明しきれない，各店舗の特徴によるものと考えられる．

5.3 残差の層別

残差を週ごとに層別し，その相関を調べて見ると，意外なことが判明した．

	40	41	42	43	44	45	46	47	48	49	50	51	52
40	1.00												
41	0.69	1.00											
42	0.63	0.74	1.00										
43	0.65	0.69	0.67	1.00									
44	0.68	0.68	0.65	0.69	1.00								
45	0.62	0.66	0.63	0.68	0.69	1.00							
46	0.63	0.66	0.64	0.65	0.68	0.69	1.00						
47	0.64	0.66	0.65	0.67	0.67	0.68	0.70	1.00					
48	0.60	0.61	0.59	0.64	0.65	0.67	0.63	0.72	1.00				
49	0.62	0.66	0.64	0.64	0.69	0.67	0.68	0.69	0.70	1.00			
50	0.62	0.61	0.58	0.63	0.65	0.65	0.68	0.65	0.68	0.70	1.00		
51	0.60	0.63	0.63	0.63	0.65	0.65	0.68	0.70	0.66	0.71	0.73	1.00	
52	0.60	0.66	0.63	0.60	0.66	0.63	0.67	0.67	0.65	0.70	0.70	0.74	1.00

上の値を眺めると，どれだけ離れた週の間でも最低0.58の相関があり，さらに商品ごとに層別してもほとんど変化がない．このことは，これら9商品に関してあてはめたモデルから予測される売上量よりも定常的に多く売り上げる，あるいは逆に少ない売上しかない店の存在を示唆する．そこで，さらに店ごとに層別し，そこでの残差平均が−0.2以下の店を抜き出してみれば次の表6のようになる．ただし，Kは残差が−0.2以下であった週の回数である．また，前と同じようにNAは欠損値（Not Available）を表す．週と店で層別した平均残差の標準偏差は

表6 9商品の売上率の低い店

chain	pref	location	chain.type	残差平均	K	開店年
57	静岡	駅前	地域大	−0.5394	13	1975
42	大阪	住宅街	全国小	−0.3187	10	NA
32	静岡	駅前	全国大	−0.3145	9	1972
43	東京	駅前	全国小	−0.3092	10	1978
46	京都	住宅街	地域小	−0.2923	11	1979
54	東京	住宅街	地域大	−0.2825	9	1974
38	三重	住宅街	地域小	−0.2796	7	1978
54	埼玉	住宅街	地域大	−0.2786	12	1979
54	東京	住宅街	地域大	−0.2774	12	1981
29	神奈川	住宅街	地域小	−0.2721	8	1984
42	兵庫	商店街	全国小	−0.2694	8	1966
32	茨城	商店街	全国大	−0.2369	10	1968
52	神奈川	駅前	地域大	−0.2148	8	1980
22	埼玉	住宅街	地域小	−0.2114	7	1986
50	神奈川	商店街	全国大	−0.2026	4	1980

表7　9商品の売上率の高い店

chain	pref	location	chain.type	残差平均	K	開店年
57	兵庫	商店街	地域小	0.2039	6	1992
42	東京	住宅街	地域大	0.2135	8	1978
32	東京	駅前	全国小	0.2277	8	1972
43	茨城	住宅街	全国大	0.2385	8	NA
46	大阪	駅前	地域小	0.2535	10	1977
54	三重	住宅街	地域小	0.2595	11	1984
38	京都	住宅街	全国大	0.2689	11	1976
54	大阪	住宅街	全国大	0.2716	11	1968
54	静岡	郊外	地域大	0.2819	12	1979
29	神奈川	駅前	全国小	0.3014	12	1989
42	埼玉	住宅街	全国小	0.3019	12	1978
32	静岡	商店街	地域小	0.3080	12	1985
52	埼玉	住宅街	地域大	0.3125	13	1976
22	神奈川	駅前	地域大	0.3443	12	1990
50	東京	駅前	地域大	0.4574	13	1976
57	大阪	住宅街	全国大	0.4900	13	1982

0.1073であるので，表6に掲げた店は，実際，売上量がかなり少ない店であることになる．このことは，Kを眺めてみるとよりはっきりする．特に最初の店の売上量が極端に少なく，13週すべてで残差が-0.2以下である．その原因は，説明変量の値や，属するチェーン，面積，駐車場の台数などからは説明がつかないが，少なくとも最初の3店に関しては最後の列の（新装開店を含む）開店年から説明できる．つまり比較的古い店である．これ以外の店に関しては，店の新旧以外に，その地域の定住人口の減少，競合店の存在など，他の要因も調べる必要があろう．一方，残差平均が0.2以上の店を抜き出すと表7のようになる．この場合には，Kは残差が0.2以上である週の回数であり，最後の2店は，すべての週で0.2以上の残差をもつだけでなく，その平均も高い．つまり売上量が極端に多い．このような売上量の多い店に関しては次のような解釈ができるのではなかろうか．つまり，「駅前」以外の店で，比較的開店の古い店は7番目，8番目，13番目の店だけである．その中でももっとも古い8番目の店に関しては，実際に調べた結果，その地域最大の店であり，売上の多いのも当然と思われる店であった．もちろん，これは一つの解釈にすぎない．実際に各店舗に当たってその原因を探る必要がある．

ここでのモデル化は，立地条件，属するチェーンのタイプ，所在県以外の各店舗の特性は無視したモデル化であるので，説明しきれない各店舗の特徴が残差にあらわれていることになる．これは前節で注意した残差の裾が比較的重いことにも関係してくる．したがって，このモデルを出発点にして，さらにモデルを改良する必要のあることはいうまでもない．

6. ま と め

以上，解析したデータはきわめて身近なデータであり，しかも日々発生し，オンラインで収集されているデータである．数々の試行錯誤の結果到達したモデル(4.2)は比較的単純な線形モデルでありながら，かなりの説明力があり，各商品の価格や店舗の立地条件，属するチェーンのタイプ，所在県による売上量の変化をうまく説明している．さらに，商品ごとの週効果を導入することにより秋から冬へかけての季節性をもつ商品とその週効果をも明らかにすることができた．この意味では，最初に掲げた統計的代表性，データの継続性，結果の安定性のいずれの項目に関してもかなり合理的な説明ができたものと信ずる．ここでは，これらの変量の交

互作用については特に述べなかったが，実際には交互作用を入れたモデルも試した上でその効果があまりないため採用しなかった．

データからモデルを構築する作業はある意味で終りのない作業であり，どこかで割り切らざるを得ない．その意味では，最後に述べた残差が常に一定以上になるような店の特徴を新たな説明変量として導入する，あるいは大売出しなどのイベント情報を導入するなど，いくらでも上のモデルの改良の余地はある．いわば，ここで得られたモデルの残差こそがマーケティングのさまざまな側面をさらにデータから明らかにするための１つの出発点である．

また，ここでは行わなかった，ブランドの違いの売上量への影響，保存食品以外のデータの解析など，数々の興味深いマーケティング上の問題も残されている．理論的には一般化線形モデルへの拡張なども大いに興味深い．しかし，著者らは統計解析の応用可能性を広げ，その重要性を高める１つの現実的なステップとして以上のようなデータ解析の結果をこの段階で報告しておくことも十分意味のあることと信じここに報告する．なお，ここで解析に用いたデータは著者の１人が従来から提唱してきた「データとその記述の一体化」の一方式であるＤ＆Ｄ形式で自由に入手可能であるので，そのサポートソフトウェアと共に入手すれば，直ちに違った角度からの解析が開始できる．詳細については付属のマニュアルあるいはEJDA [2] を参照して頂きたいが，例をあげればＳでコマンド

summary(market. data)

を入力すればデータの概要が表示され，

attach(market. data)

の入力でＤ＆Ｄの大項目と，変量c. log. price, price, volume. acv (volume/acv) のデータおよび説明変量のデータフレーム factors にアクセス可能となる．Possible.analysis の内容に従えば，本論文でのモデル (4.1) と (4.2) のあてはめを再現することもできる．データ market. data とＤ＆Ｄサポートソフトウェアは stat. math. keio. ac. jp : /usr/pub/statlib/s. jpn より Anonymous FTP で market. data. d_d. z と d_d. tar. z として自由に入手できる．なお，statlib に関しては [4] を参照されたい．このデータに関し読者諸兄姉御自身によるさまざまな角度からのモデル化がなされることを期待し，また，それにもとづく御批判を仰ぎたい．

参考文献

[1] 「1992年度版日本スーパーマーケット名鑑」，1991，商業界．
[2] 「“データ解析の電子ジャーナル (EJDA)” の実働化」，1994，統計数理研究所共同研究リポート54．
[3] 片平秀貴 (1987)，「マーケティング・サイエンス」，東京大学出版会．
[4] 渋谷政昭，柴田里程 (1992)，「Sによるデータ解析」，共立出版．
[5] チェンバース，ヘイスティー編 (1994)，「Sと統計モデル」，柴田里程訳，共立出版．
[6] ベッカー，チェンバース，ウィルクス (1991)，「S言語Ⅰ，Ⅱ」，渋谷政昭，柴田里程訳，共立出版．
[7] 法政大学産業情報センター，小川孔輔編 (1993)，「POSとマーケティング戦略」，有斐閣．