4. 重回帰分析

4.1 重回帰式とは

複数個の変数 $x_1, x_2, \cdots, x_p$ に基づいて，ひとつの変数 y を推測する式

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon \qquad (4.1)$$

は，y の $x_1, x_2, \cdots, x_p$ に対する重回帰式とよばれる．ここで $x_1, x_2, \cdots, x_p$ を**説明変数**(独立変数)，y を**目的変数**(従属変数) という．目的変数 y は計測がむずかしく，計測の容易な $x_1, x_2, \cdots, x_p$ から y を推測する場合や，$x_1, x_2, \cdots, x_p$ の値から y を予測する場合に用いられる．

データは表 4.1 のように書き表せる．j 番目のデータは，対になった $p+1$ 個の値 $(x_{1j}, x_{2j}, \cdots, x_{pj}, y_j)$ として与

表 4.1　大きさ N の標本

標準番号	1	2	・・・	j	・・・	N
x_1	x_{11}	x_{12}	・・・	x_{1j}	・・・	x_{1N}
x_2	x_{21}	x_{22}	・・・	x_{2j}	・・・	x_{2N}
⋮	⋮	⋮	・・・	⋮	・・・	⋮
x_p	x_{p1}	x_{p2}	・・・	x_{pj}	・・・	x_{pN}
y	y_1	y_2	・・・	y_j	・・・	y_N

えられている．計測された N 個のデータから，重回帰式を決める係数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ の推定値 $b_0, b_1, b_2, \cdots, b_p$ を求めることになる．この推測値から y の $x_1, x_2, \cdots, x_p$ に対する重回帰式

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

を得る．次に重回帰分析の例をいくつか述べよう．

重回帰分析の例

歯の咬耗度から年齢を推定したい．そこで一本一本の歯の摩耗の程度を数量化し，p 本の歯の値 $x_1, x_2, \cdots, x_p$ から年齢 y を推測する重回帰モデル

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon$$

を考える(§4.3)．

工業製品卸売物価指数 y を輸入価格 x_1，単位労働コスト x_2，製造業生産者在庫率 x_3 によって推測する式

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\varepsilon$$

は重回帰モデルである．

ある種族の手の指の長さ y は，年齢 x とどのような関係にあるかを調べたい．回帰式

$$y=\beta_0+\beta_1 x+\varepsilon$$

は，ある狭い年齢の範囲では有効であるが，年齢の範囲が広がると，推測の精度は悪くなる．そこで多項式

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_p x^p+\varepsilon$$

による当てはめを考える．このモデルは

$$x_1=x,\ x_2=x^2,\ \cdots,\ x_p=x^p$$

とおくことにより重回帰式モデルに帰着する．

実質GNP y' を

z_1：全産業稼動率指数×民間設備資本ストック

z_2：全産業労働時間指数×全産業就業者数

z_3：生活基盤社会資本ストック／民間設備資本ストック

z_4：ヴィンテージ係数(資本の新規度)

z_5：タイムトレンド

で説明する式

$$y'=\beta_0 z_1^{\beta_1} z_2^{\beta_2} z_3^{\beta_3} z_4^{\beta_4} z_5^{\beta_5} e^{\varepsilon}$$

を考える．この場合には，両辺に対数をとると

$$\log y'=\beta_0+\beta_1 \log z_1+\beta_2 \log z_2+\beta_3 \log z_3+\beta_4 \log z_4+\beta_5 \log z_5+\varepsilon$$

となり

$$y=\log y',\quad x_1=\log z_1,\quad x_2=\log z_2,\quad x_3=\log z_3,\quad x_4=\log z_4,\quad x_5=\log z_5$$

とおくと，上式は重回帰モデル

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\varepsilon$$

になる．

以後この章では観測値 x と残差項 ε は次の仮定を満たすものとする.

1) x の値 $(x_{1j}, x_{2j}, \cdots, x_{pj})$, $j=1, 2, \cdots, N$ は固定された変数値（非確率変数）とする.

2) ε_j $(j=1, 2, \cdots, N)$ の期待値は 0 である:

任意の x の値 $(x_{1j}, x_{2j}, \cdots, x_{pj})$ に対し, $y_j=\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\cdots+\beta_p x_{pj}+\varepsilon_j$ としたとき, y_j の期待値は $\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\cdots+\beta_p x_{pj}$ である. これは, y_j の期待値が回帰平面上にあることを意味している.

3) ε_j $(j=1, 2, \cdots, N)$ の分散は等しく一定である:

この仮定は ε_i の分散が等しければよいのであって, 同一分布に従う必要はない.

4) ε_j と ε_k $(j \neq k,\ j, k=1, 2, \cdots, N)$ は無相関である.

残差 ε に正規分布を仮定すると, 正規分布の性質から, 条件 3 は ε_j $(j=1, 2, \cdots, N)$ は同じ正規分布に従うことを意味し, 条件 4 は ε_j と ε_k とが互いに独立であることを意味する. 推定量の区間推定や仮説検定を論じる際には, この正規分布の仮定を入れる.

重回帰式はいろいろな分野でよく利用されている分析モデルである. まず初めに, データから重回帰式を決める係数を, どのような考え方で求めるのか知るために, 次節では説明変数が一つだけの場合について説明することにする.

4.2　1 変数の場合の回帰式

対になった N 個のデータを

標準番号	1	2	・・・	j	・・・	N
x_1	x_{11}	x_{12}	・・・	x_{1j}	・・・	x_{1N}
y	y_1	y_2	・・・	y_j	・・・	y_N

で表す．回帰モデルは

$$y_j=\beta_0+\beta_1 x_{1j}+\varepsilon_j \qquad (j=1,2,\cdots,N) \tag{4.2}$$

と書ける．$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_N$ は互いに無相関で，平均 0，分散 σ^2 である．つまり，ε_j の平均と分散は共通で等しいことを前提としている．この ε_j は**残差**，あるいは**回帰からの偏差**とよばれ，(4.2) 式から

$$\varepsilon_j=y_j-(\beta_0+\beta_1 x_{1j}) \qquad (j=1,2,\cdots,N) \tag{4.3}$$

と表される．

N 個のデータから，残差の平方和

$$E=\varepsilon_1{}^2+\varepsilon_2{}^2+\cdots+\varepsilon_N{}^2 \tag{4.4}$$

が最小になるように，係数 β_0 と β_1 を決める．(4.4) 式は残差平方和とよばれ

$$E=\{y_1-(\beta_0+\beta_1x_{11})\}^2+\{y_2-(\beta_0+\beta_1x_{12})\}^2+\cdots+\{y_N-(\beta_0+\beta_1x_{1N})\}^2 \tag{4.5}$$

と書ける．上式の j 番目の平方は，展開して β_0 についてまとめると

$$\begin{aligned}\{y_j-(\beta_0+\beta_1x_{1j})\}^2&=y_j{}^2-2(\beta_0+\beta_1x_{1j})y_j+(\beta_0+\beta_{1j})^2\\&=\beta_0{}^2-2(y_j-\beta_1x_{1j})\beta_0+\beta_1{}^2x_{1j}{}^2-2\beta_1x_{1j}y_j+y_j{}^2\end{aligned} \tag{4.6}$$

β_0 の 2 次式になっている．これを N 個 $(j=1,2,\cdots,N)$ 加え合わせた E も，やはり β_0 の 2 次式である．

2次関数の最小問題と考えて解き，Eを最小とする β_0 の値を b_0, β_1 の値を b_1 で表せば

$$b_0=\bar{y}-b_1\bar{x}_1$$

$$b_1=\frac{(x_{11}-\bar{x}_1)(y_1-\bar{y})+(x_{12}-\bar{x}_1)(y_2-\bar{y})+\cdots+(x_{1N}-\bar{x}_1)(y_N-\bar{y})}{(x_{11}-\bar{x}_1)^2+(x_{12}-\bar{x}_1)^2+\cdots+(x_{1N}-\bar{x}_1)^2}$$

を得る．ここで，$\bar{x}, \bar{y}$ は x の平均，y の平均のことである．よってデータから推定した回帰式は

$$\hat{y}=b_0+b_1x_1$$

となる．

出産した人に「赤ちゃんは何gでしたか」ときくのはよく耳にする．しかし身長をきく人はまれである．そこで体重を知ったとき，身長yのおおよその値を推測する回帰式を求めよう．新生児45人を抽出し，体重と身長を計測したところ，次のような数値を得たとする．

新生児の体重(g)と身長(cm)

体重x_1	2880	2725	2365	2185	1560	2155	2250	2730	2310	2670	2430	2010
身長y	51.0	48.5	45.5	44.4	43.1	45.9	46.1	48.4	43.8	46.6	44.7	52.0
体重x_1	2690	3525	3510	3500	4300	3715	3150	3009	4100	3065	3345	3675
身長y	47.5	52.0	52.3	54.0	51.3	53.2	53.5	48.5	53.5	47.4	50.4	51.4
体重x_1	4040	2790	1930	1930	1845	1780	1900	2340	2155	2325	3600	3330
身長y	54.8	48.0	46.5	44.8	45.5	44.5	48.5	48.0	47.6	46.0	50.1	48.6
体重x_1	3360	3400	2365	3365	1955	1950	2835	2225	1930			
身長y	50.3	48.1	49.0	49.5	44.8	45.7	48.1	48.0	43.9			

これより回帰係数 b_1 は

$$b_1=\frac{(2880-2738)\,(51.0-48.3)+\cdots+(1930-2738)\,(43.9-48.3)}{(2880-2738)^2+\cdots+(1930-2738)^2}$$

$$=\frac{142\times 2.7+\cdots+(-808)\times(-4.4)}{142^2+\cdots+(-4.4)^2}=0.00344$$

であり，定数項 b_0 は

$$b_0=\bar{y}-b_1\,\bar{x}_1=48.3-0.00344\times 2738=38.9$$

となる．よって求める回帰直線は

$$\hat{y}=0.00344\,x_1+38.9$$

である．データと回帰直線をかいたのが図 4.1 である．体重が 3,350 g であれば

$$\hat{y}=0.00344\times 3350+38.9=50.4$$

となり，50 cm ほどと推察することになる．

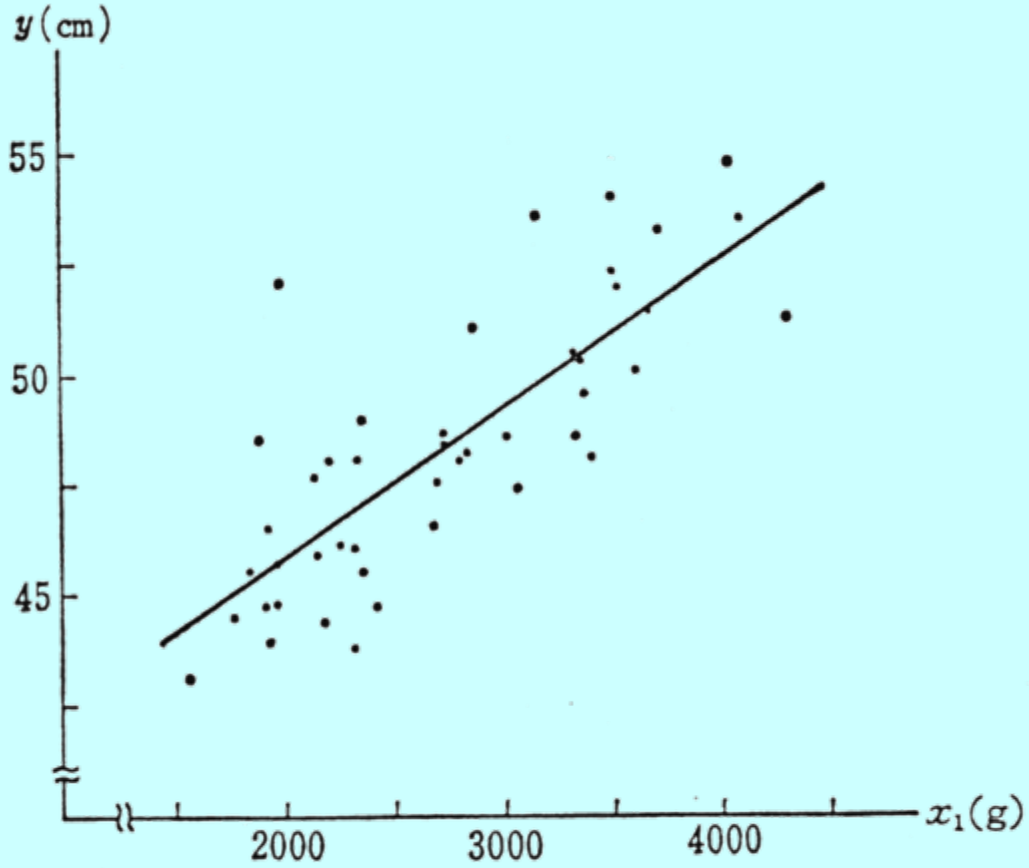

図 4.1　新生児の身長 y の体重 x_1 に対する回帰直線

4.3　多変数の重回帰分析

初めに説明変数が二つ x_1, x_2 の場合を述べる．重回帰式は

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$$

であり，データは表 4.2 のように表される．

表 4.2　2 変数の標本

標本番号	1	2	・・・	j	・・・	N
x_1	x_{11}	x_{12}	・・・	x_{1j}	・・・	x_{1N}
x_2	x_{21}	x_{22}	・・・	x_{2j}	・・・	x_{2N}
y	y_1	y_2	・・・	y_j	・・・	y_N

重回帰モデル

$$y_j=\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}+\varepsilon_j \qquad (j=1,2,\cdots,N) \tag{4.7}$$

の残差 ε_j は互いに無相関で，平均は 0，分散 σ^2 とする．(4.7) 式から

$$\varepsilon_j=y_j-(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j}) \qquad (j=1,2,\cdots,N) \tag{4.8}$$

であり，残差平方和は

$$\begin{aligned} E&=\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_N^2 \\ &=\{y_1-(\beta_0+\beta_1 x_{11}+\beta_2 x_{21})\}^2+\{y_2-(\beta_0+\beta_1 x_{12}+\beta_2 x_{22})\}^2 \\ &\qquad +\cdots+\{y_N-(\beta_0+\beta_1 x_{1N}+\beta_2 x_{2N})\}^2 \end{aligned} \tag{4.9}$$

と書ける．これを最小にする $\beta_0, \beta_1, \beta_2$ の値 b_0, b_1, b_2 は，1変数の場合ほどは容易でないが，数学的に解くことができ，次のようになる．

$$\begin{aligned} b_0&=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2 \\ b_1&=a^{11}a_{1y}+a^{12}a_{2y} \\ b_2&=a^{12}a_{1y}+a^{22}a_{2y} \end{aligned} \tag{4.10}$$

ここで，$\bar{y}$ は y の平均，$\bar{x}_1$ は x_1 の平均，$\bar{x}_2$ は x_2 の平均であり，分散共分散は

$$
\begin{aligned}
a_{11} &= (x_{11}-\bar{x}_1)^2+(x_{12}-\bar{x}_1)^2+\cdots+(x_{1N}-\bar{x}_1)^2 \\
a_{22} &= (x_{21}-\bar{x}_2)^2+(x_{22}-\bar{x}_2)^2+\cdots+(x_{2N}-\bar{x}_2)^2 \qquad (4.11) \\
a_{yy} &= (y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_N-\bar{y})^2 \\
a_{12} &= (x_{11}-\bar{x}_1)(x_{21}-\bar{x}_2)+(x_{12}-\bar{x}_1)(x_{22}-\bar{x}_2)+\cdots+(x_{1N}-\bar{x}_1)(x_{2N}-\bar{x}_2) \\
a_{1y} &= (x_{11}-\bar{x}_1)(y_1-\bar{y})+(x_{12}-\bar{x}_1)(y_2-\bar{y})+\cdots+(x_{1N}-\bar{x}_1)(y_N-\bar{y}) \\
a_{2y} &= (x_{21}-\bar{x}_2)(y_1-\bar{y})+(x_{22}-\bar{x}_2)(y_2-\bar{y})+\cdots+(x_{2N}-\bar{x}_2)(y_N-\bar{y})
\end{aligned}
$$

また

$$
a^{11}=\frac{a_{22}}{(a_{11}a_{22}-a_{12}{}^2)}, \qquad a^{12}=\frac{-a_{12}}{(a_{11}a_{22}-a_{12}{}^2)}, \qquad a^{22}=\frac{a_{22}}{(a_{11}a_{22}-a_{12}{}^2)}
$$

である．

データから推定した回帰式は

$$\hat{y}=b_0+b_1x_1+b_2x_2 \tag{4.12}$$

となる．b_0, b_1, b_2 は**偏回帰係数**とよばれる．　このときの残差平方和(最小値)は

$$E=(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+\cdots+(y_N-\hat{y}_N)^2$$

$$=a_{yy}-b_1a_{1y}-b_2a_{2y}$$

と書き表される．

歯の咬耗度から年齢を推定する問題を考えよう．下顎の第2大臼歯の咬耗度を x_1，犬歯の咬耗度を x_2 とする．咬耗度の数量化は専門家の意見と，データのいろいろな角度からの考察に基づいて，エナメル質の摩耗の局面が狭い範囲で独立している場合は1，エナメル質の大部分が摩耗している場合は2，象牙質の摩耗が進んで露出している場合は3，象牙質のかなりの部分が広くあるいは強く摩耗している場合は4という数値を割当てた．表4.3は32人についてのデータである．

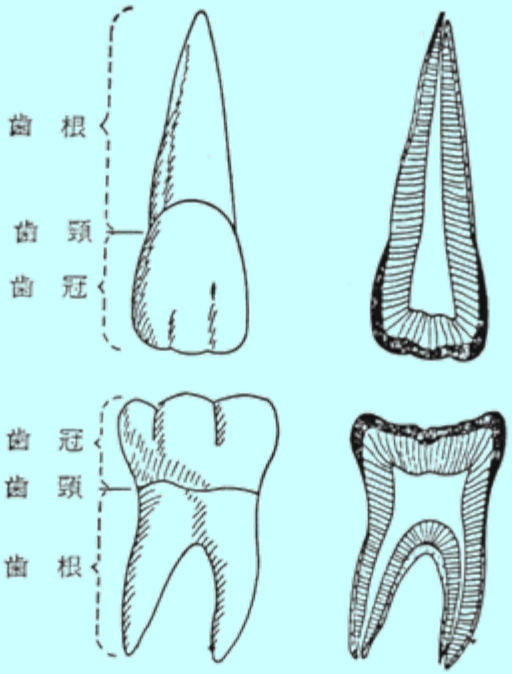

図 4.2　歯の外景と内景(断面図ではエナメル質は黒，象牙質は線影，セメント質は点影で表してある)

表 4.3　大臼歯(x_1)と犬歯(x_2)の咬耗度と年齢(y)

x_1	1	1	1	1	1	2	1	1	2	2	2	2	2	2	2	3	2	2
x_2	3	1	2	1	3	1	2	3	1	3	2	3	3	3	2	3	3	3
y	21	22	23	25	26	27	28	30	31	32	33	34	36	37	38	40	41	44

x_1	4	2	2	3	4	2	4	3	3	4	3	3	3	4
x_2	3	3	3	4	1	3	4	2	3	3	4	4	4	4
y	44	45	46	47	48	50	51	52	53	54	56	57	58	59

図 4.3 は横軸に摩耗の程度を，縦軸に年齢をとって書いたものである．

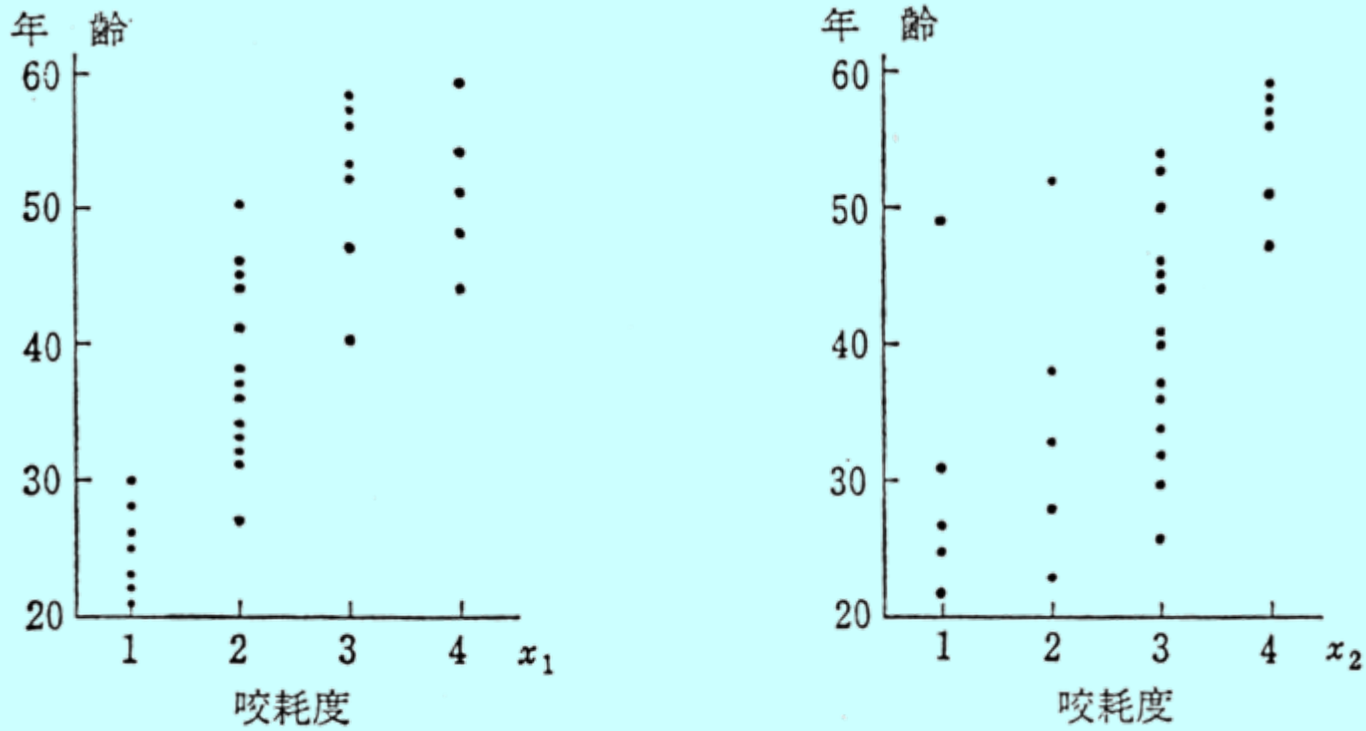

図 4.3 咬耗度と年齢

重回帰式を求めるために，データを（4.11）式に代入して平均，分散共分散等を計算すると

$$\bar{x}_1=2.3 \qquad a_{11}=39.5 \qquad a_{12}=14.4 \qquad a_{1y}=324.6$$

$$\bar{x}_2=2.7 \qquad\qquad\quad a_{22}=26.5 \qquad a_{2y}=225.8$$

$$\bar{y}=40.3 \qquad a_{yy}=4194$$

$$a^{11}=0.0316 \qquad a^{12}=-0.017 \qquad a^{22}=0.047$$

となり，（4.10）式から偏回帰係数

$$b_0=11.6 \qquad b_1=6.4 \qquad b_2=5.1$$

を得る．よって，歯の咬耗度から年齢を推定する重回帰式は

$$\hat{y}=11.6+6.4\,x_1+5.1\,x_2$$

となる．この式に基づいて計算した推定年齢を表4.4に示す．

表 4.4　実年齢(y)と推定年齢($\hat{y}$)

y	21	22	23	25	26	27	28	30	31	32	33	34
$\hat{y}$	33.2	23.1	28.1	23.1	33.2	29.4	28.1	33.2	29.4	39.6	34.5	39.6
y	36	37	38	40	41	44	44	45	46	47	48	50
$\hat{y}$	39.6	39.6	34.5	45.8	39.6	39.6	52.3	39.6	39.6	51.0	42.2	39.6
y	51	52	53	54	56	57	58	59				
$\hat{y}$	57.4	40.9	45.9	52.3	51.0	51.0	51.0	57.4				

年齢yを横軸にとり推定年齢yを縦軸にとって，表4.4の値$(y, \hat{y})$に基づいて32個の点をかいたのが図4.4である．観測値yと回帰による推定値$\hat{y}$との相関係数Rは

$$R=\frac{(y_1-\bar{y})(\hat{y}_1-\bar{y})+\cdots+(y_N-\bar{y})(\hat{y}_N-\bar{y})}{\sqrt{(y_1-\bar{y})^2+\cdots+(y_N-\bar{y})^2}\sqrt{(\hat{y}_1-\bar{y})^2+\cdots+(\hat{y}_N-\bar{y})^2}}$$

であり，これから

$$R=0.875$$

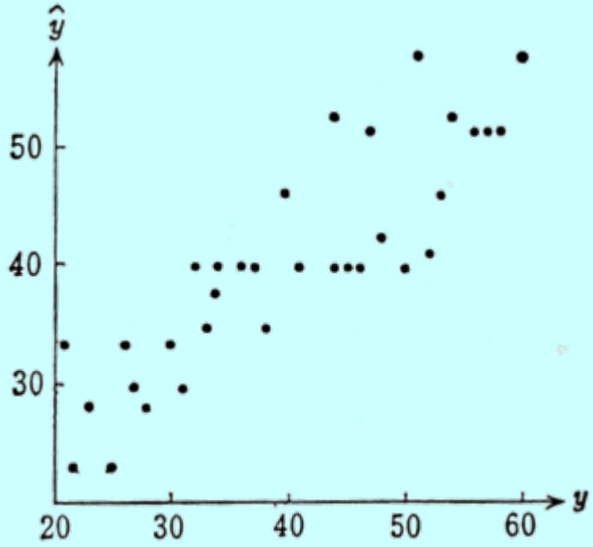

図 4.4　実際年齢 y と推定年齢 $\hat{y}$

相関係数 R は y と $\hat{y}$ との直線的関連性の強さであり，その絶対値が1に近ければ近いほど，点は直線のまわりに散らばっていることになり，予測の精度がよいことになる．この R を回帰分析では**重相関係数**という．この重相関係数の2乗 R^2 を**決定係数**あるいは**寄与率**という．重相関係数については，§4.5で述べる．

説明変数が二つの場合で述べたが，一般に説明変数が p 個の場合も同様である．この場合の重回帰モデルは

$$y_j=\beta_0+\beta_1x_{1j}+\beta_2x_{2j}+\cdots+\beta_px_{pj}+\varepsilon_j \quad (j=1,2,\cdots,N) \qquad (4.13)$$

と書ける．これまでと同じように残差

$$\varepsilon_j=y_j-(\beta_0+\beta_1x_{1j}+\beta_2x_{2j}+\cdots+\beta_px_{pj}) \qquad (j=1,2,\cdots,N) \quad (4.14)$$

を考え，残差平方和 $\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_N^2$ を最小にするような $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ の値 $b_0,b_1,b_2,\cdots,b_p$ を求める．それより重回帰式は，

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$$

となる．実際にデータから偏回帰係数 $b_0,b_1,b_2,\cdots,b_p$ を求めるには，２変数の場合と同様に $x_1,x_2,\cdots,x_p,y$ のそれぞれの平均や分散共分散などを求める必要があるが，詳しいことは省略する．

付録 データ6から，28本の歯すべてを用いたとき，年齢yを推定する重回帰式は

$$\hat{y}=20.0+1.30x_1-0.09x_2+2.58x_3-0.82x_4+2.39x_5-0.16x_6$$
$$+1.12x_7+0.41x_8+0.25x_9+0.86x_{10}+0.29x_{11}+0.04x_{12}$$
$$+0.72x_{13}+2.75x_{14}+0.86x_{15}-0.80x_{16}+1.14x_{17}+2.83x_{18}$$
$$-1.64x_{19}-0.69x_{20}+1.27x_{21}-0.45x_{22}-0.12x_{23}+0.20x_{24}$$
$$-0.04x_{25}-0.49x_{26}+1.00x_{27}+0.72x_{28}$$

である．そのときの重相関係数Rは

$$R=0.891$$

であり，決定係数は0.794になる．実際には，すべての歯を用いて年齢を推定することは賢明でなく，28本の中の一部分を利用して推測式をつくることになる．この説明変数の選択は重要な問題であり，これに関する詳しいことは§4.8で述べる．

4.8　説明変数の選択

§4.3では歯の咬耗度による年齢推定の問題を考え，28本すべてを用いたときの重回帰式を示した．しかしながら，この場合28変数 $x_1, x_2, \cdots, x_{28}$ すべてを用いるのは賢明ではない．変数を多く取入れて重回帰式をつくれば，残差平方和は小さくなり，確かに重相関係数は大きくなる．その意味では予測の精度は上っているようにみえるが，他方データの関数である偏回帰係数 b_j の分散は大きくなり，重回帰式の安定性は悪くなる．説明変数すべてを用いた場合と，そのうちの一部を用いた場合とを比較して，予測の精度がそれほど変わっていなければ，説明変数の一部を用いたときの重回帰式で十分間にあい，またその方が重回帰式は安定している．私達はできるかぎり少数個の互いに相関の低い説明変数を用いて，予測精度が全変数を用いたときに比べてあまり変わらない重回帰式を求めたいのである．この節ではたくさんある変数のなかから，どのような変数の組合せを選ぶかという問題を扱う．変数選択の方法としては代表的な**変数増減法**と**変数減増法**について述べる．

変数選択において，まず初めに目的変数 y と説明変数 x_i との相関係数の大きさをみることは，相関係数が大きければ大きいほど x_i は y に関する情報をもっていることであるから参考にはなる．説明変数の間で相互に関連しあっていることを考えると，これだけでは不十分である．たとえば，y と x_i および y と x_j との相関係数は大きく，また x_i と x_j とは高い相関をもっているとする．この場合には x_i の y に関してもっている情報と，x_j の y に関してもっている情報とはよく似たものであり，y の説明変数としては x_i か x_j のいずれかを用いることで十分であろう．これらのことから y との相関係数の大きさだけではなく，説明変数相互間の関連性の強さまで考慮に入れて $x_1, x_2, \cdots, x_p$ のなかからどのような組合せを選ぶかを決めることになる．

重回帰式

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

において，残差 ε は平均 0，分散 σ^2 の正規分布に従うとする．これは y が平均 $\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$ の正規分布に従うことを意味する．いま表 4.1 のようにデータが得られているとしよう．偏回帰係数 β_i の推定値 b_i は，正規分布に従う $y_1, y_2, \cdots, y_N$ の 1 次式で表されるから，正規分布に従い

$$F=\frac{(b_i-\beta_i)^2}{b_i \text{ の分散}}$$

は自由度 $(1, N-p-1)$ の F 分布に従うことを示すことができる．いま，説明変数 $x_1, x_2, \cdots, x_p$ のなかから q 個 $x_{(1)}, x_{(2)}, \cdots, x_{(q)}$ が選ばれているとする．このとき，この q 個の説明変数に基づく重回帰式を

$$y=\beta_{(0)}+\beta_{(1)}x_{(1)}+\cdots+\beta_{(q)}x_{(q)}$$

で表す．データから計算した式を

$$y=b_{(0)}+b_{(1)}x_{(1)}+\cdots+b_{(q)}x_{(q)}$$

とする．もし $x_{(i)}$ の係数 $b_{(i)}$ が小さいならば，$x_{(i)}$ は y を説明する変数として役に立っていないとして取除く．この係数が大きいならば $x_{(i)}$ は除去しない．

偏回帰係数 $b_{(i)}$ が小さいとか大きいとかは

$$F=\frac{b^2{}_{(i)}}{b_{(i)}\text{ の分散}}$$

の値で決める．これは自由度 $(1, N-q-1)$ の F 分布に従う．説明変数を除去する際の F の基準値を F_{OUT} で表すと，F 値が F_{OUT} より小さいときは $x_{(i)}$ は除去する．$b_{(1)}$ の F 値，$b_{(2)}$ の F 値，…，$b_{(q)}$ の F 値と F_{OUT} とをそれぞれ比較して，F_{OUT} より小さいのが二つ以上あるときには，いちばん小さい F 値に対応する変数のみを除去する．これが説明変数を除去する規則である．

説明変数 $x_{(1)}, x_{(2)}, \cdots, x_{(q-1)}$ が選ばれていて，残りの変数のなかの一つを追加したときに，決定係数が最大になるような変数 $x_{(q)}$ について考えよう．このとき，偏回帰係数 $b_{(q)}$ が大きいならば $x_{(q)}$ は y を説明する変数として役に立っているとして追加する．また，小さいときには追加しない．前と同じで，この場合も偏回帰係数が大きいとか小さいとかは

$$F=\frac{b^2{}_{(q)}}{b_{(q)}\text{ の分散}}$$

の値で決める．これも自由度 $(1, N-q-1)$ の F 分布に従う．説明変数を追加する際の F の基準値を F_{IN} で表すと，F 値が F_{IN} より大きければ $x_{(q)}$ を取入れる．この変数を除去する基準値 F_{OUT} と追加する基準値 F_{IN} とは，データ以外のほかの理由による特別な要請がないならば，ともに 2.0 にするのがよい．このことは標本数 N が大きいときは，F 分布の有意水準 16,7% の点を F_{OUT}, F_{IN} の値としていることになる．$F_{OUT}=F_{IN}=2.0$ がよいというのは情報量基準から導びくことができるが，それについての詳しいことは次節で述べる．

歯の咬耗度による年齢推定を例にして説明しよう．その変数選択を説明する前に，データの特徴について詳しく説明することにする．データは次のような分類番号で与えられている．前に述べたように，分類1はエナメル質の局面が狭い範囲で独立している場合，2はエナメル質の大部分が摩耗している場合，3は象牙質の摩耗が進んで部分的に露出している場合，4は象牙質のかなりの部分が広くあるいは強く摩耗している場合，5は欠如している場合である．

それぞれの歯の咬耗度と年齢との関連性を把握するために，咬耗度を横軸にとって書いたのが図4.9である．

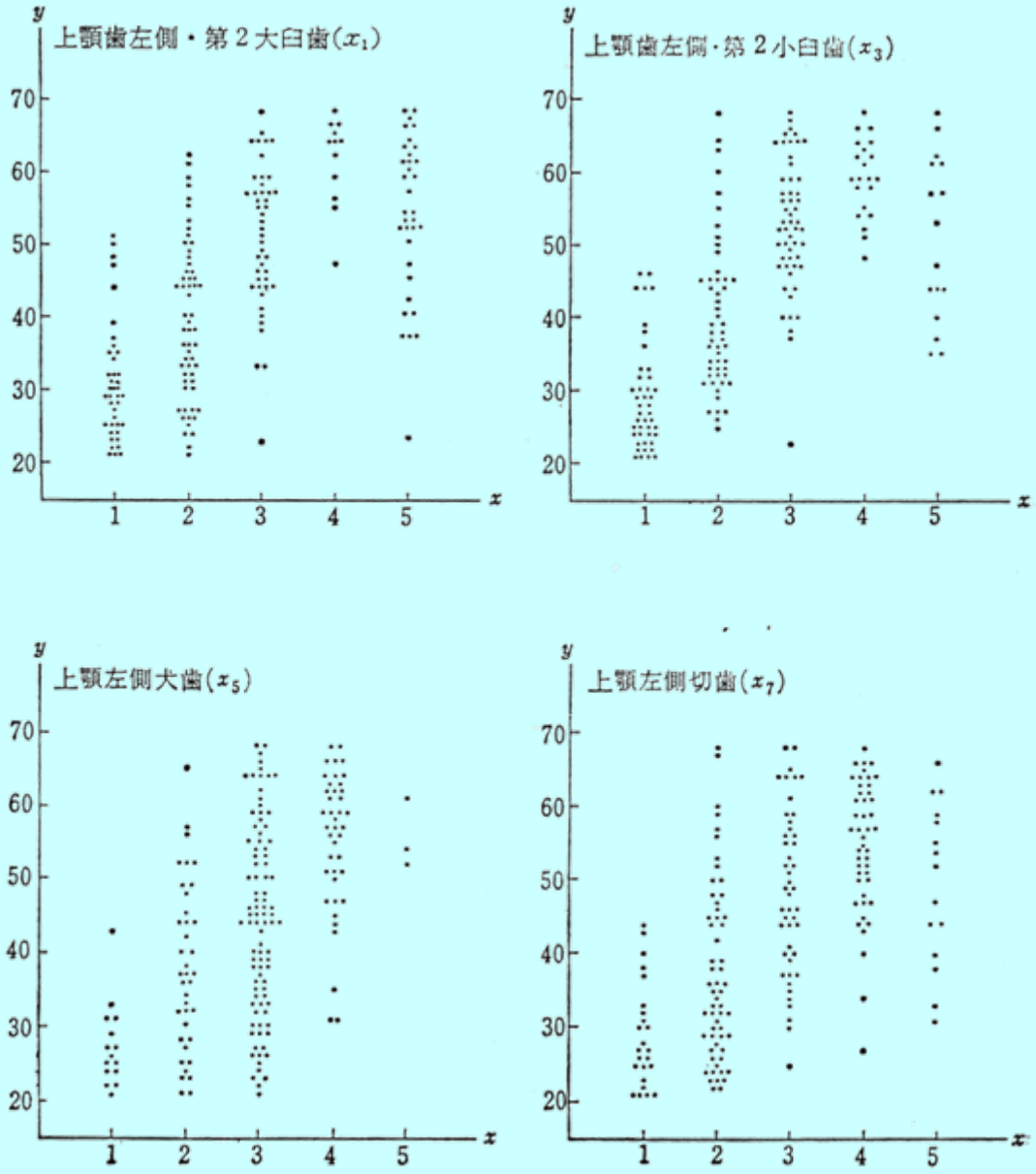

図 4.9　歯の摩耗度(x)と年齢(y)

回帰式による年齢推定という立場からこの散布図をみた場合に，５段階の分類をそのまま１から５という形で数量化し，重回帰分析を適用してよいかという疑問にぶつかる．とくに５という段階は，除去したほうが全体として直線的な関係があるように思える．しかしながら，分類５の入っているデータは多く，取除くことはできない．この欠測値をどう補うかが問題となる．欠測値を補うのに，分類１から４を数量化したあとに，最小２乗法を用いた補い方などが考えられるが，いろいろな観点から検討した末に，ここでは４という段階と同一視して分析することにした．実際には段階３の歯がなんらかの原因で欠如して５になるか，あるいは段階４であった歯が欠如して５になるかは紙一重であり，数量化は3.5がよいのではという専門家の意見もあった．このような経過を経て，分類１には1，分類２には2，分類３には3，分類４，５には４という数値を割りふることにした．数量化１類による数量化についても文献［9］のなかで議論されている．

この数量化に基づいて，28 変数と y とによる相関行列を表 4.7 に示す．データ数 N は 189 である．

年齢 y との相関係数をみると，最も相関の高いのは上顎左側小臼歯（x_3）の 0.73，上顎右側第 2 大臼歯 x_{14} の 0.70，下顎右側第 1 小臼歯 x_{18} の 0.70，以下略して，大臼歯 x_1 の 0.68，小臼歯 x_{17} の 0.65，切歯 x_7 の 0.64，小臼歯 x_{12}の 0.63，小臼歯 x_{11} の 0.62 と続き，そのほかの歯はほぼ 0.55 程度である．なお，上顎歯の方に比較的相関の高いものが集まっている．

変数選択の代表的方法は，変数増減法と変数減増法である．変数増減法は，最初に y との相関係数が最大のものを選び，これを出発点として変数の追加と除去という上記の手続きを繰返し行う．また，変数減増法は p 変数のすべてを取入れた重回帰式から出発し，変数の除去と追加の手続きを繰返し行う．

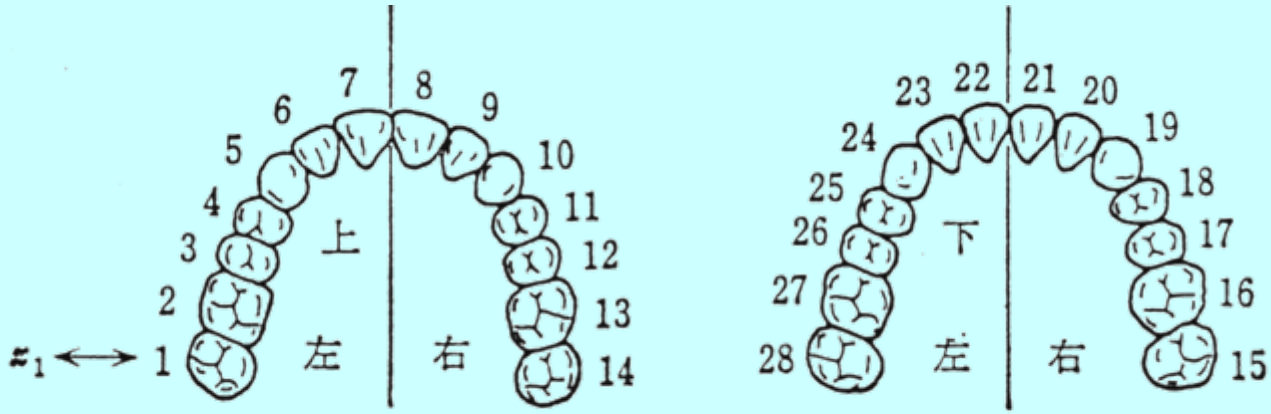

図 4.10　歯と変数名との対応関係

表 4.7　相関行列

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	y
x_1	1																												
x_2	.65	1																											
x_3	.61	.56	1																										
x_4	.48	.46	.65	1																									
x_5	.41	.44	.43	.50	1																								
x_6	.40	.46	.52	.49	.46	1																							
x_7	.46	.48	.55	.45	.51	.67	1																						
x_8	.46	.43	.50	.46	.46	.60	.81	1																					
x_9	.38	.43	.47	.46	.38	.74	.57	.60	1																				
x_{10}	.45	.46	.48	.49	.67	.56	.53	.46	.49	1																			
x_{11}	.60	.54	.61	.57	.45	.52	.50	.41	.41	.55	1																		
x_{12}	.58	.50	.70	.55	.38	.53	.48	.43	.43	.48	.65	1																	
x_{13}	.52	.51	.43	.42	.30	.39	.30	.29	.28	.32	.40	.49	1																
x_{14}	.64	.48	.56	.44	.39	.39	.46	.44	.39	.40	.49	.53	.46	1															
x_{15}	.40	.29	.42	.45	.45	.37	.41	.42	.31	.38	.43	.37	.30	.48	1														
x_{16}	.38	.33	.35	.29	.29	.31	.30	.30	.28	.22	.34	.36	.34	.37	.49	1													
x_{17}	.53	.47	.61	.54	.39	.52	.50	.45	.42	.43	.55	.50	.36	.41	.52	.39	1												
x_{18}	.55	.54	.60	.52	.44	.58	.57	.53	.47	.51	.52	.55	.36	.47	.44	.34	.67	1											
x_{19}	.48	.44	.52	.57	.72	.49	.57	.50	.38	.63	.45	.42	.28	.44	.35	.21	.50	.57	1										
x_{20}	.45	.43	.56	.53	.51	.71	.67	.59	.58	.56	.52	.53	.29	.39	.34	.21	.54	.62	.65	1									
x_{21}	.42	.40	.52	.49	.41	.69	.65	.60	.56	.53	.49	.53	.32	.37	.36	.18	.47	.53	.50	.72	1								
x_{22}	.38	.38	.50	.41	.34	.57	.57	.49	.54	.44	.48	.43	.26	.37	.29	.09	.45	.51	.43	.66	.73	1							
x_{23}	.42	.43	.54	.47	.40	.58	.57	.54	.65	.51	.53	.48	.33	.47	.37	.26	.48	.60	.49	.72	.65	.64	1						
x_{24}	.43	.42	.45	.50	.62	.54	.54	.49	.48	.71	.53	.41	.27	.41	.38	.20	.48	.51	.70	.63	.49	.50	.54	1					
x_{25}	.48	.42	.58	.47	.32	.48	.44	.43	.35	.42	.54	.53	.35	.40	.37	.24	.56	.62	.44	.53	.44	.46	.53	.50	1				
x_{26}	.52	.42	.57	.47	.31	.44	.45	.47	.39	.50	.55	.48	.29	.41	.43	.32	.58	.60	.40	.48	.41	.41	.51	.50	.69	1			
x_{27}	.43	.37	.51	.38	.22	.32	.34	.30	.27	.32	.44	.44	.23	.47	.44	.57	.44	.43	.29	.30	.29	.24	.34	.28	.37	.46	1		
x_{28}	.46	.35	.51	.47	.47	.29	.43	.43	.30	.41	.46	.39	.31	.44	.58	.50	.53	.45	.43	.34	.35	.29	.38	.40	.45	.51	.54	1	
y	.68	.59	.73	.57	.57	.57	.64	.59	.51	.59	.62	.63	.49	.70	.58	.41	.65	.70	.56	.57	.57	.50	.57	.54	.56	.56	.53	.59	1

変数選択に際しては $F_{IN}=F_{OUT}=2.0$ で年齢推定式をつくる．この基準値で変数選択を行うと，28 個が 10 個まで減少する．変数増減法で最後の段階で変数の入れ換わりが一度あったほかは，一度取込まれた変数が再び追い出されたことも，また変数減増法でいえば除去された変数が再び取込まれたこともなかった．変数選択の途中のステップを省略し，変数の取入れられた順序だけを記すと次のようになる．

変数増減法　$x_3 \rightarrow x_{14} \rightarrow x_{18} \rightarrow x_5 \rightarrow x_{28} \rightarrow x_7 \rightarrow x_1 \rightarrow x_{15} \rightarrow x_{19} \rightarrow x_{21} \rightarrow x_{17} \rightarrow (x_{15})$

変数減増法　$x_{12} \rightarrow x_{25} \rightarrow x_2 \rightarrow x_{23} \rightarrow x_6 \rightarrow x_{24} \rightarrow x_9 \rightarrow x_{11} \rightarrow x_{22} \rightarrow x_8 \rightarrow x_{26} \rightarrow x_4 \rightarrow$

$x_{10} \rightarrow x_{20} \rightarrow x_{16} \rightarrow x_{13} \rightarrow x_{15}$

また全変数を用いた場合と最終段階で選ばれた 10 個の変数の場合とについて，回帰係数と F 値を表 4.8 に与え，簡単な説明を加えることにする．

表 4.8　28 変数での重回帰分析と変数選択によって選ばれた 10 変数による重回帰分析の結果

全変数 28 個の場合						最終段階		
変数番号	回帰係数	F 値	変数番号	回帰係数	F 値	変数番号	回帰係数	F 値
(1)	1.30	2.86	15	0.85	1.68	1	1.47	5.09
2	−0.09	0.02	16	−0.80	1.36	3	2.38	11.54
(3)	2.58	9.15	(17)	1.14	2.50	5	2.87	11.62
4	−0.82	1.10	(18)	2.83	11.82	7	1.22	3.13
(5)	2.39	5.43	(19)	−1.64	1.88	14	3.27	28.62
6	−0.16	0.03	20	−0.69	0.54	17	1.15	3.08
(7)	1.12	1.38	(21)	1.27	1.71	18	2.75	14.19
8	0.41	0.20	22	−0.45	0.26	19	−2.11	4.53
9	0.25	0.09	23	−0.12	0.02	21	1.08	2.38
10	0.86	0.82	24	0.20	0.04	28	1.09	3.79
11	0.29	0.12	25	0.04	0.00			
12	0.04	0.00	26	−0.49	0.44			
13	0.72	1.43	27	1.00	2.12			
(14)	2.75	15.67	(28)	0.72	1.11			
		寄与率	0.794				0.783	

1) 最終的に残ったものが上顎歯左側半分のしかも1本おきのものであり，あとはそれらと相関の最も低い部分である下顎歯の右側半分の1本おきのもの，すなわち互い違いの位置にあるものである．

2) 大臼歯4本のうち3本が残っている．

3) 最終的に残ったものは，yとの単相関の高いものはほぼ選ばれているものの，すべてがその順序通りではない．

4) 最初の段階において有意でなかったものが，必ずしも棄却はされていない．

5) x_{19}と年齢yとは正の相関をしているにもかかわらず，係数は負であるという特性を下顎の右側犬歯はもっている．

6) 最初の段階で上顎歯右側第2小臼歯x_{12}が除去されたが，これのF値は当然最低ではあったが，yとの相関は高い方であった．

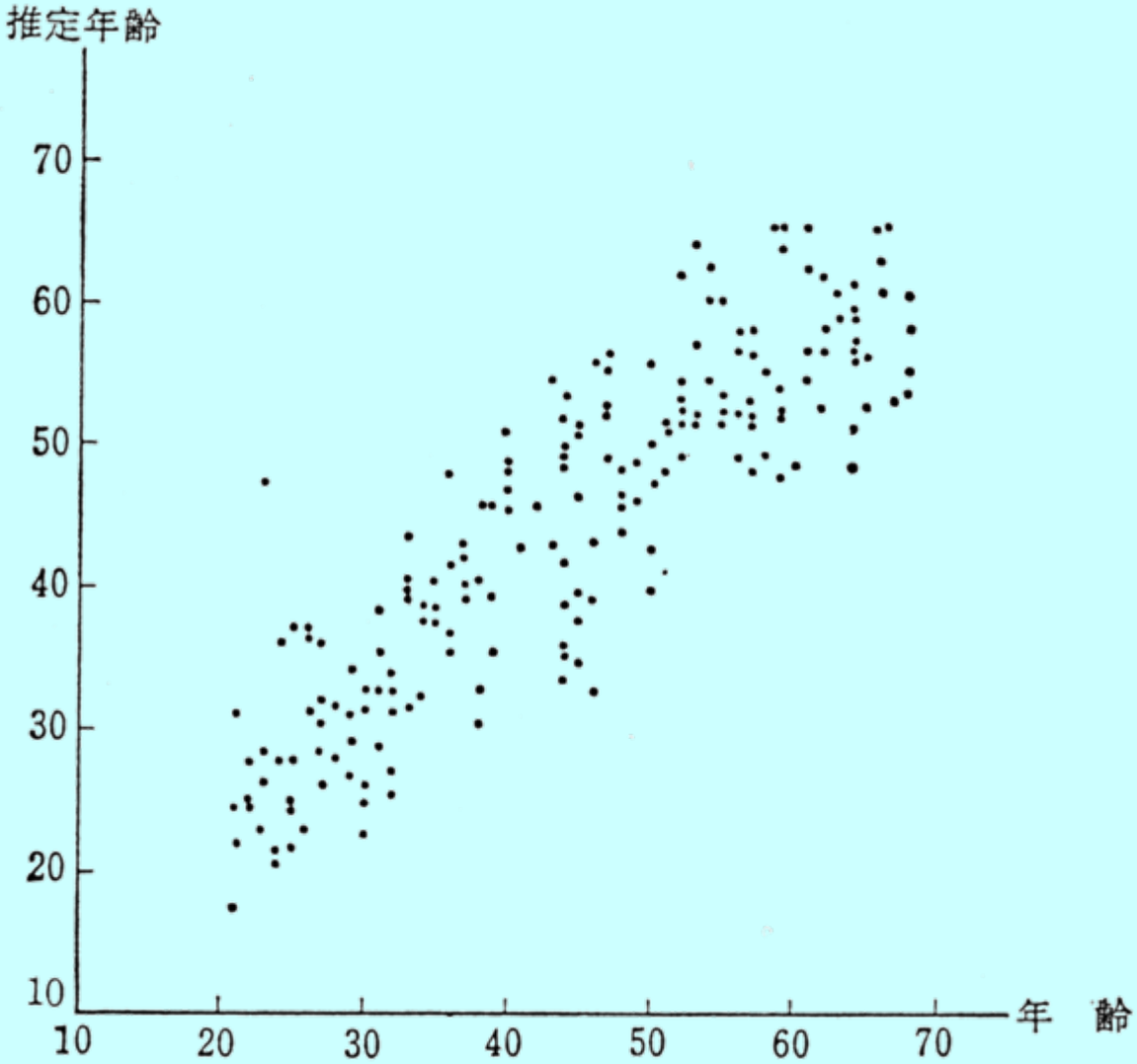

図 4.11　最終段階で選ばれた10本の歯による推定年齢と年齢との相関図

最終段階で選択された10本による寄与率は0.78であり，説明変数が18本も減少しているにもかかわらず0.01しか寄与率は落ちていない．普通，1割から2割減少することが多いのだが，この場合は各変数の相関が0.4～0.6程度にもかかわらず，1％程度しか落ちていないのはなぜだろうか．そのひとつの理由は，一つ一つの変数の散らばり方がパターンとして似ているからであろう．最終的に選ばれた10個の変数を除いた18個はほとんど寄与する力はなく，18個は除いてもよいといえる．

選択された10本の歯のなかでの x_{19}（表4.8），また，それと対称な位置にある10本の歯による重回帰式（表4.9）での x_{24} は，それぞれの係数がともに負になっていることに気がつく．上顎の犬歯が下顎の犬歯に及ぼす影響はほかの歯の上下顎歯の咬み合わせとは異なり，その動きを妨げようとする力が働くことを考えると，この符号の意味は偶然ではないといえよう．

この例では変数増減法と変数減増法との結果が一致した．数理的には一致するという保障は何もない．一般によく用いられるのは計算時間が少なくてすむ変数増減法である．

表 4.9　選択した変数と左右対称の位置にある10本の歯に基づく重回帰分析

変数番号	回帰係数	F値
1	2.36	10.50
8	1.85	7.08
10	2.30	6.46
12	1.76	5.46
14	3.12	19.69
15	2.07	11.85
22	0.95	1.70
24	−0.15	0.02
25	1.12	1.73
26	0.23	0.09
寄与率	0.727	

変数選択で $F_{IN}=F_{OUT}=2.0$ という基準で選ばれた10本の歯による寄与率は0.783であったが，実際に年齢推定を行う際には0.783より0.01程度小さい，すなわち，0.773より大きくなる10本の歯の組合せに基づく年齢推定式を，すべて求めておくとよいであろう．0.773以上と述べたが，これはなんら根拠のある数字ではなく，場合によっては0.763以上としてもよいであろうし，また9本あるいは8本の歯の組合せまで含めていろいろな場合の推定式を求めておくことは実用上有益である．