重回帰分析における変数選択

芳賀敏郎

1981, 6, 26

No.

目　次

0	まえがき	1
1	変数選択の規準としての予測平方和	8
1.1	予測平方和の導出	8
1.2	数値例	15
1.3	予測平方和とその他の規準との関係	18
(1)	予測平方和の期待値	18
(2)	予測平均2乗誤差	19
(3)	自由度調整済重相関係数と自由度2重調整済相関係数	19
(4)	F統計量による打切規準	21
1.4	多項式回帰による検討	23
(1)	例1	23
(2)	例2	28
2	予測平方和(PSS)と残差平方和(RSS)の関係	41
2.1	説明変数がない場合	42
2.2	説明変数が1個で正規分布の場合	42
(1)	PSS と RSS の分布	43
(2)	回帰の有意性検定	44
(3)	検出力	47
2.3	説明変数が1個で非正規分布の場合	49
(1)	比 PSS/RSS の分布	49
(2)	第一種の誤りの確率	52
2.4	説明変数が1個で，異常値の混入した場合	53
2.5	説明変数が2個の場合	57
(1)	モデルと変数選択の規準	57
(2)	各変数が正規分布の場合	59
(3)	各変数が非正規分布の場合	61
(4)	各変数に異常値の混入した場合	64

3　最良回帰式を探すアルゴリズム　67
3.1　総当り法　68
3.2　改良した逐次法　77
3.3　RSS で予備選択をする方法　82
3.4　3つの方法の比較　83
3.5　対話による変数選択　95

4　数量化理論I類への適用　102
4.1　アイテム選択の方法　102
(1)　総当り法　105
(2)　逐次法　107
4.2　アイテム間に交互作用のある場合　110

あとがき　112

付録　コンピュータプログラム　114

参考文献　122

0 まえがき

化学工業や組立工業における製品品質を表わす特性値は，原材料や中間製品について測られる特性値と製造工程でとられる種々の計測値との関数として予測できる筈である。この関数は通常，重回帰式として表わされ，最終製品の品質特性を目的変数 y，原料特性や工程の条件を表わす計測値を説明変数 $x_1, x_2, \cdots, x_p$ と書くと，y の $x_1, x_2, \cdots, x_p$ に対する重回帰モデルは

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad (1)$$

となる。ここで β_0 を定数項，$\beta_1, \beta_2, \cdots, \beta_p$ を偏回帰係数と呼び，ε は $x_1, x_2, \cdots, x_p$ では説明しきれない y の変動部分を表わす。ε は平均 0，分散 σ^2（一定）の正規分布に従うと仮定することが多い。

ところで，式(1)のようなモデルが確定的に成立するという保証は現実にはほとんどあり得ない。実験室で行われる化学反応や物理実験に対しては，その構造を表わす理論式が定められる場合もある。しかし生産現場でとられるデータについてそのような理論モデルを設定できることはまずあり得ない。とすれば，ここで求めようとしているのは，式(1)のような形の経験式である。経験式であれば，（x_1, x_2, x_3）の代りに，これらと相関の高い（x_4, x_5）を用いて重回帰式を作ることもできるから，式(1)は一義的に定まるのではなく，適当なものを便宜的に定めればよいことになる。そこで，現在入手できる候補変数を全部取り上げてその数を p とすると，p は30にも50に

も達するのが普通であるが，その中から適当な q 個（$q<p$）の変数を選べばよいことになる。こうして，単純ではあるが，予測に十分役立つ重回帰式を求めようとする試みは古くから多くの実務家によって遂行されてきた。

式（1）は $x_1, x_2, \cdots, x_p$ の1次式であるが，実際の計測値を $u_1, u_2, \cdots$ と表わすとき，$x_1=u_1$，$x_2=u_2$，$x_3=u_3$，$x_4=u_1^2$，$x_5=u_1u_2$，$x_6=\log u_3$，$\cdots$ というように，2乗や積，あるいは対数変換などを施した関数をそれぞれ説明変数 x_j とすることもできる。それゆえ，説明変数 x_j は計測値 u_k（$k=1,2,\cdots$）の任意の関数であり，式（1）は u_k については非線形の項を含んでもよいが，未知パラメータ $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ については線形であるということが重回帰分析適用の条件となる。このような非線形の項も候補変数に入れると，その総数 p は益々大きくなり，100を越えることもある。それゆえ，その適当な部分集合（変数の数 $q<p$）を選ぶ方法が「変数選択」または「モデル選択」と呼ばれ，古くから検討されてきた。

式（1）の未知パラメータ $\{\beta_j\}$ $(j=0,1,2,\cdots,p)$ は $\{x_1, x_2, \cdots, x_p; y\}$ についてのn組のデータから最小二乗法（ε の分布が正規分布のときは最尤法に一致する）によって推定し，$\{b_j\}$ $(j=0,1,2,\cdots,p)$ とするのが普通である。このとき，y を推定する重回帰式は

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p \qquad (2)$$

と書け，この回帰式のあてはめの良さを評価する規準としては，次の2つが

用いられる。

1° 残差平方和 Residual Sum of Squares (RSS)

観測値 y_i と回帰推定値 $\hat{y}_i$ ($i=1,2,\cdots,n$) との差の2乗和で

$$RSS = \sum (y_i - \hat{y}_i)^2 \qquad (3)$$

と書ける。

2° 重相関係数 Multiple Correlation Coefficient (R)

観測値 y_i と回帰推定値 $\hat{y}_i$ (これ自身は式(2)に示すように, $x_1, x_2, \cdots, x_p$ の1次式) との相関

$$R = r(y, \hat{y}) = r(y, [x_1, x_2, \cdots, x_p]) \qquad (4)$$

この1°, 2°の間には明確な関数関係があり, y の全変動を

$$S_T = \sum (y_i - \bar{y})^2$$

$$\bar{y} = \sum y_i / n$$

で表わすと,

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{RSS}{S_T} \qquad (5)$$

が成立する。S_T は変数選択に無関係な定数であるから, RSS → 小 なら R → 大 という関係がある。

今からおよそ10年以上も前であるが, このような重回帰式の計算にコンピュータが用いられはじめた頃は, p が10～30位の重回帰式が良く用いられ, R^2 が0.9以上にもなって, データに非常に良く適合すると考えられ

た。鉄鋼業におけるプロセスコントロールでは，pが30位の重回帰式は数行にも十数行にもわたって表記され，$R^2 > 0.9$で非常に良い予測式が得られたと報告された。しかし，それが良く適合したのは手許にある1組のデータについてであって，これを用いて将来の値を予測すれば必ずしも良い予測値を与えないことがわかってきた。このことから，多数の変数を含む複雑なモデルは，予測式を推定するために使ったデータを説明するのには役立つが，その再現性は必ずしも保証されず，かえって単純なモデルほど再現性があって，将来の予測に用いても十分役に立つことが認識されるようになってきた。そこで，変数選択の問題の重要性が益々強調されるようになった。

この重回帰分析における変数選択に係る問題は次の2つの観点から論ずることができる。

1° 重回帰式の良さを評価する規準

2° ある規準のもとで，最良の重回帰式を探すアルゴリズム

本論文はこの1°に対して「予測平方和PSS」および「自由度二重調整済重相関係数 R^{**}」という新しい規準を提案し，その統計的性質を他の規準との比較において詳細に論ずるとともに，2°に関しても種々のアルゴリズムを工夫した結果をまとめたものである。

重回帰式の良さを評価する規準は，上述のRSS（またはR^2）が通常用いられているが，この規準では変数の数pを大きくするとRSSは単調に減少し，（Rは単調に増大し），$p = n-1$で$RSS = 0$，$R = 1$となる。

それゆえ，「変数の数 p は多ければ多いほど良い」という結論になる。そこで，変数の数を q から $q+1$ にしたときの RSS の減少分（R^2 の増大分）があらかじめ定められた値より小さい変数は増加する価値がないと考える。そのような変数を増すことに対するペナルティーを与えて，変数増加を打切る stopping rule を導入する。このようにして説明変数の数が無制限に増加するのを防ぐのであるが，この打切り規準が極めて恣意的である。そのようなペナルティー を与える規準として，Mallow の Cp，赤池の AIC，自由度調整済重相関係数などが提案されている。本論文では，これらと PSS の期待値との関係を論ずるとともに，自由度調整済重相関係数という規準を新しく提案する。PSS は，これらのように，RSS に罰金を課する規準とは異なり，RSS とは全く異なる立場から最良の部分集合を選ぶものである。それは一種の Jackknife 法と見られる手法で，手許の n 組のデータの中で予測の模擬をするのである。すなわち，No. 1 の観測値 y_1 を予測するときは，No. 1 を除いた残りの $(n-1)$ 組のデータから重回帰式を計算し，これに No. 1 の説明変数の値 $(x_{11}, x_{12}, \cdots, x_{1p})$ を代入して予測値 $\hat{y}_1^*$ を求める。これは式 (2) から求めた $\hat{y}_1$ とは異なる。次に y_2 を予測するときには，No. 2 のデータを除いて，$(n-1)$ 組のデータから改めて重回帰式を計算し，これに $(x_{21}, x_{22}, \cdots, x_{2p})$ を代入して，$\hat{y}_2^*$ を求める。以下このようにして，n 回重回帰式を計算して，これから予測値 $\hat{y}_i^*$ と観測値 y_i の差の2乗和

$$PSS = \Sigma (y_i - \hat{y}_i^*)^2 \qquad (6)$$

を求めるのである。この PSS 規準の重要な特長は，変数の数 p を 1, 2, … と大きくして行くと，はじめは減少するが，やがて増大に転じ，極小値が存在することである。それゆえ，PSS 規準を用いれば，RSS 規準と異なり，何の打ち切り規準も必要としない。

この PSS が何らかの意味で最適な規準であることを証明する数学的根拠は未だ明らかにされていない。この考え方は Allen (1971) によって示されたが，式 (6) のままでは計算量が膨大になるので，これを簡単に計算する方法を，芳賀，竹内，奥野 (1976) が与えた。(同じ結果は途中の証明ぬきで Hocking (1972) にも示されていることが後に判った)。本論文ではこの PSS の性質をいろいろのモデルのもとで検討し，他の規準による選択との比較を行なった。

2° のアルゴリズムについては，現在 Stepwise method (逐次変数選択法) が標準的な方法として BMDP などに採用されている。これは変数の増加・減少の各ステップで，最良の回帰式を 1 つずつ求めて行くのに対して，筆者は，候補の重回帰式の中の best k 個を選ぶ方法を提案する。この方法によって，最良の重回帰式を見落す危険を小さくするとともに，次善，3 善の代替重回帰式を示し，技術的判断を加える余地を残すことができる。また，コンピュータと対話しながら，技術的判断を加えつつ適切な重回帰式を探索する方法も提案する。これらを効率よく行なうアルゴリズムを開発した。

最後に，説明変数がカテゴライズされた（分類された）変数を用いる重回帰分析と考えられる，いわゆる数量化Ⅰ類について，そのアイテムの選択にPSSと同じ考え方を導入し，その具体的計算法を示した。

1. 変数選択の規準としての予測平方和

1.1 予測平方和の導出

n個の観測値ベクトル $\underset{\sim}{y}$ に対して，次の重回帰モデルを仮定する：

$$\underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times(p+1)}{\underset{\sim}{X}}\ \underset{(p+1)\times 1}{\underset{\sim}{\beta}} + \underset{n+1}{\underset{\sim}{\varepsilon}}, \qquad (1.1)$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

$$\mathrm{rank}\, X = p+1 \leqq n$$

成分ごとには

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i=1,2,\cdots,n \qquad (1.2)$$

と書く。

ここで，x_j $(j=1,2,\cdots,p)$ は説明変数，$x_0 \equiv 1$ で，$(p+1)$個の未知パラメータのうち $\beta_1, \beta_2, \cdots, \beta_p$ は偏回帰係数，β_0 は定数項と呼ばれる。

説明変数のとる値を並べた行列 $\underset{\sim}{X}$ $(n\times(p+1))$，または，その1列を除いた $X^*(n\times p)$ を以下"計画行列"と呼ぶ。

このn組のデータから最小二乗法によって未知パラメータ $\underset{\sim}{\beta}$ の推定値 $\underset{\sim}{b}$ を求める。すなわち

$$\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}) \Rightarrow \min \qquad (1.4)$$

より，X の rank は $p+1$ であるから $X'X$ の逆行列は存在し，

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y} \qquad (1.5)$$

を得る。これから $\underset{\sim}{y}$ の推定値

$$\hat{\underset{\sim}{y}} = \underset{\sim}{X}\,\underset{\sim}{b}$$

を得る。$\underset{\sim}{X}$ の第 i 行を

$$\underset{\sim}{x}_i' = (x_{i0}, x_{i1}, x_{i2}, \cdots, x_{ip})$$

と書くと，y_i の推定値 $\hat{y}_i$ は

$$\hat{y}_i = \underset{\sim}{x}_i'\,\underset{\sim}{b} \qquad (1.6)$$

で，残差平方和 RSS (Residual Sum of Squares) は

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= (\underset{\sim}{y} - \hat{\underset{\sim}{y}})'(\underset{\sim}{y} - \hat{\underset{\sim}{y}}) \qquad (1.7)$$

で定義される。

y と $(x_1, x_2, \cdots, x_p)$ との重相関係数 R およびその2乗である寄与率 R^2 は

$$R^2 = 1 - \frac{RSS}{S_T} \qquad (1.8)$$

で与えられる。S_T は y の偏差平方和 $(\underset{\sim}{y} - \bar{\underset{\sim}{y}})'(\underset{\sim}{y} - \bar{\underset{\sim}{y}})$ である。

p 個の説明変数の中から，q 個の説明変数を選んで作った回帰式の残差平方和を RSS_q とする。q 個の説明変数にさらに1個説明変数を追加して得られる回帰式の RSS_{q+1} と RSS_q との間には常に

$$RSS_{q+1} \leq RSS_q \qquad (1.9)$$

の関係がある。

本来 $x_{q+1} \sim x_p$ は y を説明するのに実質的には役に立たないと考えられる変数であっても，これらを回帰式に含めると，残差平方和は減少し，寄与率は増加する。

これから，次のことが言える。

i) 残差平方和を小さくする（または寄与率を大きくする）という規準で変数を選択すると，無意味な説明変数が無制限にとり込まれる。

ii) 寄与率は，真の寄与率より大きい方に偏りを持ち，過大に評価される。

このような結果になることを避けるために，従来いくつかの方法が考えられてきた。

その一つに"相互検証法（cross validation）"がある。それは次のような方法である。

i) n組のサンプルをテンダムに大きさ $n/2$ の2群に分割する。

ii) 一方の群（解析群）のデータを使って回帰式を推定する。

iii) 上に得た回帰式を予測式として他方の群（評価群）のデータの予測値を求め，その予測の良さを評価する。

iv) 解析群と評価群を入れかえて同じ操作を繰り返す。

この方法によれば予測の精度も公正に評価できる。しかし，$n/2$ 組のデータだけで解析するのは効率が良くない。

そこで，次のような手法を考える。

i) n組のサンプルを，n−1組のデータからなる解析群と，1組のデータだけの評価群に分割する。

評価群のデータの組番号を i とする。

ii) 解析群の（n−1）組のデータから偏回帰係数と定数項を推定する。

推定値を $\underset{\sim}{b}_i^*$, b_0^* とする。

iii) この係数を使って，y_i の予測値 $\hat{y}_i^*$ を求める。

iv) i)～iii) の手順を $i=1,2,\cdots,n$ と変えて反復し，y_i と予測値 $\hat{y}_i^*$ との差の2乗和を求める。

こうして得られた2乗和を，"予測平方和"（Prediction Sum of Squares）[注1] と呼ぶことにする。

ここに導いた $\underset{\sim}{b}_i^*$ と PSS の関係を式に表わすと次のようになる。

$$\underset{\sim}{b}_i^* = (\underset{\sim}{X}'\underset{\sim}{X} - \underset{\sim}{x}_i\underset{\sim}{x}_i')^{-1}(\underset{\sim}{X}'\underset{\sim}{y} - \underset{\sim}{x}_i y_i) \quad [注2] \quad (1.10)$$

$$\hat{y}_i^* = \underset{\sim}{x}_i'\,\underset{\sim}{b}_i^* \quad (1.11)$$

$$PSS = \sum_{i=1}^{n}(y_i - \hat{y}_i^*)^2 \quad (1.12)$$

本研究の初期の段階では式（1.10）～（1.12）によって PSS を計算していたが，np^3 のオーダーの計算量を必要とし，ミニコンには大きな負担であった。その後，予測残差 $y_i - \hat{y}_i^*$ と回帰残差 $y_i - \hat{y}_i$ との間には次の関係のあることを導き，計算量を np^2 のオーダーに下げることができた。

（芳賀，竹内，奥野（1976））

$$y_i - \hat{y}_i^* = (y_i - \hat{y}_i)/(1-c_i) \qquad (注3) \qquad (1.13)$$

$$c_i = \underset{\sim}{x}_i' (X'X)^{-1} \underset{\sim}{x}_i$$

$$= \frac{1}{n} + \frac{D_i^2}{n-1} \qquad (注4) \qquad (1.14)$$

ここに，D_i は $x_1, x_2, \cdots, x_p$ の p 次元空間における $\underset{\sim}{x}_i$ と $\overline{\underset{\sim}{x}}$ との Mahalanobis (汎) 距離と呼ばれるものである。また，c_i は回帰推定値 $\hat{y}_i$ の分散を与える式

$$V[\hat{y}_i] = c_i \sigma^2 \qquad (1.15)$$

の係数に相当している。

注1) この考え方は，判別関数にも適用され，The leaving-one-out method と呼ばれている。(P. A. Lachenbruch (1975))

注2) $X'X - \underset{\sim}{x}_i \underset{\sim}{x}_i'$ non-singular でなければならない。したがって，説明変数の数 p は $n-2$ を越えることはできない。

注3) 式 (1.13) の証明

n 組のデータによる正規方程式

$$X'X\underset{\sim}{b} = X'\underset{\sim}{y}$$

を，i 番目を除いた $n-1$ 組のデータに対する正規方程式

$$(X'X - \underset{\sim}{x}_i \underset{\sim}{x}_i')\underset{\sim}{b}_i^* = X'\underset{\sim}{y} - \underset{\sim}{x}_i y_i$$

に代入する。

$$(X'X - \underset{\sim}{x}_i \underset{\sim}{x}_i')\underset{\sim}{b}_i^* = X'X\underset{\sim}{b} - \underset{\sim}{x}_i y_i$$

整理して

$$X'X(b_i^* - b) = x_i(x_i' b_i^* - y_i)$$
$$= x_i(\hat{y}_i^* - y_i)$$

両辺の左から $x_i'(X'X)^{-1}$ を乗じて

$$x_i'(b_i^* - b) = x_i'(X'X)^{-1}x_i(\hat{y}_i^* - y_i)$$
$$\hat{y}_i^* - \hat{y}_i = c_i(\hat{y}_i^* - y_i)$$
$$y_i - \hat{y}_i^* = (y_i - \hat{y}_i)/(1 - c_i)$$

(証明終)

式(1.13)は，n組のサンプルをn−1組の解析群と1組の評価群に分けたときのものである。これをn−m個の解析群とm個の評価群に拡張することができる。以下m=2の場合について 式(1.13)に対応する式を導く。

i=1,2 を評価群とする。評価群の計画行列を X_0（大きさ 2×(p+1)），対応する目的変数の値 (y_1, y_2) を y_0 とする。

$$(X'X - X_0'X_0)b^* = X'y - X_0y_0$$
$$= X'Xb - X_0'y_0$$
$$X'X(b^* - b) = X_0'(X_0b^* - y_0)$$
$$= X_0'(\hat{y}_0^* - y_0)$$
$$X_0(X'X)^{-1}(X'X)(b^* - b)$$
$$= X_0(X'X)^{-1}X_0'(\hat{y}_0^* - y_0)$$
$$X_0(b^* - b) = C(\hat{y}_0^* - y_0)$$
$$\hat{y}_0^* - \hat{y}_0 = C(\hat{y}_0^* - y_0)$$
$$y_0 - \hat{y}_0 = (I - C)(y_0 - \hat{y}_0^*)$$
$$y_0 - \hat{y}_0^* = (I - C)^{-1}(y_0 - \hat{y}_0)$$

要素を書くと,

$$\begin{pmatrix} y_1 - \hat{y}_1^* \\ y_2 - \hat{y}_2^* \end{pmatrix} = \begin{pmatrix} 1-c_{11} & -c_{12} \\ -c_{12} & 1-c_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \end{pmatrix}$$

となる。

注4) 式(1.14)の誘導 ($x_0 \equiv 1$ を除いた $\underset{\sim}{x}, \underset{\sim}{X}$ を $\underset{\sim}{x}^*, X^*$ とする).

$$\underset{\sim}{X}'\underset{\sim}{X} = \begin{pmatrix} n & \Sigma \underset{\sim}{x}^{*\prime} \\ \Sigma \underset{\sim}{x}^* & \underset{\sim}{X}^{*\prime}\underset{\sim}{X}^* \end{pmatrix}$$

$\underset{\sim}{X}'\underset{\sim}{X}$ の1行1列の要素(n)を軸にして掃き出し，第1列に逆行列の第1列を入れる。

$$\begin{pmatrix} 1/n & \overline{\underset{\sim}{x}}^* \\ -\overline{\underset{\sim}{x}}^* & \underset{\sim}{S} \end{pmatrix}$$

ここに $\underset{\sim}{S}$ は平方和積和行列で次のように書くことができる.

$$S = \underset{\sim}{X}^{*\prime}\underset{\sim}{X}^* - \Sigma \underset{\sim}{x}^* \cdot \Sigma \underset{\sim}{x}^{*\prime}/n$$

$$= \underset{\sim}{X}^{*\prime}(\underset{\sim}{I} - \underset{\sim}{U})\underset{\sim}{X}^*$$

ただし，I は単位行列，U は要素がすべて $1/n$ の行列である。

さらに，平方和・積和行列 $\underset{\sim}{S}$ の対角要素について掃き出し，逆行列で置きかえると，$(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ が次のように求められる。

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \begin{pmatrix} 1/n + \overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1}\overline{\underset{\sim}{x}}^* & -\overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1} \\ -\underset{\sim}{S}^{-1}\overline{\underset{\sim}{x}}^* & \underset{\sim}{S}^{-1} \end{pmatrix}$$

これから，(1.14')式は

$$(1, \underset{\sim}{x}_i^{*\prime})(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\begin{pmatrix} 1 \\ \underset{\sim}{x}_i^* \end{pmatrix}$$

$$= (1, \underset{\sim}{x}_i^{*\prime}) \begin{pmatrix} 1/n + \overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1}\overline{\underset{\sim}{x}}^* - \overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1}\underset{\sim}{x}_i^* \\ -\underset{\sim}{S}^{-1}\overline{\underset{\sim}{x}}^* + \underset{\sim}{S}^{-1}\underset{\sim}{x}_i^* \end{pmatrix}$$

$$= \frac{1}{n} + \overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1}\overline{\underset{\sim}{x}}^* - 2\overline{\underset{\sim}{x}}^{*\prime}\underset{\sim}{S}^{-1}\underset{\sim}{x}_i^* + \underset{\sim}{x}_i^{*\prime}\underset{\sim}{S}^{-1}\underset{\sim}{x}_i^*$$

$$= \frac{1}{n} + (\underset{\sim}{x}_i^* - \overline{\underset{\sim}{x}}^*)'\underset{\sim}{S}^{-1}(\underset{\sim}{x}_i^* - \overline{\underset{\sim}{x}}^*)$$

として導かれる。

1.2 数値例

このPSSは，その導かれた根拠から期待されるように，回帰式に説明変数を追加して行っても，PSSが減少するとは限らないという性質をもっている。簡単な例によって示そう。

表1.1は$n=30$, $p=5$の説明変数およびyのデータと，その基本統計量を示す。

表1.1 データと基本統計量

DATA

	1	2	3	4	5	y
1	8.2500	9.8900	10.8800	8.7300	8.8700	8.3900
2	10.7700	11.4700	11.0200	12.3300	11.8800	10.9300
3	10.9400	9.8200	10.2700	10.9000	10.3600	11.4600
4	9.3400	9.0000	8.8600	8.2800	8.0000	8.4400
5	10.4600	9.9600	8.6200	10.5400	9.8800	10.1400
6	11.5400	10.1700	9.8300	11.9900	11.3700	11.2700
7	9.4200	10.2300	11.4400	10.2900	9.5800	9.6700
8	9.7800	9.9300	10.7400	10.0100	9.8500	10.3900
9	11.3500	10.4700	12.2900	11.8900	10.9200	12.4900
10	9.4100	9.4700	10.3600	9.3600	8.9600	9.9300
11	10.8300	8.1700	8.1400	8.9700	7.5700	9.4900
12	9.6100	10.4600	10.2300	10.1200	10.4100	9.8800
13	9.9500	10.5300	10.9400	10.1300	10.5600	11.5800
14	11.3200	10.3900	11.9800	11.9800	11.5000	12.6800
15	10.3400	9.7100	9.8400	10.3200	9.1800	10.4100
16	8.3100	9.5300	11.4100	9.4900	8.6500	8.2900
17	8.7700	11.5800	10.8600	9.6000	9.9900	10.2200
18	10.2600	8.0500	8.7800	8.8300	8.9600	8.6400
19	7.0500	8.6900	8.9900	7.0400	8.2800	5.9300
20	8.6100	9.3200	8.5600	7.8200	8.5500	7.7800
21	10.2600	8.0300	7.9200	8.4900	8.1400	7.7800
22	8.3400	9.4200	9.6300	8.3500	8.7200	7.3100
23	9.6900	9.6000	10.2600	8.9800	10.1000	8.4800
24	8.6600	8.4400	8.4200	7.8100	7.8300	7.4900
25	8.0200	10.4100	10.3100	8.2400	8.7300	8.0000
26	9.5200	10.0100	9.7000	8.2000	9.4300	10.1500
27	8.8100	8.8500	8.9000	7.5200	8.0100	6.2600
28	9.3300	9.0600	9.0100	8.8800	8.1900	8.1900
29	9.5600	8.2100	8.4100	7.8900	8.0000	8.7000
30	9.3600	9.5500	9.1500	8.3800	8.7000	9.6400
MEAN	9.5953	9.6140	9.8840	9.3786	9.3390	9.3337
SIGMA	1.0879	0.9230	1.1364	1.4340	1.1761	1.6930
MINIMUM	7.0500	8.0300	7.9200	7.0400	7.5700	5.9300
MAXIMUM	11.5400	11.5800	12.2900	12.3300	11.8800	12.6800

R MATRIX

	1	2	3	4	5	6
1	1.0000	0.1711	0.2043	0.7687	0.5599	0.7871
2	0.1711	1.0000	0.7515	0.6491	0.7898	0.6149
3	0.2043	0.7515	1.0000	0.6602	0.7187	0.6190
4	0.7687	0.6491	0.6602	1.0000	0.8856	0.8691
5	0.5599	0.7898	0.7187	0.8856	1.0000	0.7840
6	0.7871	0.6149	0.6190	0.8691	0.7840	1.0000

5個の説明変数の組合せでできる $2^5-1=31$ 個の回帰式について計算したRSSとPSSを表1.2に示す。

表1.2　2^5-1 個の回帰式の比較

q	x_1	x_2	x_3	x_4	x_5	RSS	R^2	R^{*2}	R^{**2}	PSS
1	1					31.63	0.620	0.606	0.593	35.58
		2				51.69	0.378	0.356	0.335	59.36
			3			51.27	0.383	0.361	0.340	58.65
				4		*20.34	*0.755	*0.747	*0.738	*23.97
					5	32.03	0.615	0.601	0.588	37.60
2	1	2				*11.88	*0.857	*0.847	*0.837	*14.74
	1		3			13.41	0.839	0.827	0.816	15.98
	1			4		17.46	0.790	0.774	0.760	21.72
	1				5	17.36	0.791	0.776	0.761	21.61
		2	3			46.99	0.435	0.393	0.354	58.11
		2		4		19.96	0.760	0.742	0.725	25.20
		2			5	32.03	0.615	0.586	0.559	38.73
			3	4		20.03	0.759	0.741	0.724	24.75
			3		5	31.50	0.621	0.593	0.567	38.52
				4	5	20.26	0.756	0.738	0.721	26.42
3	1	2	3			*9.79	*0.882	*0.869	**0.856	**12.31
	1	2		4		11.84	0.858	0.841	0.826	15.85
	1	2			5	11.88	0.841	0.841	0.825	15.44
	1		3	4		13.10	0.842	0.824	0.807	17.13
	1		3		5	12.61	0.848	0.831	0.814	15.65
	1			4	5	16.33	0.804	0.781	0.760	22.87
		2	3	4		19.90	0.764	0.733	0.707	26.88
		2	3		5	31.33	0.623	0.580	0.539	40.45
		2		4	5	19.95	0.760	0.732	0.706	26.92
			3	4	5	19.90	0.761	0.733	0.707	26.88
4	1	2	3	4		*9.41	*0.887	**0.869	*0.852	12.91
	1	2	3		5	9.55	0.885	0.867	0.850	*12.68
	1	2		4	5	11.82	0.858	0.835	0.814	16.81
	1		3	4	5	12.61	0.848	0.824	0.801	18.49
		2	3	4	5	19.89	0.761	0.723	0.687	28.69
5	1	2	3	4	5	**9.36	**0.887	*0.864	*0.842	*14.18

R^*は自由度調整，R^{**}は自由度2重調整相関係数で，その定義式は（1.23）と（1.24）に示す。

ここでは，取り入れた変数の数 q（$=1,2,\cdots,5$）ごとに，その中で最小のRSSとPSSには*印が，全体として最小の値には**印がつけられて

いる。これから，次のことがわかる。

i) 当然のことながら，変数の数 q を大きくすると RSS は単調減少し，$q=p=5$ で最小になる。寄与率 R^2 は単調に増加する。

ii) PSS は $q=3$ になるまでは単調減少し，3 で最小となるが，それ以上説明変数を追加するとかえって大きくなる。

iii) $q=4$ で最良の回帰式は RSS 規準では x_1, x_2, x_3, x_4 であるが，PSS 規準では x_1, x_2, x_3, x_5 である というように，q を固定したとき RSS と PSS を最小にする変数の組合せは必ずしも一致しない。

この iii) のような事態の起こる原因については，本論文で更にくわしく検討する。

1.3 予測平方和とその他の規準との関係

(1) 予測平方和の期待値

式(1・1)の回帰モデル

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$
$$\varepsilon_i \in N(0, \sigma^2)$$

に対して,

$$E[RSS] = E\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]$$
$$= \sum_{i=1}^{n} (1 - c_i)\sigma^2$$
$$= (n - p - 1)\sigma^2 \qquad (1.16)$$

であることはよく知られている。

PSSの期待値は次のように評価される。

$$E[PSS] = E\left[\sum_{i=1}^{n} (y_i - \hat{y}_i^*)^2\right]$$
$$= E\left[\sum \{(y_i - \hat{y}_i)/(1 - c_i)\}^2\right]$$
$$= \sum \sigma^2/(1 - c_i) \qquad (1.17)$$

コーシー・シュバルツの不等式

$$\sum_{i=1}^{n} \frac{1}{1 - c_i} \geq \frac{n^2}{\sum (1 - c_i)} = \frac{n^2}{n - p - 1} \qquad (1.18)$$

を用いると,

$$E[PSS] \geq \frac{n^2}{n - p - 1}\sigma^2$$
$$> (n + p + 1)\sigma^2 > (n - p - 1)\sigma^2 \qquad (1.19)$$

となり, PSSの期待値の下限はE[RSS]より大きいことがわかる。

(2) 予測平均2乗誤差 (MSEP, Mean Squares Error of Prediction)

説明変数が $x_i'(x_{i1}, x_{i2}, \cdots, x_{ip})$ をとるとき，将来得られるであろう測定値を y_i^0 とする（β の推定に用いた y_i とは異なる測定値）。$\hat{y}_i$ をこの y_i^0 の予測値と考えるとき，その予測平均2乗誤差（分散）は

$$V[y_i^0 - \hat{y}_i] = (1 + c_i)\sigma^2 \qquad (1.20)$$

となる。[y_i^0 と $\hat{y}_i$ は独立であることと，式(1.14)による]

この予測平均2乗誤差を n 個の x_i について加え合せたものを TMSEP (Total MSEP) とすると

$$TMSEP = \sum_{i=1}^{n}(1 + c_i)\sigma^2 = (n + p + 1)\sigma^2 \qquad (1.21)$$

を得る。これと，上に求めた RSS, PSS の期待値との間には

$$E[PSS] > TMSEP > E[RSS] \qquad (1.22)$$

の関係が成立する。

これから，TMSEP は PSS の期待値の下限を与えると言ってもよい。

(3) 自由度調整重相関係数 R^* と2重調整重相関係数 R^{**}

重相関係数の2乗 R^2 は，回帰による平方和 S_R を使って次のように表わされる。

$$R^2 = 1 - \frac{RSS}{S_T} = \frac{S_T - RSS}{S_T} = \frac{S_R}{S_T}$$

S_R の期待値は，$p\sigma^2 + \sum_{j=1}^{p}\sum_{j'=1}^{p}\beta_j\beta_{j'}S_{jj'}$ であって，その中に誤差成分 $p\sigma^2$ が含まれ，R^2 は真の寄与率に対して平均的に過大評

価となる。R^2から誤差成分を除くために，その推定値である誤差分散 $RSS/(n-p-1)$ の p 倍を引く。

$$R^{*2} = \frac{S_T - RSS - p(RSS/(n-p-1))}{S_T}$$
$$= 1 - \frac{RSS/(n-p-1)}{S_T/(n-1)}$$
$$= 1 - Ve/V_T$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2) \qquad (1.23)$$

としたR^*は"自由度調整相関係数"と呼ばれている。ここに，Ve，V_Tは $q=p$ および $q=0$ としたときの分散の不偏推定量である。

R^2を定義する式（1.8）の分子と分母を，説明変数の数 p を p と 0 としたときのTMSEPの不偏推定量で置きかえると，

$$R^{**2} = 1 - \frac{(n+p+1)Ve}{(n+1)V_T}$$
$$= 1 - \frac{(n+p+1)}{(n+1)}\frac{(n-1)}{(n-p-1)}(1-R^2) \qquad (1.24)$$

となる。R^{**}は"自由度2重調整重相関係数と呼ばれる。（T. Okuno, K. Takeuchi & T. Haga (1977)）。明らかに $R > R^* > R^{**}$ で，R^{*2} と R^{**2} は変数を増しても大きくなるとは限らないし，負の値をとることすらある。

表1・2には，各回帰式について，R^2, R^{*2}, R^{**2} が記されている。

R^2 は RSS と同様，$q=5$ のとき最大となっている。

R^{*2} は $q=4$ のとき最大となっている。

R^{**2} は PSS と同じく，$q=3$ のとき最大となっている。

(4) F統計量による打切り規準

変数選択のアルゴリズムにおいては，すでに取り入れられている $q-1$ 個の変数に，さらに1変数を加えて q 変数とするとき，または，q 変数の中から1変数を除いて $q-1$ 変数にするとき，その取捨は，次のF統計量によって判断されるのがふつうである。

$$F = \frac{RSS_{q-1} - RSS_q}{RSS_q/(n-q-1)} = \frac{R_q^2 - R_{q-1}^2}{(1-R_q^2)/(n-q-1)} \tag{1.25}$$

$R_q^{*2} > R_{q-1}^{*2}$ のとき q 番目の変数を追加するという規準は，

$$R_q^{*2} - R_{q-1}^{*2} = -\frac{n-1}{n-q-1}(1-R_q^2) + \frac{n-1}{n-q}(1-R_{q-1}^2)$$

$$= \frac{n-1}{n-q}\left[(R_q^2 - R_{q-1}^2) - \frac{1-R_q^2}{n-q-1}\right] > 0$$

$$\frac{R_q^2 - R_{q-1}^2}{(1-R_q)/(n-q-1)} - 1 > 0 \tag{1.26}$$

式(1.25)を利用して

$$F > 1 \tag{1.27}$$

という規準と同等であることがわかる。

また，$R_q^{**2} > R_{q-1}^{**2}$ を変数選択の基準とすることは，

$$R_q^{**2} - R_{q-1}^{**2} = -\frac{(n+q+1)(n-1)}{(n+1)(n-q-1)}(1-R_q^2) + \frac{(n+q)(n-1)}{(n+1)(n-q)}(1-R_{q-1}^2)$$

$$= \frac{(n-1)(n+q)}{(n+1)(n-q)}\left[(R_q^2 - R_{q-1}^2) - \frac{2n}{n+q}\frac{(1-R_q)^2}{n-q-1}\right] > 0 \tag{1.28}$$

から

$$F > \frac{2n}{n+q} \fallingdotseq 2 \quad (n \gg q) \tag{1.29}$$

と同等となる。

R^{**} は TMSEP の不偏推定値から導かれたものであり，TMSEP は PSS の期待値の下限になっている。これから，PSS による変数選択は，平均的に $F > 2$ の規準より幾分きびしいものと予想される。（§1.4 参照）

1.4 多項式回帰による検討

PSSとRSSとの関係をさらに詳しく調べるために，簡単な多項式モデルについて PSS，RSS等を計算した。

(1) 例1.

次のモデルを考える。

$$y_i = \beta_1 P_{i1} + \beta_2 P_{i2} + \beta_3 P_{i3} + \beta_4 P_{i4} \quad (i=1,2,\cdots,5) \qquad (1.30)$$

ここに P_{ij} は x の j 次の直交多項式 $\left(P_{ij} = \sum_{k=0}^{j} a_{jk} x^k\right)$ である。

表1.3は P_{ij} の値を示したものである。

表1.3 直交多項式

i	P_1	P_2	P_3	P_4
1	-2	2	-1	1
2	-1	-1	2	-4
3	0	-2	0	6
4	1	-1	-2	-4
5	2	2	1	1
$\sum X^2$	10	14	10	70

β_1 と β_2 は平方和の1次および2次成分がそれぞれ250，200となるように決めた。すなわち

$$\beta_1 = \sqrt{250/10} = 5.00$$
$$\beta_2 = \sqrt{200/14} = 3.78$$

とする。

平方和の3次成分と4次成分の和を200として，β_3 と β_4 を変えて，2次

項がモデルに取り入れられるかどうかをいろいろの規準で比較した。

まず，3次または4次成分のいずれかが200になる場合を比較する。すなわち，

$$\beta_3 = \sqrt{200/10} = 4.47$$
$$\beta_4 = \pm\sqrt{200/70} = \pm 1.69$$

とし，(β_3, β_4) を I $(0, -1.69)$，II $(4.47, 0)$ および III $(0, 1.69)$ と変えた3つのモデルについて比較した結果を表1・4と図1・1に示す。

表1・4の上半分は1次式を，下半分は2次式をあてはめたときの $1/(1-c_i)$，残差 $y-\hat{y}$ と $y-\hat{y}^*$ を与えている。

2乗和の行は $\Sigma(y-\hat{y})^2$ と $\Sigma(y-y^*)^2$ すなわち RSS と PSS を示している。

これからわかるように，いずれのモデルでも RSS は1次式のとき400，2次式のとき200である。これから，2次の項のF比を求めると，

$$F = \frac{RSS_1 - RSS_2}{RSS_2/(n-3)} = \frac{400-200}{200/(5-3)} = 2.00$$

となる。したがって，RSS を規準にするとき，この3つのモデルでは全く同じ結論になる。

それに対して PSS のほうは，3つのモデルでかなりの違いを見せている。

モデル [I] と [III] では，2次式を当てはめた時の残差の符号が逆になっているだけであるため，2次の PSS は共に1058で等しい。

しかるに，1次の PSS は956，1533と大きく異なる。したがって，モデル [I] では1次式を，モデル [III] では2次式を当てはめたとき PSS

が小さくなる。モデル〔Ⅲ〕の1次式のPSSは$1/(1-c_i)$の大きい両端の残差の影響を受けて大きくなったものである。

モデル〔Ⅱ〕では，2次式を当てはめたときのPSSが非常に大きい。2次式の$i=1$と5に対する$1/(1-c_i)$が8.75と極めて大きくモデル〔Ⅱ〕のデータは2次式を当てはめてもe_1とe_5が小さくならないからである。

表1.4 残差平方和と予測平方和の計算

モデル (β_3, β_4)		〔Ⅰ〕 (0, −1.69)		〔Ⅱ〕 (4.47, 0)		〔Ⅲ〕 (0, 1.69)	
i	$1/(1-c)$	$y-\hat{y}$	$y-\hat{y}^*$	$y-\hat{y}$	$y-\hat{y}^*$	$y-\hat{y}$	$y-\hat{y}^*$
〈1次式〉							
1	2.500	5.87	14.68	3.09	7.73	9.25	23.13
2	1.429	3.00	4.29	5.18	7.40	−10.54	−15.06
3	1.250	−17.70	−22.13	− 7.52	9.40	2.58	3.22
4	1.429	3.00	4.29	−12.72	−18.18	−10.54	−15.06
5	2.500	5.87	14.68	12.02	30.05	9.25	23.13
	2乗和	400	956	400	1438	400	1533
〈2次式〉							
1	8.750	−1.69	−14.79	−4.47	−39.11	1.69	14.79
2	1.591	6.76	10.75	8.94	14.22	−6.76	−10.75
3	1.944	−10.14	−19.71	0.00	0.00	10.14	19.71
4	1.591	6.76	10.75	−8.94	−14.22	−6.76	−10.75
5	8.750	−1.69	−14.79	4.47	39.11	1.69	14.79
	2乗和	200	1058	200	3467	200	1058

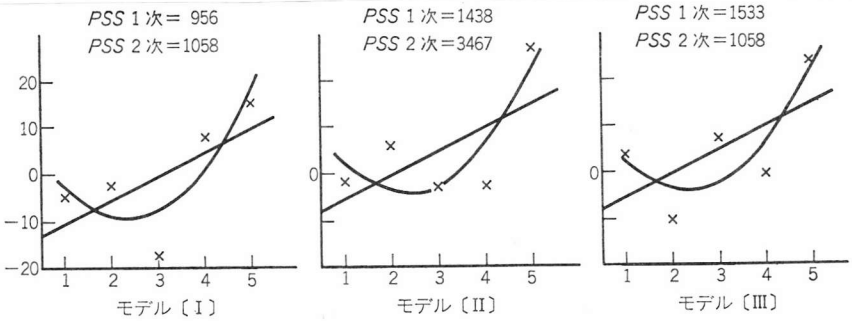

図1.1 PSSとRSSの関係

2次の係数β_2の大きさを変えて，1次のPSSが2次のPSSより大きくなる境界とそれに対するF比を求めた結果を，表1.5に示す。

表 1.5

β_3		0.00	2.23	3.16	3.87	4.47	3.87	3.16	2.23	0.00	
β_4		−1.69	−1.46	−1.19	−0.84	0.00	0.84	1.19	1.46	1.69	
											RSS
PSS2		1058	1660	2263	2865	3467	2865	2263	1660	1058	200
F	β_2										
0.0	0.00	383	431	480	528	577	528	480	431	383	200
0.2	1.20	378	439	501	569	663	660	630	597	560	220
0.4	1.69	426	492	561	636	749	675	743	715	684	240
0.6	2.07	484	553	627	708	835	866	850	827	800	260
0.8	2.39	545	618	696	782	921	964	954	934	910	280
1.0	2.67	610	686	766	857	1007	1061	1055	1039	1018	300
1.2	2.93	677	755	839	934	1094	1157	1155	1142	1124	320
1.4	3.16	745	826	912	1011	1180	1252	1254	1244	1228	340
1.6	3.38	815	897	987	1089	1266	1347	1352	1344	1331	360
1.8	3.59	885	970	1062	1167	1352	1441	1449	1444	1432	380
2.0	3.78	956	1043	1138	1246	1438	1534	1546	1543	1533	400
2.2	3.96	1029	1117	1214	1325	1525	1627	1642	1641	1634	420
2.4	4.14	1101	1192	1291	1404	1611	1720	1737	1739	1733	440
2.6	4.31	1175	1267	1368	1484	1697	1813	1833	1837	1832	460
2.8	4.47	1248	1342	1445	1564	1783	1905	1928	1934	1931	480
3.0	4.63	1323	1418	1523	1644	1869	1998	2022	2030	2029	500
3.2	4.78	1397	1494	1601	1725	1956	2090	2117	2126	2127	520
3.4	4.93	1472	1571	1679	1805	2042	2181	2211	2222	2224	540
3.6	5.07	1547	1648	1757	1886	2128	2273	2305	2318	2321	560
3.8	5.21	1623	1725	1836	1967	2214	2365	2399	2413	2418	580
4.0	5.35	1699	1802	1915	2048	2300	2456	2492	2508	2515	600
4.2	5.48	1775	1879	1994	2129	2386	2547	2585	2603	2611	620
4.4	5.61	1851	1957	2073	2210	2473	2638	2678	2698	2707	640
4.6	5.73	1928	2035	2153	2292	2559	2729	2771	2793	2803	660
4.8	5.86	2005	2113	2232	2373	2645	2820	2864	2887	2898	680
5.0	5.98	2082	2191	2312	2455	2731	2911	2957	2981	2994	700
5.2	6.09	2159	2269	2392	2536	2817	3002	3050	3075	3089	720
5.4	6.21	2236	2348	2472	2618	2904	3092	3142	3169	3184	740
5.6	6.32	2314	2427	2552	2700	2990	3183	3234	3263	3279	760
5.8	6.44	2391	2505	2632	2782	3076	3273	3327	3356	3374	780
6.0	6.55	2469	2584	2712	2864	3162	3364	3419	3450	3468	800
6.2	6.65	2547	2663	2792	2946	3248	3454	3511	3543	3563	820
6.4	6.76	2625	2743	2873	3028	3335	3544	3603	3636	3657	840
6.6	6.87	2703	2822	2953	3110	3421	3634	3695	3729	3751	860
6.8	6.97	2781	2901	3034	3193	3507	3725	3786	3822	3845	880
7.0	7.07	2860	2981	3115	3275	3593	3815	3878	3915	3939	900

表1.4では β_3 と β_4 のいずれか一方のみを含む場合を取り上げたが，表1.5では両方を含む場合も検討した。3次と4次の平方和成分の和がいずれも200になるように，また3次成分と4次成分が 200,0；150,50；100,100；50,150；0,200になるように β_3 と β_4 を決めてある。

3行目のPSS2は，2次式に対するPSS，右端は2次式に対するRSSである。

4行目以下は，分散比Fが0.2きざみになるように2次の係数 β_2 を変え

て，1次式に対するPSSとRSSを求めたものである。

表中で，太字の数値は表1.4で取りあげた数値である。

1次のPSSが2次のPSSを越える境界に区切り（———）を入れてある。2次式が採用されるときのF比の値が，1.2の場合もあるし，6.8のときもある。

PSSによる変数選択の規準はF比による規準とは必ずしも一致しないことがわかる。

（2） 例2

例1と同様に目的変数 y の値を x の9次の多項式で定め，これを x の1次から7次までの多項式で推定するときの変数選択の問題をとりあげる．（8次および9次項を誤差とみなす）．

y を求めるためには規準化した直交多項式 P_{ij} を用いて

$$y_i = \sum_{j=1}^{9} \beta_j P_{ij} \quad (i=1,2,\cdots,n)$$

とする。ここに $n=10$ で，P_{ij} には次の性質がある。

$$\sum_{i=1}^{n} P_{ij}^2 = 1$$

$$\sum_{i=1}^{n} P_{ij} P_{ij'} = 0 \quad (j \neq j')$$

また P_{ij} は次の式により逐次に求めた。（M. Sibuya & T. Haga (1960)）（このプログラムを，巻末に付す）。

初期値を

$$P_{i,0} = 0$$

$$P_{i,1} = 1/\sqrt{n}$$

とすると，j 次の直交多項式 P_{ij} は $(j-1)$ 次と $(j-2)$ 次の直交多項式から次式で導かれる。

$$a = \sum_{i=1}^{m} x_i P_{i,j-1}^2$$

$$b = \sum_{i=1}^{n} x_i P_{i,j-1} P_{i,j-2}$$

$$p_{i,j} = (x_i - a) P_{i,j-1} - b P_{i,j-2}$$

$$P_{i,j} = p_{i,j} / \sqrt{\Sigma p_{i,j}^2}$$

β として次の3つの系列を取りあげた。

j	β [I]	β [II]	β [III]	V [I] [II]	V [III]
1	10.0000	10.0000	10.0000	100.0000	100.0000
2	5.0000	5.0000	7.0711	25.0000	50.0000
3	2.5000	-2.5000	5.0000	6.2500	25.0000
4	1.2500	-1.2500	3.5355	1.5625	12.5000
5	0.6250	0.6250	2.5000	0.3906	6.2500
6	0.3125	0.3125	1.7678	0.0977	3.1250
7	0.1563	-0.1563	1.2500	0.0244	1.5625
8	0.0781	-0.0781	0.8839	0.0061	0.7812
9	0.0391	0.0391	0.6250	0.0015	0.3906

j 次回帰による分散 V_j は

$$V_j = \beta_j^2 \sum_i P_{ij}^2 = \beta_j^2$$

であり，各モデルについて，その具体的な値を上表の右に示す。

$x_i = i$ として，y を x の多項式

$$y = \beta_0' + \beta_1' x + \beta_2' x^2 + \cdots + \beta_9' x^9$$

と表わすときの偏回帰係数 β_j' は次のようになる。

j	(I)	(II)	(III)
0	-15.217	-19.492	-207.74
1	33.151	50.407	568.15
2	-36.805	-56.183	631.44
3	22.109	32.771	376.72
4	-7.966	-11.372	-134.71
5	1.804	2.453	30.237
6	-0.2586	-0.3305	-4.2923
7	0.0228	0.0270	0.3738
8	-0.0011	-0.0012	-0.0182
9	0.00002	0.00002	0.00038

また，x と y の関係を図1.2に示す。

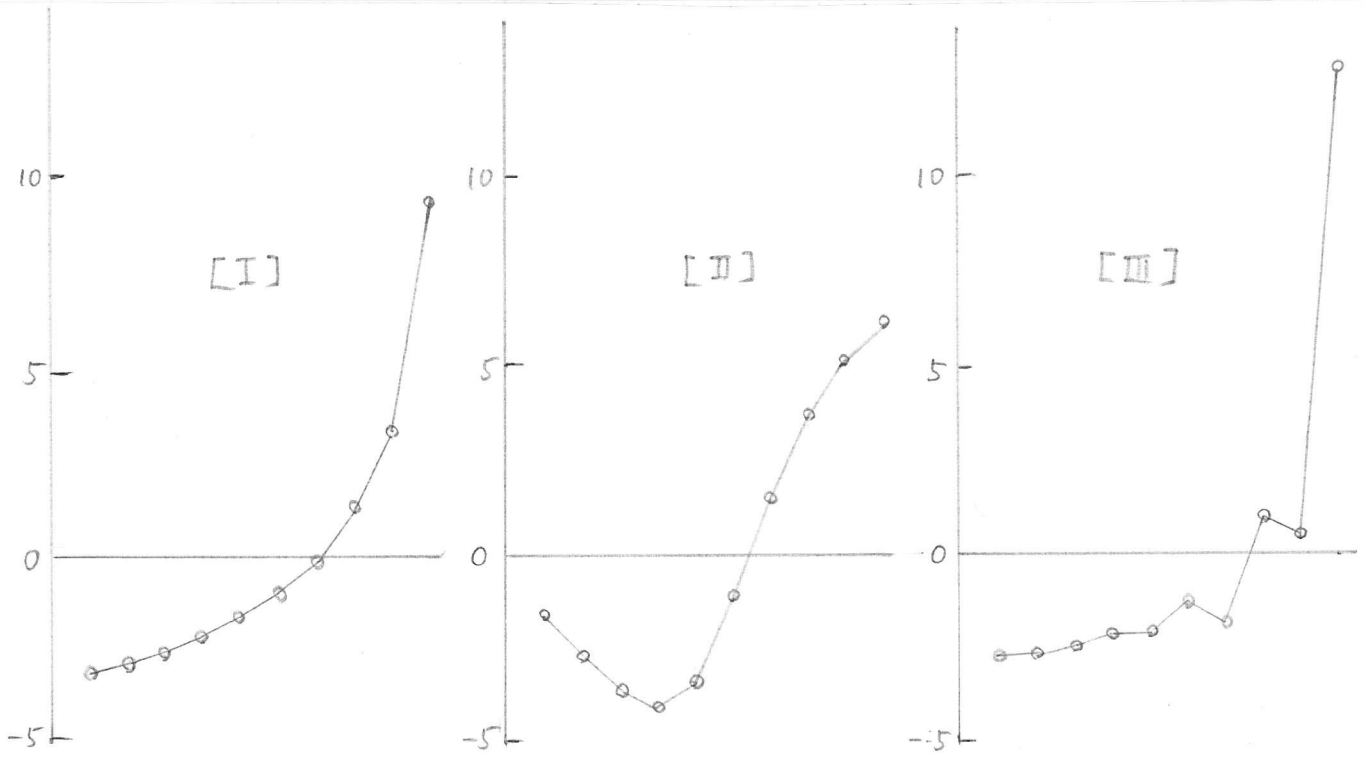

図 1.2

そのようにして求めたyに対して，2組の説明変数x_{ij}を考えて回帰式を求めた。

A： x_iのj乗をj番目の説明変数とする。

$$x_{ij} = x_i^j \quad (x_i = i)$$

B： j次の直交多項式P_{ij}をj番目の説明変数とする。

$$x_{ij} = P_{ij}$$

以下，上記のAとBを単にベキ乗式と直交多項式と呼ぶことにする。

ベキ乗式の説明変数の相関行列を表1.6に示す。

表 1.6 x_i の相関係数

	1	2	3	4	5	6	7
1	1.0000						
2	.9746	1.0000					
3	.9284	.9872	1.0000				
4	.8817	.9615	.9927	1.0000			
5	.8394	.9325	.9769	.9954	1.0000		
6	.8023	.9040	.9579	.9850	.9969	1.0000	
7	.7700	.8773	.9360	.9718	.9896	.9978	1.0000

2組の説明変数と3つの目的変数との相関係数を表1.7に示す。

表1.7 x と y との相関係数(r)

(*印は $max(r)$ を示す)

モデル		x_1	x_2	x_3	x_4	x_5	x_6	x_7
[I]	A	.8660	.9410	.9749	.9884	.9916*	.9890	.9834
	B	.8660*	.4330	.2165	.1083	.0541	.0271	.0135
[II]	A	.8660	.9410	.9516*	.9337	.9051	.8736	.8427
	B	.8660*	.4330	-.2165	-.1083	.0541	.0271	-.0135
[III]	A	.7078	.8020	.8602	.8980	.9236	.9413	.9537*
	B	.7078*	.5005	.3539	.2502	.1769	.1251	.0885

7つの説明変数のすべての部分集合についてPSSとRSSを計算し，取り入れた変数の数ごとに最良の変数組合せを求めた。

モデル[I]に対する結果を表1.8に示す。

ベキ乗式の場合

1変数の最良回帰式はPSSでは x_6，RSSでは x_5 である。(*印を付す)

2変数のとき，PSSとRSSで，同じ変数が選択される。

3変数のとき，PSSでは x_5, x_6, x_7 が，RSSでは x_3, x_6, x_7 が選ばれる。RSS最小の組合せ(x_3, x_6, x_7)のPSSは2.91で，これはPSSの小さい方から5番目である。逆にPSS最小の組合せ(x_5, x_6, x_7)

表 1.8　モデル［I］

べき乗式			直交多項式		
説明変数	PSS	RSS	説明変数	PSS	RSS
5	4.89(2)	2.24*	1	66.96*	33.33*
6	3.91*	2.91(2)			
1 7	3.67*	0.51*	12	33.12*	8.33*
3 67	2.91(5)	0.12*	123	18.13*	2.08*
567	0.44**	0.17(7)			
1 567	6.85*	0.07*	1234	12.52*	0.52*
23 567	8.71(4)	0.03**	12345	11.57**	0.13*
2345 7	8.57*	0.05(6)			
			123456	15.33*	0.03*
1234567	32.30*	0.01*	1234567	32.30*	0.01**

のPSSは，小さい方から7番目である。（カッコ内の数字で示す。）

3変数のときPSSが最小となり（＊＊印で示す），それ以上変数を増すとPSSが増加する。

5変数（$x_2, x_3, x_5 \sim x_7$）の回帰式からどの変数を除いても，残差平方和増加量に対する分散比Fは2以上である。6変数の回帰式にはF比が2以下の変数が含まれるので，$F_{OUT}=2.0$規準で選択されるのは5変数の回帰式（＊＊印で示す）である。

（6変数の組合せのいくつかは桁落ちのため計算できなかった。）

直交多項式の場合

PSSとRSSで同じ変数の組が選ばれるが，PSSでは5変数の，RSSでは全変数の回帰式が選択される。

いずれの場合も，PSSの方が取り入れられる変数の数が少ない。

取り入れる変数の数が同じグループ内で，PSSの順序とRSSの順序にはそれ程大きな差は見られなかった。

PSS規準で選ばれた回帰式で，直交多項式のPSSの値がベキ乗式のPSSの値より小さくならないのは，興味がある。

曲線の滑かさがモデル〔I〕に近いモデル〔II〕の結果を表1.9に示すが，ほとんど同じ傾向が見られる。

表1.9　モデル〔II〕

ベキ乗式			直交多項式		
説明変数	PSS	RSS	説明変数	PSS	RSS
1	23.56*	12.60*	1	56.90*	33.33*
67	24.27(6)	5.85*	12	25.40*	8.33*
56	9.81*	6.27(3)			
234	4.41(4)	0.53*	123	14.07*	2.08*
1 34	2.75*	0.74(3)			
3456	0.25(2)	0.04**	1234	10.00*	0.52*
2 4 67	0.16**	0.05(3)			
1 345 7	1.49(6)	0.04*	12345	9.90**	0.13*
23456	0.62*	0.04(3)			
			123456	14.08*	0.03*
1234567	32.30*	0.01*	1234567	32.30*	0.01**

モデル〔III〕の結果を表1.10に示す。

ベキ乗式。PSSでもRSSでも4変数が選ばれるが，変数の組合せが異なる。

直交多項式の場合は，PSSでは3変数が，RSSでは全変数が選ばれる。

ここまではモデル〔I〕，〔II〕と同じ傾向であるが，モデル〔III〕で説明変数がベキ乗式の場合は，PSSとRSSの順序の逆転が多いという現象が見られる。

表1.10　モデル［III］

べキ乗式			直交多項式		
説明変数	PSS	RSS	説明変数	PSS	RSS
7	92.17**	18.04*	1	207.32*	99.61*
67	273.09(31)	13.11*	12	194.03*	49.61*
1 7	139.90*	17.70(8)			
567	289.73(35)	5.15**	123	192.57**	24.61*
12 7	185.05*	13.37(35)			
4567	593.28(35)	3.76*	1234	247.51(3)	12.11*
1234	247.51*	12.11(35)	123 7	197.48*	21.48(3)
34567	713.21(21)	2.94*	12345	422.05(10)	5.86*
12345	422.05*	5.86(21)	123 67	219.85*	19.92(6)
			123456	1054(4)	2.73*
			12345 7	338.62*	7.42(3)
1234567	4197*	1.17*	1234567	4197*	1.17**

2変数のすべての組合せについてPSSとRSSの値をRSSの大きさの順に並べたのが表1.11である。両者の関係を図1.3に示す。

右下の右上りの曲線はx_1を含む変数組合せを結んだものである。また，右の点はx_1の単回帰式のPSSとRSSである。x_1の単回帰式にx_2，x_3，…，x_7のいずれか1つを追加すると，PSSもRSSも減少する。またx_1との相関が低い変数x_7がRSSの減少に効果的であることがわかる。

表1.11　2変数の組合せに対するPSSとRSS

変数	PSS	RSS	変数	PSS	RSS	変数	PSS	RSS
6 7	273	13.1	1 7	140	17.7	3 4	176	25.2
5 7	242	14.1	3 6	182	18.6	1 5	155	26.3
4 7	213	15.1	4 5	191	19.4	2 4	170	28.7
5 6	223	15.6	2 6	164	20.1	1 4	165	32.3
3 7	187	16.1	1 6	147	21.5	2 3	175	34.4
2 7	163	17.0	3 5	178	21.6	1 3	178	40.0
4 6	202	17.1	2 5	166	24.0	1 2	194	49.6

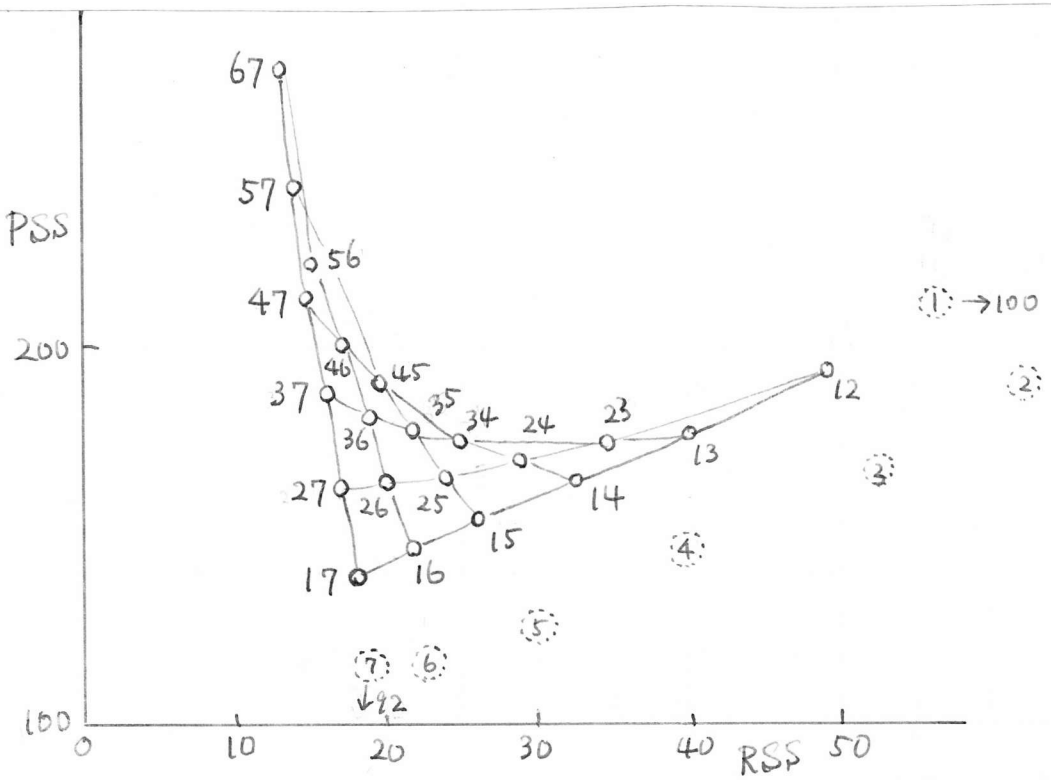

図1.3 モデル[Ⅲ]べき乗式で2変数の場合

左のほぼ鉛直の曲線は x_7 を含むものである。x_7 に，$x_1, x_2, \cdots, x_6$ のいずれを加えても，RSS は減少し，PSS は増加する。RSS の減少に役立つ変数程 PSS の増加量が大きいという逆の関係がある。

まず，x_1 に $x_2, x_3, \cdots, x_7$ のいずれかを追加した場合についてくわしくしらべる。RSS と PSS の基礎となる $e_i\ (=y_i-\hat{y}_i)$，$1/(1-c_i)$ と PSS，RSS の値を表1.12に示す。$i=9$ と10の PSS 成分 $\{e_i/(1-c_i)\}^2$ の値が表の下につけ加えてある。これから PSS の大部分は $i=9$ と10によるものであることがわかる。これらの値の大小関係をまとめると次のようになる。

表 1.12　x_1 に $x_2 \sim x_7$ を追加した場合

i	x_1		x_1, x_2		x_1, x_3		x_1, x_4	
	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i
1	1.528	2.23	2.619	-1.46	2.216	-1.25	2.017	-1.04
2	1.331	1.28	1.387	0.05	1.401	-0.22	1.402	-0.32
3	1.213	0.35	1.224	0.96	1.215	0.59	1.213	0.32
4	1.146	-0.46	1.243	1.38	1.207	1.16	1.183	0.90
5	1.115	-1.43	1.289	1.03	1.264	1.09	1.234	0.99
6	1.115	-1.75	1.289	0.71	1.307	1.05	1.304	1.20
7	1.146	-2.93	1.243	-1.09	1.284	-0.56	1.316	-0.17
8	1.213	-1.79	1.224	-1.17	1.243	-0.70	1.268	-0.24
9	1.331	-3.47	1.387	-4.71	1.372	-4.64	1.356	-4.45
10	1.528	7.98	2.619	4.29	3.172	3.48	3.868	2.82
RSS		99.61		49.61		39.99		32.32
PSS		207.32		194.03		177.63		165.01
$\{e_9/(1-c_9)\}^2$		21.34		42.64		40.58		36.36
$\{e_{10}/(1-c_{10})\}^2$		148.60		126.33		121.88		119.03

i	x_1, x_5		x_1, x_6		x_1, x_7	
	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i
1	1.904	-0.85	1.834	-0.69	1.786	-0.55
2	1.398	-0.35	1.392	-0.33	1.387	-0.30
3	1.214	0.14	1.215	0.03	1.216	-0.04
4	1.169	0.66	1.161	0.48	1.156	0.34
5	1.208	0.83	1.188	0.65	1.173	0.48
6	1.289	1.22	1.269	1.17	1.250	1.07
7	1.335	0.08	1.343	0.23	1.342	0.30
8	1.296	0.17	1.322	0.52	1.347	0.80
9	1.343	-4.19	1.335	-3.90	1.331	-3.59
10	4.731	2.28	5.799	1.84	7.116	1.49
RSS		26.27		21.49		17.70
PSS		154.82		146.57		139.90
$\{e_9/(1-c_9)\}^2$		31.70		27.14		22.85
$\{e_{10}/(1-c_{10})\}^2$		116.42		113.84		112.36

	x_1, x_2		x_1, x_3		...		x_1, x_7
RSS	49.61	>	39.99	>	...	>	17.70
PSS	194.03	>	177.63	>	...	>	134.90
$\lvert e_9 \rvert$	4.71	>	4.64	>	...	>	3.59
$1/(1-e_9)$	1.39	>	1.37	>	...	>	1.33
$\lvert e_{10} \rvert$	4.29	>	3.48	>	...	>	1.49
$1/(1-c_{10})$	2.62	<	3.17	<	...	<	7.12
$e_{10}/(1-c_{10})$	11.24	>	11.04	>	...	>	10.60

$i=9$ について，$x_2, x_3, \cdots, x_7$ の順に $|e_9|$ と $1/(1-c_9)$ が共に減少するので，RSSの成分 e_9^2 とPSSの成分 $\{e_9/(1-c_9)\}^2$ は同じ方向に変化する。$i=10$ について，$|e_{10}|$ は減少し，$1/(1-c_{10})$ は増加するが，$|e_{10}|$ の減少速度の方が大きいので，$\{e_{10}/(1-e_{10})\}^2$ は減少する。したがって，PSS は RSS と同じ方向に変化する。

それに対して，x_7 に $x_1, x_2, \cdots, x_6$ のいずれかを追加したときの結果を表1.13に，その一部を次に示す。

	x_7, x_1		x_7, x_2		...		x_7, x_6
RSS	17.70	>	17.00	>	...	>	13.11
PSS	139.90	<	162.99	<	...	<	272.17
$\lvert e_9 \rvert$	3.59	>	3.41	>	...	>	2.59
$1/(1-c_9)$	1.33	<	1.40	<	...	<	1.83
$e_9/(1-c_9)$	4.78	>	4.78	>	...	>	4.74
$\lvert e_{10} \rvert$	1.49	>	1.34	>	...	>	0.78
$1/(1-c_{10})$	7.12	<	8.64	<	...	<	19.72
$e_{10}/(1-c_{10})$	10.60	<	11.57	<	...	<	15.38

$x_1, x_2, \cdots, x_6$ の順に e_{10} は減少するが，$1/(1-c_{10})$ の方がそれ以上に増加するため，$e_{10}/(1-c_{10})$ が増加する。PSSの中で大きな割合を占める $i=10$ で e_{10}^2 と $\{e_{10}/(1-c_{10})\}^2$ の順序が逆転するので，$i=1 \sim 9$ の如何にかかわらず，PSS と RSS とが逆の動きをする。

表1.13

i	x_7		x_7, x_1		x_7, x_2		x_7, x_3	
	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i
1	1.155	-0.23	1.786	-0.55	1.457	-0.66	1.336	-0.71
2	1.155	-0.08	1.387	-0.30	1.362	-0.45	1.313	-0.53
3	1.155	0.09	1.216	-0.04	1.254	-0.18	1.261	-0.29
4	1.154	0.36	1.156	0.34	1.174	0.23	1.194	0.12
5	1.151	0.41	1.173	0.48	1.155	0.46	1.152	0.39
6	1.142	0.90	1.250	1.07	1.219	1.15	1.187	1.16
7	1.124	0.08	1.342	0.30	1.367	0.48	1.350	0.62
8	1.112	0.56	1.347	0.80	1.462	1.04	1.555	1.27
9	1.239	-3.73	1.331	-3.59	1.402	-3.41	1.490	-3.21
10	5.098	1.63	7.116	1.49	8.637	1.34	10.586	1.18
RSS		18.04		17.70		17.00		16.07
PSS		92.17		139.90		162.99		187.41
$\{e_9/(1-c_9)\}^2$		21.34		22.85		22.85		22.85
$\{e_{10}/(1-c_{10})\}^2$		69.05		112.36		133.86		156.00

i	x_7, x_4		x_7, x_5		x_7, x_6	
	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i	$1/(1-c_i)$	e_i
1	1.281	-0.73	1.252	-0.74	1.234	-0.75
2	1.276	-0.57	1.250	-0.59	1.233	-0.60
3	1.253	-0.36	1.240	-0.40	1.228	-0.42
4	1.206	0.02	1.210	-0.04	1.209	-0.08
5	1.157	0.30	1.164	0.22	1.169	0.15
6	1.164	1.13	1.151	1.07	1.145	1.01
7	1.316	0.70	1.278	0.73	1.245	0.73
8	1.614	1.48	1.642	1.64	1.643	1.76
9	1.591	-3.00	1.705	-2.79	1.830	-2.59
10	13.018	1.03	16.023	0.90	19.722	0.78
RSS		15.07		14.07		13.11
PSS		213.44		241.52		272.17
$\{e_9/(1-c_9)\}^2$		22.75		22.66		22.47
$\{e_{10}/(1-c_{10})\}^2$		179.83		207.94		236.54

〔補足〕

本文中でモデル(I)に対して，直交多項式により，ベキ乗式の方がよく適合することを述べた。

これに関連する3つの回帰式に対する残差を表1.14に，グラフを図1.5に示す。

(x_5, x_6, x_7)は，PSSが最小となるベキ乗式回帰である。説明変数が同数の3個の直交多項式回帰で，PSSが最小となるのが(x_1, x_2, x_3)である。

直交多項式でPSSが最小になるのは$(x_1, x_2, x_3, x_4, x_5)$である。

直交多項式は，1次，2次と低次の方から順次取り入れられる。$i=9, 10$で急にyが変化するのに追随することができないので，$i=9, 10$における残差が大きくなり，多項式のあてはめでは，高次の項を取り入れることにより，全データに対して平均的に良く当てはまる回帰式を得ることができる。

一般に，説明変数は相互に相関の低い方が良いので，多項式を当てはめるに当っては，$(x-\bar{x})^i$とするか，直交多項式を用いるのが良いと言われている。しかし，この例が示すように，曲線の形によっては，変換しない方が良い場合もあり得る。

表 1.14

x_{ij}	x^j	P_j	P_j
変数の組	5,6,7	1,2,3	1,2,3,4,5
RSS	0.17	2.08	0.13
PSS	0.44	18.13	11.57
1	-0.26	0.32	0.03
2	-0.03	-0.30	-0.10
3	0.14	-0.35	0.07
$y-\hat{y}$ 4	0.18	-0.09	0.08
5	0.06	0.24	-0.05
6	-0.08	0.45	-0.11
7	-0.13	0.30	-0.02
8	0.14	-0.14	0.24
9	-0.05	-1.02	-0.19
10	0.01	0.60	0.05

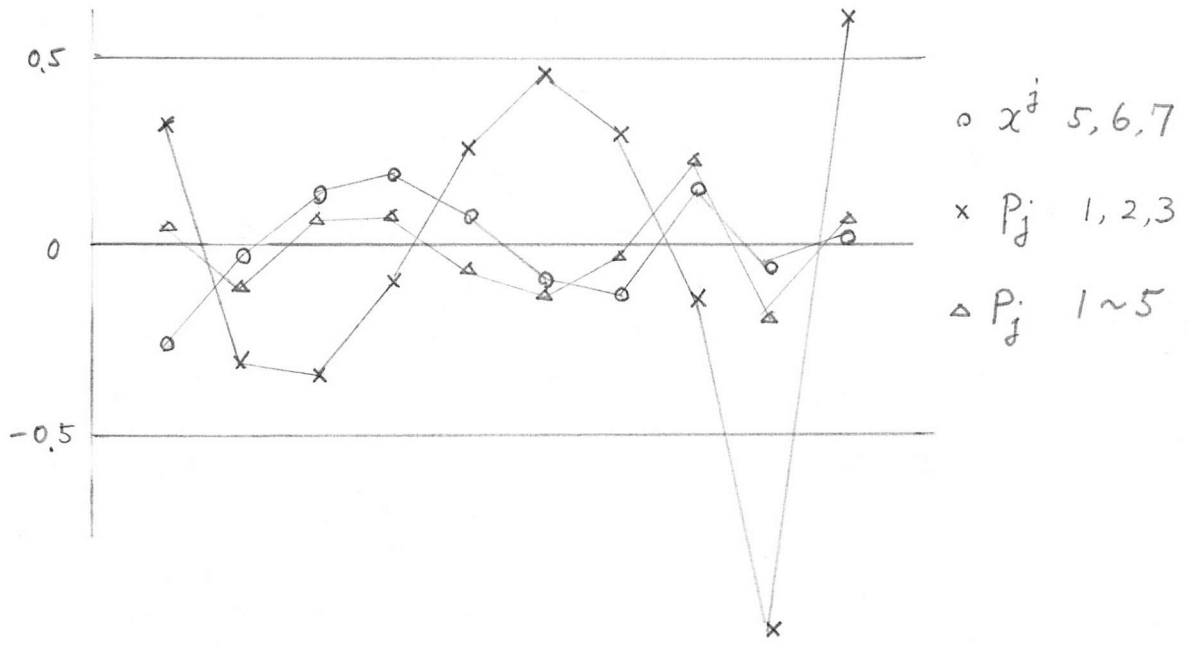

図 1.5

2. 予測平方和(PSS)と残差平方和(RSS)の関係

前節で導いたPSSの挙動がRSSのそれとどのように違うかをシミュレーション実験によってしらべた。

この種のシュミレーションでは，回帰モデル，計画行列，誤差の大きさ，等の選び方によって，異なる結果が得られる。シミュレーション実験によって普遍性のある結論を得るためには，モデルの選定に十分注意をはらわなければならない。この研究に当っては，まず比較的単純なモデルを対象として，正しい結論が導かれるように努めた。

なお，シミュレーションの繰返し回数は特に断わらない限り，1000回とした。

2.1 説明変数がない場合

$p=0$ すなわち常数項のみの場合の PSS と RSS をそれぞれ PSS_0, RSS_0 で表わすと,

$$PSS_0 = \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{1 - 1/n} \right)^2$$
$$= \left(\frac{n}{n-1} \right)^2 \sum (y_i - \bar{y})^2$$
$$= \left(\frac{n}{n-1} \right)^2 RSS_0 \qquad (2.1)$$

となる。すなわち, PSS_0 と RSS_0 の比は常に

$$\frac{PSS_0}{RSS_0} = \left(\frac{n}{n-1} \right)^2 \qquad (2.2)$$

となる。したがって, PSS_0 の期待値は

$$E[PSS_0] = \left(\frac{n}{n-1} \right)^2 E[RSS_0]$$
$$= \frac{n^2}{n-1} \sigma^2 \qquad (2.3)$$

であって, 式(1.19)で求めた PSS の期待値の下限に等しい。

2.2 説明変数が1個で, 正規分布の場合

回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (2.4)$$

において, x_i と e_i がいずれも平均 0, 分散 1 の正規分布に従う場合を検討する。以下, このモデルで求めた PSS と RSS をそれぞれ, PSS_1, RSS_1 で表わす。n 組 ($n=10, 20$ または 40) の独立な標準正規乱数 (x_i, e_i) を発生し, まず $\beta_0=0$, $\beta_1=0$ すなわち $y_i = e_i$ のときの PSS と RSS の性質をしらべた。

(1) PSS_1 と RSS_1 の分布

PSS_1 の基本統計量（平均，標準偏差，歪度，尖度），PSS_1 と RSS_1 の相関係数および，比 PSS_1/RSS_1 の基本統計量を表2.1に示す。シミュレーションの精度を確かめるために，RSS_1 の基本統計量についてその実験値と理論値を併記してある。

PSS_1/RSS_1 の分布形については，後で取りあげる。

表2.1 RSSとPSSの統計量

	n	10	20	40
RSS	平均	8.04(8.0)	17.77(18.0)	38.29(38.0)
	標準偏差	4.15(4.0)	5.87(6.0)	8.29(8.72)
	$\sqrt{b_1}$	1.09(1.0)	0.51(0.67)	0.44(0.46)
	b_2	4.82(4.50)	3.12(3.67)	3.12(3.32)
PSS	平均	13.05(12.50)	22.84(22.22)	42.49(42.11)
	標準偏差	7.14	7.31	9.23
	$\sqrt{b_1}$	1.26	0.51	0.45
	b_2	5.42	3.10	3.14
$r(PSS, RSS)$		0.936	0.988	0.996
PSS/RSS	平均	1.622	1.241	1.110
	標準偏差	0.280	0.059	0.019
	$\sqrt{b_1}$	2.64	1.33	1.07
	b_2	14.93	6.03	5.28

PSSの平均の行の括弧内の値は，PSSの期待値の下限（式(1.18)）である。

(2) 回帰の有意性検定

PSS_1 が PSS_0 より小さいとき「説明変数を回帰式にとり入れる」という方法を"PSS規準"と呼ぶことにする。

一般には，§1.3 に述べた"F規準"

$$F = \frac{RSS_0 - RSS_1}{RSS_1/(n-2)} \qquad (2.5)$$

が用いられる。

$\beta_1 = 0$ のとき，PSS_1 が PSS_0 より小さくなる確率（PSS規準の第1種の誤りの確率 α）をシミュレーションより求め，F規準の α と比較した。その結果を表2.2に示す。

表 2.2

n	10	20	40
$Pr[PSS_1 < PSS_0]$	0.22	0.19	0.18
$Pr[F > 2.0]$			
経験確率	0.21	0.18	0.18
理論確率	0.20	0.18	0.17

表2.2の下半分は，式(2.5)のFの値が2以上になる確率でシミュレーションで求めた経験確率と，F分布から計算した理論確率である。

PSS規準の α は，nの値によらず，規準値を2.0としたF規準に近い。ただし，すでに述べたように PSS_1 と RSS_1 との間には代数的関係がないので，2つの規準による個々の検定結果が一致するというわけではない。

式(2.5) に式(2.2)を代入すると

$$F = \frac{PSS_0\left(\frac{n-1}{n}\right)^2 - RSS_1}{RSS_1/(n-2)}$$

$$PSS_0 = RSS_1\left(\frac{n}{n-1}\right)^2\left(1+\frac{F}{n-2}\right) \qquad (2.6)$$

が得られる。$F=F_0$のとき, PSS_1 が PSS_0 より小さくなる確率は

$$Pr\left[PSS_1 < PSS_0 \mid F=F_0\right]$$

$$= Pr\left[\frac{PSS_1}{RSS_1} < \left(\frac{n}{n-1}\right)^2\left(1+\frac{F_0}{n-2}\right)\right] \qquad (2.7)$$

となる。

図2.1は PSS_1/RSS_1 ($n=40$) の経験累積分布のグラフである。横軸には, PSS_1/RSS_1 と, それに対応する(2.7)式のFの値の2つが目盛ってある。

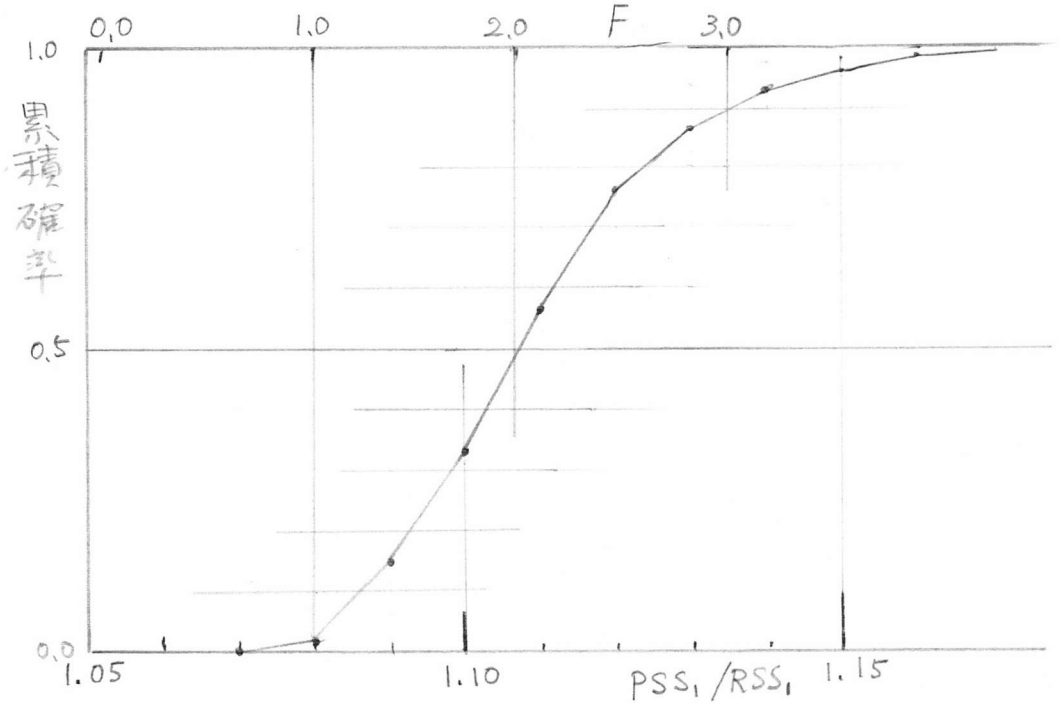

図2.1 PSS_1/RSS_1 の累積分布

$F=2.0$のとき，PSS 規準で説明変数が回帰式に取り入れられる割合は約50％である。Fが1.5であっても，20％位の割合でPSS 規準で変数が取り入れられる。逆に，$F=3.0$であっても，PSS 規準では回帰式にとり入れられないことが10％位起こる。

PSS_1/RSS_1の値が極端に小さい場合と大きい場合のxとyの散布図を図2.2 に示す。

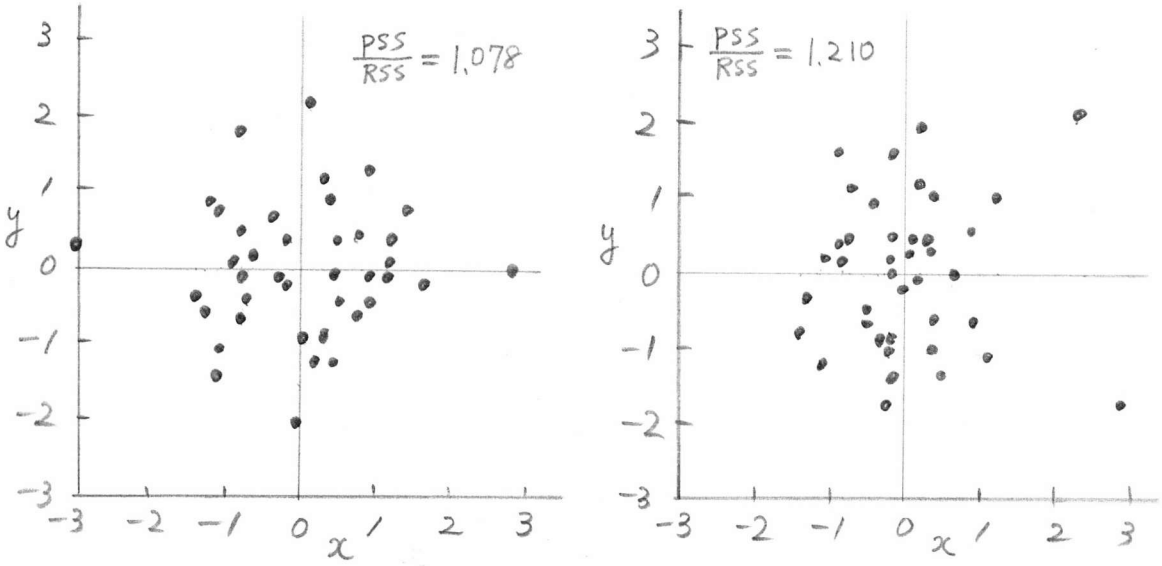

図2.2 PSS/RS の小さい場合と大きい場合の例

図2.2の右から分るように残差$|y_i-\hat{y}|$の大きい点が，$|x_i-\bar{x}|$の大きいところにあるとき PSS/RSS が大きくなる。これは$1/(1-c_i)$が$|x_i-\bar{x}|$と共に増加することから当然のこととして理解できる。

(3) 検出力

回帰モデルの(2.4)式で、β_1を0.00, 0.05, 0.10, ···, 1.00と変えて、$Pr[PSS_1 < PSS_0]$を求めると、PSS規準で説明変数が回帰式に取り入れられる確率(以下"検出力"と呼ぶことにする)が求められる。

F規準で規準値を2.0としたときの検出力と共に、表2.3と図2.3に実験結果を示す。これから、β_1が大きいところでPSS規準の検出力はF規準の検出力より幾分低下するが、両者間に実用上問題にする程の差はない。

表 2.3 検出力の比較

β_1	n = 10		n = 20		n = 40	
	PSS	F>2	PSS	F>2	PSS	F>2
0.00	.217	.209	.194	.188	.182	.179
0.05	.232	.220	.212	.191	.198	.197
0.10	.245	.232	.236	.233	.264	.254
0.15	.255	.259	.286	.281	.358	.353
0.20	.276	.287	.360	.333	.457	.450
0.25	.306	.312	.441	.417	.576	.563
0.30	.346	.351	.496	.497	.668	.685
0.35	.386	.383	.571	.569	.760	.777
0.40	.424	.442	.627	.648	.838	.859
0.45	.469	.488	.703	.720	.895	.909
0.50	.516	.536	.758	.783	.940	.949
0.55	.596	.591	.797	.833	.964	.970
0.60	.604	.641	.839	.860	.987	.987
0.65	.643	.687	.882	.898	.992	.996
0.70	.680	.723	.919	.912	.997	.998
0.75	.721	.766	.939	.944		
0.80	.750	.794	.952	.957		
0.85	.789	.819	.960	.970		
0.90	.817	.852	.972	.976		
0.95	.847	.869	.983	.985		
1.00	.872	.894	.990	.993		

以上、xとeの分布が正規分布に従うとき、PSS_1とRSS_1との間には高い相関があり、PSS規準とF規準(規準値=2.0)とはほとんど差がないことがわかる。

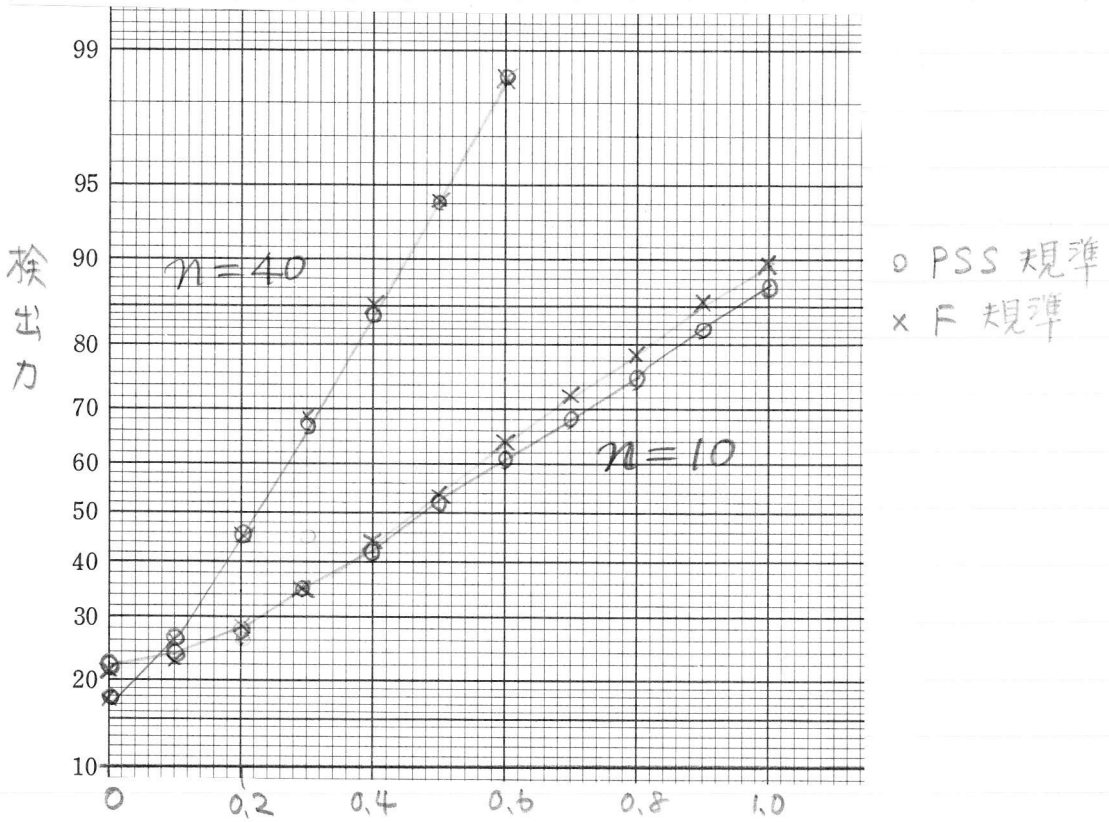

図 2.3　検出力の比較

〔注〕 β を変えるごとに新しく乱数 y を発生させて，PSS，RSS を計算し直すことなく，次の方法で検出力を求めた。

互いに独立な正規乱数 x_i，e_i（$i=1,2,\cdots,n$）を固定したとき，2つのモデル

$$\begin{aligned} y_i &= e_i \\ y_i^* &= \beta x_i + e_i \end{aligned} \qquad (2.8)$$

の PSS，RSS の間には次の関係がある。

$$\begin{aligned} RSS_0^* &= RSS_0 + \beta^2 \Sigma (x_i-\bar{x})^2 + 2\beta \Sigma (x_i-\bar{x})(y_i-\bar{y}) \\ RSS_1^* &= RSS_1 \\ PSS_0^* &= \left(\frac{n}{n-1}\right)^2 RSS_0^* \\ PSS_1^* &= PSS_1 \end{aligned} \qquad (2.9)$$

RSS_1 を計算するために $\Sigma (x_i-\bar{x})^2$，$\Sigma (x_i-\bar{x})(y_i-\bar{y})$ が求められているので，上の関係を使って，RSS_0^* と PSS_0^* を求めることができる。

2、3　説明変数が1個で非正規分布の場合

前記のモデル

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (2.4)$$

において，x_i と e_i が正規分布でないとき，特に分布の裾の広がりが変化したとき，PSS と RSS の関係がどのように変るかをしらべた。

x_i と e_i の分布として，次の4つの分布を取りあげ，16の組合せについて比較した。β_2 は各分布の尖度である。

a	一様分布	$\beta_2 = 1.8$
b	正規分布	$\beta_2 = 3.0$
c	t分布（$f=6$）	$\beta_2 = 13.5$
d	t分布（$f=2$）	$\beta_2 = \infty$

x_i または e_i のいずれかの分布が，自由度2の分布（d）の場合は裾の影響が大きく，1000回の反復では断定的な結論を下すことができなかったが，参考値として，記録しておく。

また，サンプルの大きさnは10，20，40について実験したが40の場合のみを記す。

（1）　比　PSS_1/RSS_1 の分布

尖度が∞になる自由度2のt分布を除くと，xとeの分布の尖度が大きくなっても比 PSS_1/RSS_1 の平均値は殆んど変化しないが，その標準偏差，

歪度，尖度は大きくなる。（表2.4）。

xの分布とeの分布の影響を比較すると，xの分布の影響の方が著しく大きい。

図2.3には，xの分布を正規分布に固定してeの分布を変えたときの，図2.4には，eの分布を正規分布に固定してxの分布を変えたときのPSS_1/RSS_1の経験累積分布を比較したものである。

eが正規分布で，xが自由度6のt分布のとき，$F_0=3.0$であっても，$PSS_1 < PSS_0$となる確率が15%位あることがわかる。逆に$F_0=1.0$であっても PSS 規準で有意となる確率が 5%近くある。

表2.4 PSS_1/RSS_1の分布およびPSS_1とRSS_1の相関係数

PSS_1/RSS_1	x \ e	U	N	t(6)	t(2)
平均	U	1.1082	1.1080	1.1088	1.1087
	N	1.1105	1.1095	1.1098	1.1089
	t(6)	1.1104	1.1102	1.1110	1.1128
	t(2)	1.1760	1.1576	1.1503	1.1695
標準偏差	U	0.0087	0.0118	0.0159	0.0261
	N	0.0147	0.0186	0.0219	0.0376
	t(6)	0.0203	0.0271	0.0304	0.0511
	t(2)	1.2026	0.3973	0.3235	0.5765
歪度	U	0.3064	0.5176	1.2524	0.9570
	N	0.9079	1.0769	1.3678	1.9006
	t(6)	2.7850	3.1602	2.2945	3.7781
	t(2)	30.3995	18.4767	12.4999	17.6009
尖度	U	3.250	3.846	8.056	4.530
	N	5.484	5.283	5.949	8.102
	t(6)	23.873	24.520	11.893	30.086
	t(2)	946.192	423.418	199.515	470.048
r (PSS_1, RSS_1)	U	0.9977	0.9978	0.9978	0.9975
	N	0.9953	0.9959	0.9908	0.9989
	t(6)	0.9924	0.9927	0.9949	0.9984
	t(2)	0.1874	0.5263	0.7173	0.9965

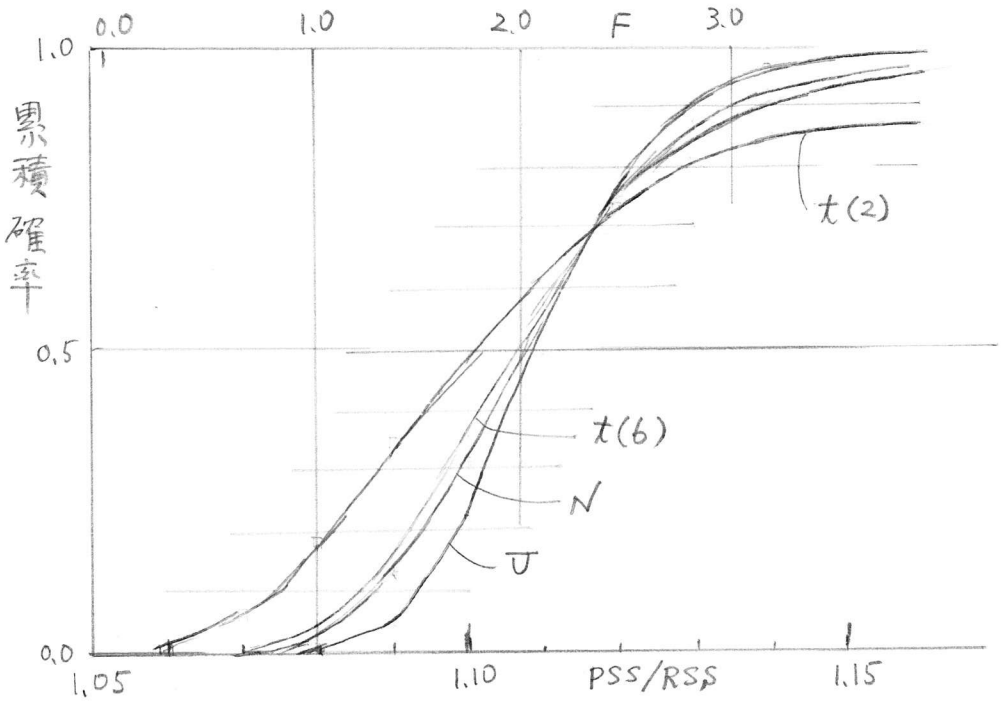

図 2.3　eの分布の影響
（xの分布は正規分布に固定）

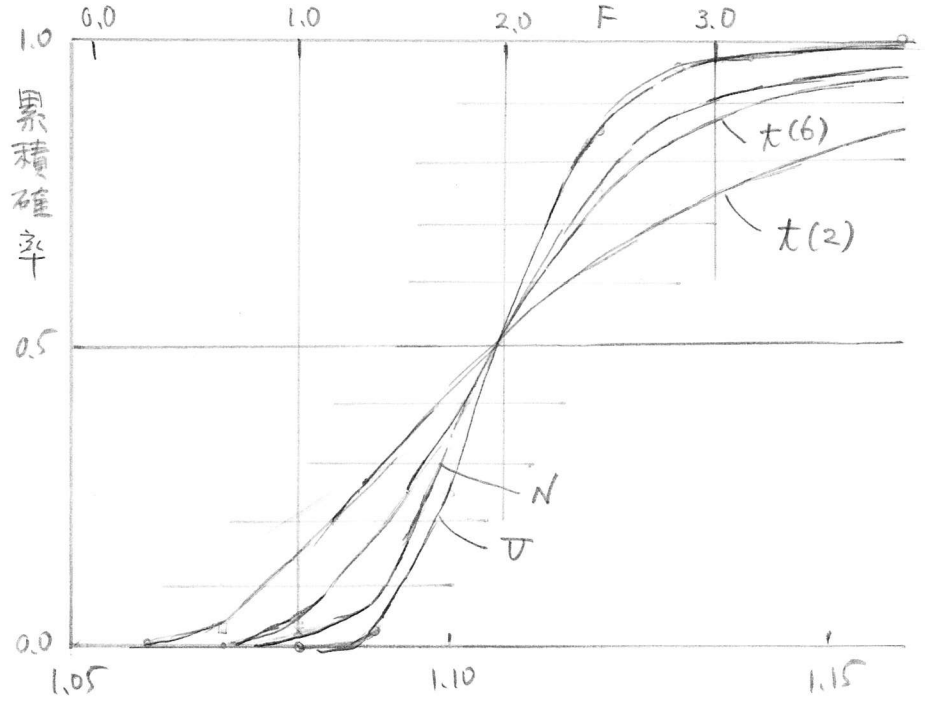

図 2.4　xの分布の影響
（eの分布は正規分布に固定）

（2） 第1種の誤りの確率

xとeの分布の変化に対する，PSS規準とF規準（規準値2.0）の第1種の誤りの経験確率を表2.5に示す。（$n=40$）

表 2.5 xとeの分布の違いによる危険率の変化

x \ e	PSS規準				F規準			
	U	N	t(6)	t(2)	U	N	t(6)	t(2)
U	.168	.168	.168	.165	.160	.168	.179	.176
N	.164	[.196]	.158	.180	.150	[.182]	.163	.169
t(6)	.173	.173	.180	.125	.165	.155	.166	.150
t(2)	.205	.217	.187	.146	.155	.193	.175	.153

表2.5でPSS規準とF規準との間に経験確率には有意差は見られない。また，この程度の差であれば，実用的にも問題にならないと思われる。

2．4　説明変数が1個で異常値の混入した場合

n個のデータの中に少数個の異常値が混入するとき，PSS規準がどのように影響を受けるかをF規準と比較する。

結論を先に述べると，F規準はある種の異常値の混入によって第1種の誤りの危険率αがかなり大きくなるのに対して，PSS規準は余り影響を受けない。　すなわち，

"PSS規準はある種の異常値の混入に対して，ロバストである。"

ということができる。

独立な正規乱数 $x_i \in N(0.1^2)$, $y_i \in N(0,1^2)$ をn組作り，そのうちの1組または2組についてxとeのいずれか一方，または両方をσ倍して，異常値とする。3つのモデル

$$\begin{aligned} &i) \quad x^* = \sigma x, \quad y^* = \sigma y \\ &ii) \quad x^* = \sigma x, \quad y^* = y \\ &iii) \quad x^* = x, \quad y^* = \sigma y \end{aligned} \qquad (2.10)$$

について検討した。

まず，モデルi)で，$n=10,20,40$, $\sigma=2,3,4,5$ とした場合の結果について述べる。

PSS_0, PSS_1, RSS_0, RSS_1 と F値を計算し，$Pr[PSS_1 < PSS_0]$, $Pr[F>2.0]$ を求めた。

表2.6および図2.6から一見してわかるように，F規準は異常値が，1～2個入るだけで第1種の誤りの危険度αが大きく変化するのに対して，PSS規準は安定している。αの受ける影響は，F規準がPSS規準の5倍位で，αが50%を越えることもある。

表2.6　異常値の混入による第1種の危険率αの変化（モデル i)）

異常値の個数	σ	n＝10		n＝20		n＝40	
		PSS	F	PSS	F	PSS	F
0		0.22	0.21	0.19	0.19	0.18	0.18
1	2	0.23	0.25	0.19	0.20	0.17	0.19
	3	0.23	0.35	0.22	0.34	0.18	0.28
	4	0.24	0.46	0.21	0.42	0.18	0.38
	5	0.27	0.54	0.25	0.53	0.20	0.46
2	2	0.22	0.27	0.20	0.26	0.18	0.22
	3	0.24	0.36	0.23	0.36	0.20	0.33
	4	0.27	0.45	0.26	0.48	0.25	0.47
	5	0.27	0.51	0.28	0.54	0.25	0.54

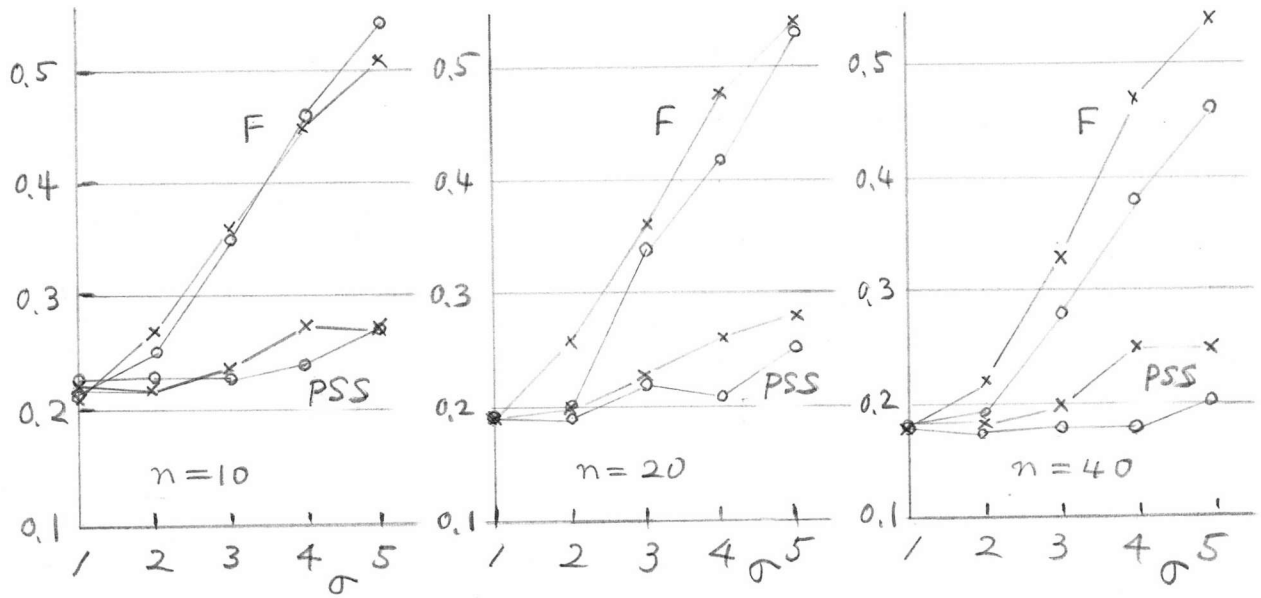

図2.6　モデル i)　　○ 1個　× 2個

ちなみに，$n=20$ で，F が $F(1, 18; 0.05)=4.41$ より大きいとき，変数をとり込むとしたときの，第1種の誤りの危険率は，シミュレーションによれば次のようになった。

異常値のσ		2	3	4	5
異常値の個数	1	0.08	0.17	0.23*	0.38*
	2	0.18	0.18	0.30*	0.39*

肩に*を付した値は，表2.6で示したPSS規準よりも危険率の高いことを示している。この結果から，F分布の5%値を使って検定しても，異常値の混入によって，α が20%を越え，PSS規準の α より大きくなる場合のあることがわかる。

次にモデルii)とiii)について述べる。異常値の混入による α の変化を表2.7と表2.8に示す。そこには，モデルi)で見られたような傾向は全くない。PSS規準もF規準も共に異常値の混入に対してロバストであるということができる。

表2.7　モデルii)　$x_1^* = \sigma x_1$

異常値の個数	σ	$n=10$ PSS	$n=10$ F	$n=20$ PSS	$n=20$ F	$n=40$ PSS	$n=40$ F
0		0.22	0.21	0.19	0.19	0.18	0.18
1	2	0.23	0.20	0.19	0.18	0.16	0.15
	3	0.19	0.19	0.18	0.17	0.18	0.16
	4	0.20	0.19	0.17	0.15	0.18	0.17
	5	0.18	0.18	0.18	0.19	0.17	0.16
2	2	0.21	0.20	0.18	0.17	0.17	0.16
	3	0.22	0.20	0.19	0.17	0.17	0.15
	4	0.19	0.20	0.17	0.17	0.18	0.16
	5	0.18	0.18	0.18	0.15	0.19	0.16

表2.8　モデルiii)

異常値の個数	σ	n=10 PSS	n=10 F	n=20 PSS	n=20 F	n=40 PSS	n=40 F
0		0.22	0.21	0.19	0.19	0.18	0.18
1	2	0.18	0.17	0.19	0.16	0.18	0.18
	3	0.19	0.20	0.17	0.16	0.15	0.15
	4	0.17	0.20	0.15	0.17	0.15	0.16
	5	0.14	0.18	0.17	0.16	0.15	0.15
2	2	0.21	0.20	0.19	0.18	0.17	0.16
	3	0.18	0.20	0.15	0.16	0.17	0.16
	4	0.17	0.20	0.15	0.17	0.17	0.17
	5	0.14	0.20	0.15	0.16	0.14	0.15

図2.2で見たように，x, yが共にその平均から飛び離れているとき，PSSはRSSと異なる挙動をとる。モデルi)は，xとyを重心から放射状に飛び出させるものであるのに対して，モデルii)とiii)は平均から水平または鉛直方向に遠ざけるものであるから，PSSがRSSと特に異なる結果を与えないと考えられる。

都道府県のデータでは東京都，北海道，沖縄県は他府県と大きく異なり，グラフの右上や左下の隅に孤立することが多い。このようなデータに対して通常の変数選択を適用すると，飛び離れた数点のあてはめを良くする変数が選ばれる傾向が見られる。それに対して，PSSは全体として予測に役立つ変数が選出される傾向があると言うことができよう。

現実のデータには，種々の理由で異常値の含まれることが多い。最近汚れたデータの取扱い方が大きな問題となり，種々のロバスト推定法が研究されている。ここに説明しているPSSは，重回帰分析の変数選択において，ある種の異常値に対するロバスト性を持っていると言うことができる。

2.5 説明変数が2個の場合

(1) モデルと変数選択の規準

x_1 が x_2 と y の変化の原因となっており，x_2 と y との間には直接の因果関係を持たないモデルを考える。このモデルは，z_1, z_2, z_3 を互いに独立に分布する分散1の確率変数とすれば，

$$\begin{aligned} x_1 &= z_1 \\ x_2 &= \alpha x_1 + z_2 \\ y &= \beta x_1 + z_3 \end{aligned} \qquad (2.11)$$

と表わすことができる。そのとき各変数間の相関係数は

$$\begin{aligned} r(x_1, x_2) &= \alpha / \sqrt{1+\alpha^2} \\ r(x_1, y) &= \beta / \sqrt{1+\beta^2} \\ r(x_2, y) &= \alpha\beta / \sqrt{(1+\alpha^2)(1+\beta^2)} \\ &= r(x_1, x_2) \cdot r(x_1, y) \end{aligned} \qquad (2.12)$$

となる。α, β が共に0でないとき，x_2 と y との間には見掛上の相関が見られる。

α, β を変えて，PSS規準およびF規準で選ばれる回帰式を比較する。$\beta \neq 0$ のとき，x_1 のみの単回帰式が選択されることが望ましい。

PSS規準では，PSS_0，$PSS_1(x_1)$，$PSS_1(x_2)$ および PSS_2 の中で $PSS_1(x_1)$ が最小になれば正しい選択となる。

F規準による選択の手順は次のようになる。

4つの RSS（RSS_0，$RSS_1(x_1)$，$RSS_1(x_2)$，RSS_2）の間の関

係を図2.7 に示す。

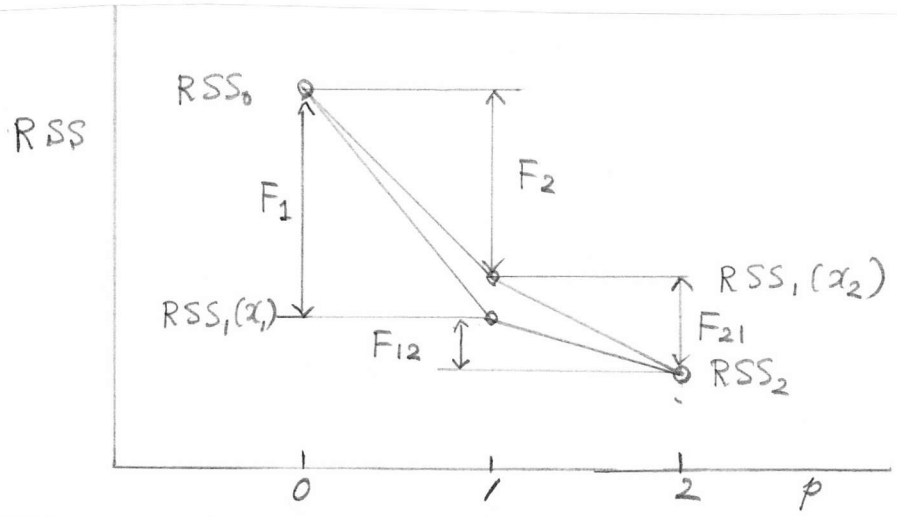

図2.7　4つのRSSとF比との関係

図中のFは，2つのRSSの差の有意性を検定するための分散比である。F_1，F_2 は単回帰の有意性を検定するためのものである。F_{12}，F_{21} は x_1 の単回帰に x_2 を追加する，または x_2 の単回帰に x_1 を追加することの有意性を検定するためのものである。これはまた，x_1 と x_2 の重回帰式から x_2 または x_1 を除くか否かの規準と考えることもできる。

F_1，F_2，F_{12}，F_{21} の値によって回帰式を選ぶ規則は次のようになる。

① F_{12}，F_{21} が共に2以上であれば x_1，x_2 の重回帰式が選ばれる。

それ以外で，

② $F_1 < F_2$ で，かつ $F_2 > 2.0$ のとき，x_1 の単回帰式が，

$F_1 > F_2$ で，かつ $F_1 > 2.0$ のとき，x_2 の単回帰式が選ばれる。

③ いずれにも当らないとき，いずれの変数も選択されない。

(2) 各変数が正規分布の場合

(2.10)式でのモデルで，zの分布がすべて正規の場合，2つの規準で選択された回帰式の割合を表2.9に示す。

なお，取りあげたαとβに対する相関係数は次のとおりである。(式(2.12))

α(またはβ)	0.5	0.7	1.0	2.0	4.0	8.0	16.0
$r=\alpha/\sqrt{1+\alpha^2}$	0.447	0.573	0.707	0.894	0.970	0.992	0.998

いくつかのα,βの組に対して，x_1, x_2, yの相関行列を次に示す。

α	1.0			2.0			8.0		
β	0.5			0.7			1.0		
x_1	1.000			1.000			1.000		
x_2	.707	1.000		.894	1.000		.992	1.000	
y	.447	.316	1.000	.573	.512	1.000	.707	.701	1.000
	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y

表2.9 選択された重回帰式の数(1000回のうち)

式(2.11)の係数		PSS				F			
α	β	—	x_1	x_2	x_1x_2	—	x_1	x_2	x_1x_2
0.0	0.0	710	130	129	31	715	131	124	30
0.0	0.5	60	784	11	145	51	795	13	141
1.0	0.5	57	727	114	102	46	746	110	98
2.0	0.5	58	608	254	80	48	624	249	79
4.0	0.5	58	514	326	102	49	513	343	95
0.0	0.7	4	840	1	155	2	845	1	152
1.0	0.7	4	824	39	133	2	836	36	126
2.0	0.7	4	734	175	87	2	742	177	78
4.0	0.7	4	597	312	87	2	603	310	85
8.0	0.7	4	510	374	112	2	508	386	104
0.0	1.0	0	844	0	156	0	847	0	153
1.0	1.0	0	844	3	153	0	847	2	151
2.0	1.0	0	824	64	112	0	827	67	106
4.0	1.0	0	661	257	82	0	674	250	76
8.0	1.0	0	551	347	102	0	550	355	95
16.0	1.0	0	486	394	120	0	494	393	113

$\alpha \neq 0$ のとき x_1 の単回帰式が選択されることが望ましい。期待されるように，x_1 が選択される割合は，β と共に増加する。また同じ β に対して α が大きくなる（r_{2y} が増す）と x_1 の選択される割合は減少し，x_2 の選択される割合が増加する。この傾向は，PSS 規準と F 規準で ほとんど 変らない。

（注）シミュレーションに当っては，§2.2 の注と同様に次の関係を利用した。

一組の乱数 z_{i1}, z_{i2}, z_{i3} $(i=1,2,\cdots,n)$ を発生する。これは，$\alpha, \beta = 0$ の x_{i1}, x_{i2}, y に相当する。

この x_1, x_2, y に対する平方和・積和，3つの残差平方和・予測平方和および4つの F 比を求める。（S_{11}, S_{22}, S_{yy}; S_{12}, S_{1y}, S_{2y}; $RSS_1(x_1)$, $RSS_1(x_2)$, RSS_2; $PSS_1(x_1)$, $PSS_1(x_2)$, PSS_2; F_1, F_2, F_{12}, F_{21}）。

同じ乱数 z_1, z_2, z_3 から $\alpha, \beta \neq 0$ の x_1, x_2, y を求め，それに対する平方和・積和等を，上記記号の肩に $*$ をつけて表わすことにする。

両者の間には次の関係がある。

$$
\begin{aligned}
S_{11}^* &= S_{11} \\
S_{22}^* &= S_{22} + \alpha^2 S_{11} + 2\alpha S_{12} \\
S_{yy}^* &= S_{yy} + \beta^2 S_{11} + 2\beta S_{1y} \\
S_{2y}^* &= S_{2y} + \alpha S_{1y} + \beta S_{12} + \alpha\beta S_{11} \\
RSS_1^*(x_1) &= RSS_1(x_1) \\
RSS_2^* &= RSS_2 \\
PSS_1^*(x_1) &= PSS_1(x_1) \\
PSS_2^* &= PSS_2 \\
F_{12}^* &= F_{12}
\end{aligned}
$$

これらから，$RSS_1^*(x_2)$, F_1^*, F_2^*, F_{21}^* が計算される。$PSS_1^*(x_2)$ だけは，x_{i2}, y_i から計算しなければならない。

（3） 各変数が非正規分布の場合

z が正規分布より裾の長い分布に従うとき，PSS 規準とF規準で選択される回帰式がどのように変るかをしらべた。非正規分布として，自由度6の t 分布をとりあげた。

式（2.11）の z_1, z_2, z_3 の分布を正規分布と t 分布のいずれをするかで，8通りの組合せがある。そのすべてについてシミュレーションをした。

シミュレーション の再現精度を見るために，1000回のシミュレーションを500回ずつの2つに分けて度数を求めた。

$\alpha=0$，$\beta=0$ の場合 PSS規準およびF規準で選択される回帰式の度数を表2.10 に示す。

表2.10 選択された重回帰式の数（z の分布の影響）

$\alpha=0$, $\beta=0$

z の分布			PSS				F			
z_1	z_2	z_3	--	x_1	x_2	x_1x_2	--	x_1	x_2	x_1x_2
N	N	N	368	54	63	15	369	54	62	15
			342	76	66	16	346	77	62	15
t	N	N	338	61	82	19	342	61	79	18
			349	61	66	24	348	63	67	22
N	t	N	348	62	75	15	366	57	63	14
			342	68	74	16	352	63	70	15
t	t	N	342	68	77	13	351	62	80	7
			338	70	76	16	351	59	76	14
N	N	t	356	62	67	15	356	61	69	14
			360	69	64	7	366	66	61	7
t	N	t	353	58	69	20	351	66	62	21
			351	71	65	13	356	65	64	15
N	t	t	347	70	73	10	355	59	77	9
			349	63	75	13	354	60	67	19
t	t	t	331	69	76	24	339	65	76	20
			328	79	74	19	346	62	72	20

また，α, β を表2.9 と同様に変えて計算したが，そのうち2組，

($\alpha=4.0$, $\beta=0.5$) と ($\alpha=20.0$, $\beta=1.0$) の結果を表2.11に示す。

いずれの場合についても，PSS規準とF規準の間には，シミュレーションの誤差以上の差が認められない。また，危険率および検出力に対する，z の分布の影響も有意ではない。

表 2.11

$\alpha = 4.0$, $\beta = 0.5$

	PSS				F			
z_1 z_2 z_3	--	x_1	x_2	x_1x_2	--	x_1	x_2	x_1x_2
N N N	21	259	165	55	19	263	168	50
	37	255	161	47	30	250	175	45
t N N	14	257	171	58	9	267	170	54
	8	263	176	53	7	267	169	57
N t N	20	278	153	49	13	295	146	46
	29	289	143	39	25	285	148	42
t t N	11	277	157	55	12	284	152	52
	5	300	139	56	4	296	151	49
N N t	65	239	139	57	52	245	149	54
	69	244	142	45	64	240	155	41
t N t	30	248	158	64	23	260	155	62
	29	256	176	39	22	249	185	44
N t t	51	262	130	57	48	259	134	59
	57	265	138	40	50	264	143	43
t t t	28	262	148	62	19	266	156	59
	29	261	163	47	29	251	168	52

(つづく)

表 2.11 (つづき)

$\alpha = 20.0$, $\beta = 1.0$

z_1	z_2	z_3	PSS				F			
			--	x_1	x_2	x_1x_2	--	x_1	x_2	x_1x_2
N	N	N	0	242	197	61	0	248	198	54
			0	244	197	59	0	246	195	59
t	N	N	0	233	193	74	0	242	184	74
			0	235	196	69	0	234	195	71
N	t	N	0	252	185	63	0	262	180	58
			0	263	184	53	0	262	183	55
t	t	N	0	253	180	67	0	254	182	64
			0	255	177	68	0	261	179	60
N	N	t	0	248	184	68	1	250	185	64
			4	249	194	53	2	254	192	52
t	N	t	0	237	187	76	0	240	190	70
			0	237	207	56	0	236	213	51
N	t	t	2	254	180	64	1	251	182	66
			1	256	188	55	0	262	187	51
t	t	t	0	249	176	75	0	243	180	77
			0	239	198	63	0	231	202	67

(4) 各変数に異常値の混入した場合

§2.4 ($p=1$ で異常値が混入した場合) と同様の検討を, $p=2$ に対して行なった。

式(2.10)に対応して, 次の8つのモデルが考えられる。

モデル番号	z_1	z_2	z_3
0	1	1	1
1	1	1	σ
2	1	σ	1
3	1	σ	σ
4	σ	1	1
5	σ	1	σ
6	σ	σ	1
7	σ	σ	σ

(x_{i1}, x_{i2}, y_i) $(i=1,2,\cdots,n)$ のうち, 1組の z について, 上表の係数を乗ずる。

モデル 0 は 対照モデルであって, 異常値は混入しない。

たとえば, モデル3は, x_2 と y に異常値が混入する。

$n=40$, $\sigma=4$ で, 異常値が1個混入した場合, $\alpha=0$, $\beta=0$; $\alpha=4.0$, $\beta=0.5$ で, 選択された回帰式の回数を表2.12に示す。

表 2.12

$\alpha = 0.0$, $\beta = 0.0$

モデル				PSS				F			
	z_1	z_2	z_3	—	x_1	x_2	x_1x_2	—	x_1	x_2	x_1x_2
0				710	130	129	31	715	131	124	30
1			*	675	142	157	26	676	141	153	31
2		*		703	127	126	44	700	125	129	46
3		*	*	675	131	149	45	493	110	316	81
4	*			682	121	148	49	700	118	146	86
5	*		*	689	158	112	41	525	318	92	65
6	*	*		693	141	134	32	705	133	131	31
7	*	*	*	613	170	164	53	394	244	250	112

$\alpha = 4.0$, $\beta = 0.5$

モデル				PSS					F				
	z_1	z_2	z_3	—	x_1	x_2	x_1x_2	$x_2+x_1x_2$	—	x_1	x_2	x_1x_2	$x_2+x_1x_2$
0				58	514	326	102	428	49	513	343	95	438
1			*	86	478	321	115	436	83	490	308	119	427
2		*		47	555	301	97	398	31	576	301	92	393
3		*	*	93	455	335	117	452	68	359	292	281	573
4	*			32	519	336	113	449	25	535	334	106	440
5	*		*	190	426	308	76	384	112	471	323	94	417
6	*	*		36	540	331	93	424	29	571	314	86	400
7	*	*	*	156	399	309	136	445	75	400	278	247	525

$\alpha=0$, $\beta=0$ の場合, x_1 と x_2 が回帰式に取り入れられる割合（$\underline{\alpha}$[注]）は, PSS規準では異常値混入のどのモデルでもほとんど影響を受けない。それに対して, F規準では, 表2.12で アンダーラインをほどこしたモデル（説明変数と目的変数の両方に異常値が入っている）で, 異常値混入の影響が顕著である。これは §2.4 の $p=1$ の場合の結果から予想されるところと一致する。

$\alpha=4.0$, $\beta=0.5$ の場合, 異常値の混入は $\underline{\alpha}$ の増加と検出力（$1-\underline{\beta}$）の低下という2つの影響を与える。x_2 が取り入れられる割合（$\underline{\alpha}$）をみると, F規準では, アンダーラインをほどこしたモデル（z_2, z_3 ; z_1, z_2, z_3）で増加しているのに対して, PSS規準は余り影響を受けない。x_1, x_2 のいずれも選択されない（表2.12の一欄）割合（$\underline{\beta}$）はPSS規準の方がF規準より大きい。

（注）第1種および第2種の誤りの確率 α, β をモデルの回帰係数 α, β と区別するために, 確率にはアンダーラインを引いてある。